Automatización del control de palas mineras mediante detección gestual con redes neuronales profundas (Minería 4.0 - Automatización y nuevas tecnologías)

Jose Esteban Malpartida Valverde[1] y Abner Franco Antezano Granados[2]

[1] Autor: MS4M, Calle Sta Rosalía 617, Lima, Perú (jose.malpartida@ms4m.com y 999155487)
[2] Coautor 1: MS4M, Calle Gerard Blanchere 135, Lima, Perú (abner.antezano@ms4m.com y 914714350)

RESUMEN

En el sector minero peruano, la operación de palas constituye una de las actividades más críticas por las exigencias físicas y cognitivas que enfrentan los operadores. Entre 2018 y mayo de 2023 se registraron 138 accidentes en 67 minas formales, resultando en 220 fallecidos, según el Ministerio de Energía y Minas. Esta problemática motivó el desarrollo de un sistema de control gestual para palas mineras basado en redes neuronales profundas, con el fin de mejorar la ergonomía, la seguridad y la eficiencia operativa.

El sistema emplea visión por computadora para interpretar gestos manuales mediante el modelo YOLO Pose, entrenado con datasets públicos y validado con altos niveles de precisión. La solución fue implementada en Unity con inferencia local y comunicación gRPC, permitiendo controlar una pala a escala. Las pruebas demostraron una precisión del 92 % y un tiempo de respuesta promedio de 30 ms, consolidando su viabilidad como alternativa innovadora en la minería 4.0.

Palabras clave: Minería 4.0, visión por computadora, detección de gestos, YOLO Pose, interfaz humano-máquina, inteligencia artificial.

1. Introducción

La minería representa un pilar fundamental de la economía peruana. No obstante, la operación de maquinaria pesada como las palas mineras presenta desafíos considerables. Los operadores están expuestos a largas jornadas laborales, condiciones ambientales extremas y la presión de evitar errores operativos que pueden tener consecuencias fatales. Frente a esta situación, se vuelve imprescindible optimizar la interacción humano-máquina.

Según Chacón Soto (2024), entre 2018 y 2023 se produjeron más de 138 accidentes en minas formales peruanas, con más de 200 muertes registradas. Esta evidencia estadística refuerza la necesidad de adoptar tecnologías emergentes como parte de la transformación digital en la minería, en línea con los principios de la Minería 4.0. Entre estas tecnologías, la visión por computadora y el reconocimiento gestual ofrecen alternativas prometedoras para sustituir o complementar mandos tradicionales, reduciendo el contacto físico, la fatiga y los errores humanos.

La propuesta consiste en una interfaz basada en gestos manuales, cuya información es procesada en tiempo real por una red neuronal profunda, permitiendo la ejecución de acciones sobre una pala minera sin necesidad de controles físicos. Este enfoque no solo mejora la ergonomía, sino también permite un entorno de operación más seguro y eficiente.

Con el objetivo de validar la efectividad y aplicabilidad del sistema propuesto en condiciones cercanas al uso real, se diseñó e implementó una evaluación experimental en un entorno controlado. Esta prueba piloto involucró a un grupo de 10 participantes, conformado por individuos con y sin experiencia previa en el sector minero, lo que permitió obtener una visión más amplia sobre la usabilidad y robustez del sistema ante distintos perfiles de usuario. Los participantes interactuaron con una pala minera a escala (ver Figura 1) exclusivamente mediante gestos manuales, que fueron capturados por un sistema de visión por computadora e interpretados por un modelo basado en aprendizaje automático. Los comandos resultantes fueron transmitidos en tiempo real al dispositivo final a través de una interfaz de comunicación basada en *gRPC* (Google Remote Procedure Call), lo cual permitió una comunicación continua y eficiente entre la aplicación en Unity y el sistema embebido.

Esta etapa experimental tuvo como propósito evaluar tres aspectos fundamentales: la precisión del reconocimiento de gestos y su mapeo al comportamiento mecánico de la pala, el tiempo de respuesta del sistema completo, y la experiencia de usuario en términos de naturalidad, control percibido y facilidad de uso. Además, se buscó identificar el grado de adaptabilidad del sistema

frente a variaciones individuales en la forma de gesticular, un aspecto relevante para garantizar su aplicabilidad en escenarios reales y con operadores diversos. La información recopilada en esta fase constituye una base sólida para el análisis de desempeño y usabilidad, también ofrece una primera aproximación a los desafíos y oportunidades de integrar interfaces gestuales en entornos industriales a escala.

Figura 1. Pala minera a escala

2. Objetivos

- Diseñar un sistema de control de palas mineras mediante detección gestual con redes neuronales profundas que permita mejorar la interacción humano-máquina en entornos operativos mineros.
- Sustituir parcialmente los mandos físicos por una interfaz gestual más intuitiva y sin contacto.
- Incrementar la seguridad operativa y minimizar los errores humanos mediante detección gestual.

3. Compilación de Datos y Desarrollo del Trabajo

La presente investigación se desarrolló en seis etapas interrelacionadas, orientadas al diseño, implementación y validación de una aplicación basada en visión por computadora e inteligencia artificial (IA), destinada a facilitar el control de palas mineras mediante gestos manuales, con el objetivo de aumentar la seguridad y reducir el riesgo de accidentes operativos.

Para ello, se siguió una estrategia compuestas de los siguientes componentes:

- *Recolección del conjunto de datos*. A través de datasets de imágenes de manos en diversas condiciones, anotadas con 21 puntos clave, el cual se utilizó para entrenar la red neuronal profunda.
- *Investigación científica*. Estudio de arquitecturas de redes neuronales, priorizando precisión y eficiencia. Se seleccionó y ajustó YOLO Pose para detectar los 21 puntos clave de la mano.
- *Implementación tecnológica*. A través del entrenamiento del modelo usando *data augmentation* y validado con métricas como mAP, logrando un rendimiento adecuado para su uso práctico.
- *Desarrollo del sistema*. Aplicación en Unity que captura gestos con una cámara RGB, ejecuta inferencias con el modelo e interactúa con la maquinaria mediante una comunicación gRPC
- *Gestos operativos*. Codificación de gestos manuales utilizando lógica heurística y umbrales de distancia para asignar comandos operativos con alta fiabilidad.
- *Estudio de usabilidad*. Evaluación de la aplicación analizando la precisión en el reconocimiento de gestos, tiempos de respuesta y facilidad de uso, mediante cuestionarios y observación directa.

3.1. Recolección del conjunto de datos

Se utilizaron tres conjuntos de datos públicos orientados al análisis de manos humanas y reconocimiento de gestos. La combinación de estos recursos permite cubrir una amplia gama de escenarios visuales y variabilidad gestual, necesaria. Los datasets seleccionados fueron: 11k Hands, Gesture Recognition Dataset y 2000 Hand Gestures Dataset.

11k Hands: El conjunto de datos 11k Hands está compuesto por 11,076 imágenes de manos humanas con resolución de 1600×1200 píxeles, capturadas a partir de 190 sujetos entre 18 y 75 años. Se registran tanto vistas palmares como dorsales de ambas manos, con un fondo blanco uniforme y distancias controladas, lo que lo hace ideal para el análisis estructural y biométrico. Incluye metadatos detallados por imagen, como edad, género, color de piel, presencia de accesorios, y condiciones particulares como esmalte de uñas o anomalías (ver Figura 2).

Figura 2. Ejemplos del conjunto de datos 11k hands

Gesture Recognition: El conjunto de datos Gesture Recognition contiene secuencias de video divididas en fotogramas (30 por muestra), con cinco gestos funcionales asociados a comandos multimedia (por ejemplo, “pulgar arriba” para subir volumen, ver Figura 3). Las imágenes tienen resoluciones variables (360×360 o 120×160) y están organizadas en carpetas por secuencia, lo que permite abordar tareas de aprendizaje temporal o reconocimiento dinámico.

Figura 3. Imagen referencial del conjunto de datos Gesture Recognition

- **2000 Hand Gestures:** El conjunto de datos 2000 Hand Gestures presenta un conjunto de 2,006 imágenes clasificadas en 8 categorías gestuales como closedFist, openPalm o singleFingerBend, con una resolución uniforme de 207×207 píxeles. Las imágenes fueron capturadas con sujetos en entornos semiestructurados, aportando diversidad a nivel de poses y condiciones de iluminación (ver Figura 4).

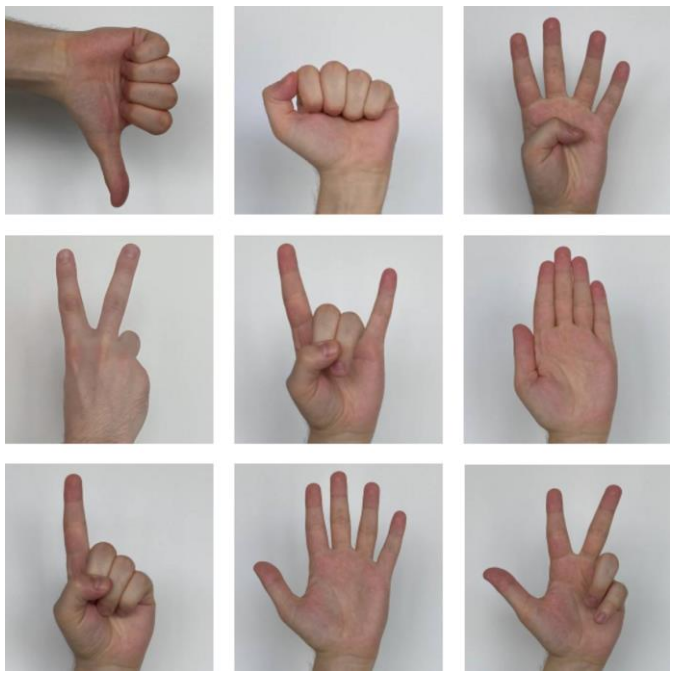

Figura 4. Ejemplos del conjunto de datos 2000 Hand Gestures

Una vez recopilados los conjuntos de datos anteriores, se procedió a prepararlos y depurarlos con el fin de estandarizar su estructura. Entre los principales pasos realizados se incluyeron:

- La normalización de coordenadas clave en los casos donde se extrajeron puntos de referencia anatómicos.
- El filtrado de imágenes con oclusiones, mala iluminación o baja calidad visual.
- La conversión a un esquema estructurado de datos por instancia, asegurando compatibilidad entre los distintos orígenes del dataset.

Este procesamiento fue crucial para garantizar la coherencia del conjunto de entrenamiento.

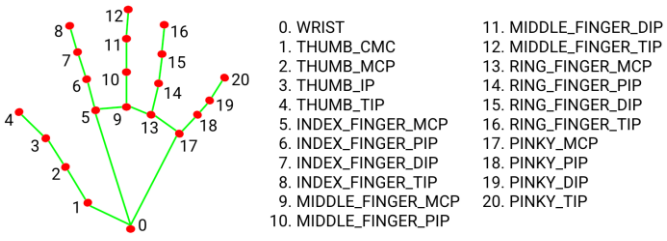

Figura 5. Anotación de los 21 puntos clave de la mano utilizados.

3.2. Investigación científica

Esta etapa consistió en una revisión y análisis comparativo de diversas arquitecturas de redes neuronales aplicadas al reconocimiento de gestos en tiempo real. El objetivo fue identificar el modelo con mejor balance entre precisión, velocidad y robustez para condiciones visuales adversas.

Se compararon arquitecturas de última generación como RTMDet y YOLO Pose. Para la evaluación se consideraron tres métricas fundamentales:

- **mAP@0.5 (mean Average Precision a IoU ≥ 0.5):** Refleja la precisión media cuando el solapamiento entre la predicción y la etiqueta real es del 50 % o más.
- **mAP@0.5:0.95:** Indica la precisión media en diferentes umbrales de superposición, proporcionando una evaluación más exhaustiva.
- **FPS (Frames per Second):** Mide la capacidad del modelo para procesar imágenes en tiempo real, esencial para sistemas interactivos.

Los resultados aprecian que el modelo de IA llamado YOLO Pose ofrece una mayor precisión en ambas métricas mAP, mientras que RT-DERT presenta un rendimiento ligeramente superior en velocidad. No obstante, ambos modelos se encuentran dentro del rango operativo requerido para aplicaciones en tiempo real (ver Tabla 1).

Modelo	mAP@0.5	mAP@0.5:0.95	FPS
YOLO Pose	83.4%	52.1%	32-45 FPS
RTMtDet	81.1%	50.2%	40-50 FPS
RT-DETR	84.5%	53.2%	22-30 FPS
PP-YOLOE+	82.3%	50.8%	30-38 FPS

Tabla 1. Comparativa de rendimiento entre modelos de detección gestual

En el contexto del control automatizado de palas mineras, donde la precisión y la fiabilidad del reconocimiento gestual son críticas, YOLO Pose se posiciona como la mejor opción. Su mayor exactitud en la detección y su rendimiento estable en tiempo real lo convierten en el modelo más adecuado para este entorno operativo.

3.3. Implementación tecnológica

Se llevó a cabo la implementación del modelo de visión por computadora que sirve como núcleo del sistema de control de palas mineras. El enfoque principal de esta sección es describir la estructura y el entrenamiento de la red neuronal, así como los resultados que validan su eficacia como base tecnológica del proyecto.

El modelo de inteligencia artificial utilizado para la detección gestual es YOLO Pose, una red neuronal profunda diseñada para el reconocimiento de objetos y la estimación de poses. Este modelo se encarga de analizar las imágenes en tiempo real para localizar las manos y estimar 21 puntos clave en cada una de ellas (como nudillos y falanges), lo que permite interpretar con precisión los gestos, movimientos y posiciones.

La Figura 6 muestra la arquitectura de la red neuronal convolucional empleada en el sistema, compuesta por tres bloques funcionales: un backbone, una unidad de fusión de características y un cabezal de detección. En primer lugar, el backbone (basado en una variante de Darknet) actúa como extractor de características, generando mapas a distintas profundidades que capturan información tanto local como contextual de la imagen de entrada. A continuación, estos mapas se integran mediante una estructura de fusión tipo PANet, que combina la información a través de una ruta *top-down*, que lleva detalles semánticos desde capas profundas hacia resoluciones más altas, y una ruta *bottom-up*, que refuerza la precisión espacial desde las capas iniciales. Esta combinación permite construir una pirámide de características enriquecidas, capaz de mejorar la detección frente a variaciones en escala, posición o iluminación, condiciones frecuentes en entornos industriales como el minero.

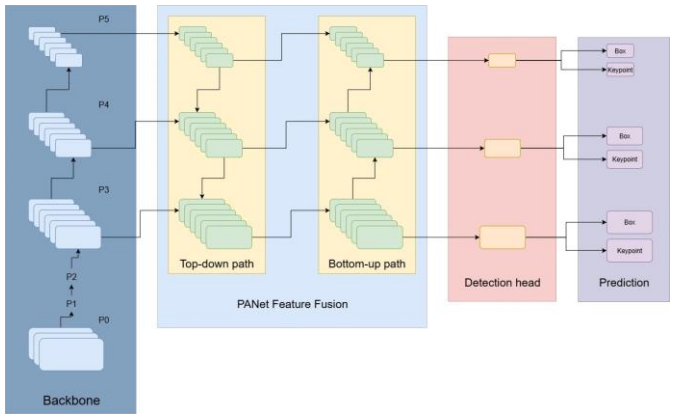

Figura 6. Arquitectura de la red neuronal convolucional.

El entrenamiento del modelo, un proceso crucial para su rendimiento, se llevó a cabo durante 200 épocas. Para fundamentar la elección y la efectividad del modelo, se evaluaron métricas de rendimiento cruciales tanto para las cajas

delimitadoras (B) como para los puntos clave (P) de la mano. La precisión mide la exactitud del modelo al determinar qué detecciones son correctas, mientras que el recall evalúa la capacidad del modelo para encontrar todas las detecciones relevantes. La métrica mAP (mean Average Precision) proporciona un resumen de la precisión y el recall en diferentes niveles de confianza, con mAP50 centrado en detecciones bien ubicadas y mAP50-95 evaluando el rendimiento en un rango más amplio de exigencia.

Los resultados finales de esta evaluación arrojaron métricas favorables para la detección de las cajas delimitadoras y las estimación de los puntos clave de las manos (ver Tabla 2).

Nombre	Valor
metrics/precision(B)	0.9782
metrics/recall(B)	0.97684
metrics/mAP50(B)	0.99097
metrics/mAP50-95(B)	0.89858
metrics/precision(P)	0.98873
metrics/recall(P)	0.98822
metrics/mAP50(P)	0.99349
metrics/mAP50-95(P)	0.99298

Tabla 2**.** Métricas de rendimiento del modelo.

La implementación tecnológica del sistema, basada en la estructura del modelo YOLO Pose y sus resultados de entrenamiento, proporciona una base sólida y validada para el control gestual de las palas mineras. El rendimiento superior del modelo asegura la precisión y confiabilidad necesarias para una operación segura y eficiente.

El modelo fue probado en una estación de trabajo de alto rendimiento, cuyas especificaciones se detallan en la Tabla 3. Esta máquina cuenta con un procesador Intel® Core™ i7-13700K de 13.ª generación con 16 núcleos y 24 hilos, lo cual permitió ejecutar procesos de inferencia y entrenamiento de manera eficiente y paralela. Para el procesamiento gráfico, se utilizó una GPU NVIDIA GeForce RTX 4090 con 24 GB de memoria dedicada, clave para acelerar la detección de gestos y estimación de puntos clave mediante el modelo YOLO-Pose. Asimismo, el equipo dispone de 64 GB de memoria RAM DDR4, lo que permitió manejar sin inconvenientes grandes volúmenes de datos, mantener múltiples procesos activos y evitar cuellos de botella durante las pruebas. Esta configuración aseguró que el sistema cumpliera con los requisitos de procesamiento en tiempo real, manteniendo una latencia inferior a los 30 ms durante su operación.

Componente	Especificación
CPU	Intel® Core™ i7-13700K, 13.ª gen, 16 núcleos (8P+8E), 24 hilos, 3.4 GHz base
GPU	NVIDIA GeForce RTX 4090, 24 GB VRAM dedicada, 55.9 GB total disponible
RAM	64 GB DDR4, 2133 MHz, dual channel (2/4 ranuras en uso)

Tabla 3. Especificaciones del equipo utilizado para probar el modelo.

3.4. Desarrollo del sistema

La aplicación fue desarrollada en Unity, una de las plataformas más versátiles para la creación de contenido interactivo en 2D y 3D. Gracias a su motor gráfico en tiempo real y su capacidad multiplataforma, Unity se utilizó como entorno central para la visualización, interacción y control del sistema, combinando procesamiento visual, inteligencia artificial y comunicación remota.

Uno de los elementos centrales del sistema fue la ejecución de un modelo de visión por computadora en formato ONNX, específicamente el modelo YOLO Pose. Este modelo fue entrenado y adaptado para la detección de manos y la estimación de 21 puntos clave por cada mano, permitiendo interpretar gestos, posiciones y movimientos con alta precisión. La ejecución del modelo se llevó a cabo directamente dentro de Unity mediante Unity Sentis, una tecnología que permite realizar inferencias de modelos ONNX localmente, sin necesidad de conexión a la nube (ver Figura 7). Como entrada, el modelo recibe imágenes RGB

capturadas en tiempo real por la cámara del dispositivo, lo que habilita un análisis visual continuo y de baja latencia.

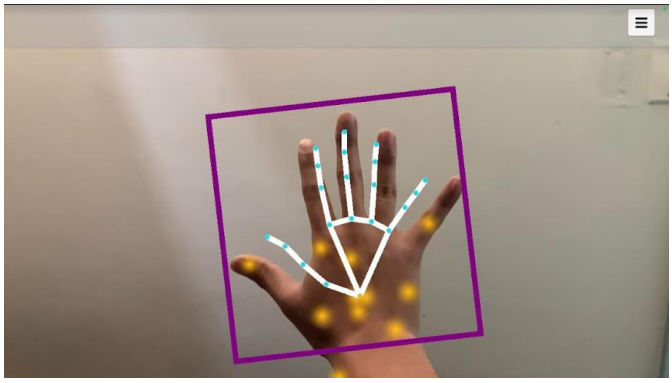

Figura 7. Imagen referencial del sistema.

Complementando esta funcionalidad, se implementó un cliente gRPC integrado en Unity, el cual cumple un rol clave en la arquitectura de comunicación del sistema (ver Figura 8). Este cliente permite una transmisión eficiente, rápida y estructurada de datos al establecer una conexión directa con un dispositivo final (endpoint), conectado a una pala minera a escala. Para esta tarea, se utilizó el modo de comunicación Client Streaming de gRPC, lo que permite enviar una secuencia continua de comandos desde Unity al dispositivo final en una única conexión persistente. Esto resulta fundamental para mantener una interacción fluida y en tiempo real, ya que los gestos manuales pueden cambiar rápidamente y requieren una respuesta inmediata del sistema.

A través del Client Stream, se transmiten comandos previamente interpretados a partir de los gestos detectados, los cuales contienen las posiciones y rotaciones deseadas del efector final (eslabón terminal) de la pala. Una vez recibidos, el dispositivo aplica un algoritmo de cinemática inversa (IK, por sus siglas en inglés) para calcular los ángulos que deben adoptar las articulaciones intermedias (como hombro, codo y muñeca del brazo mecánico) para alcanzar la posición objetivo. Este proceso considera las restricciones físicas del sistema, como los límites de rotación y longitudes de los eslabones, asegurando movimientos precisos, suaves y físicamente viables. Así, se garantiza que la pala reproduzca con fidelidad la intención del usuario, cerrando un bucle de control continuo y eficiente mediante el uso de gRPC en modalidad Client Streaming.

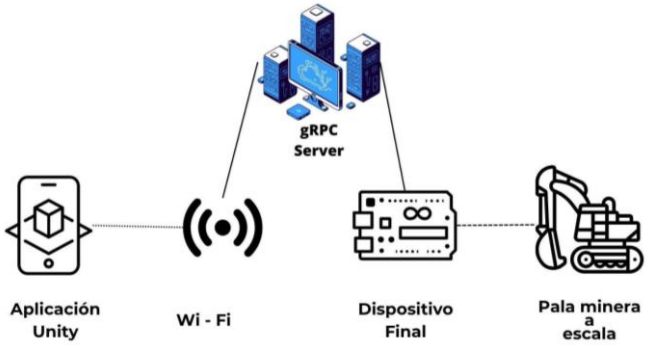

Figura 8. Diagrama de flujo del sistema.

3.5. Gestos operativos

Se implementó un sistema de reconocimiento de gestos a partir de los 21 puntos clave de la mano detectados por el modelo de IA, con el objetivo de interpretar comandos de movimiento para el efector final de la pala minera. La lógica se basó en el análisis de la posición relativa de los dedos, utilizando reglas heurísticas y umbrales geométricos para determinar qué dedos se encontraban extendidos y en qué dirección apuntaban. Este enfoque permitió una detección confiable de gestos en tiempo real, incluso en condiciones de operación variables.

Para controlar los movimientos en el espacio tridimensional (ejes X, Y, Z), se definieron seis gestos manuales únicos, cada uno vinculado a una dirección de desplazamiento específica del efector final de la pala:

Gesto 1

Este gesto requiere que el operador levante los dedos medio, índice y pulgar, con una ligera inclinación de la mano hacia el lado izquierdo (ver Figura 9). Representa un movimiento lateral hacia la izquierda (eje X negativo). La posición combinada de tres dedos y la inclinación direccional permite una detección robusta del gesto, evitando ambigüedades.

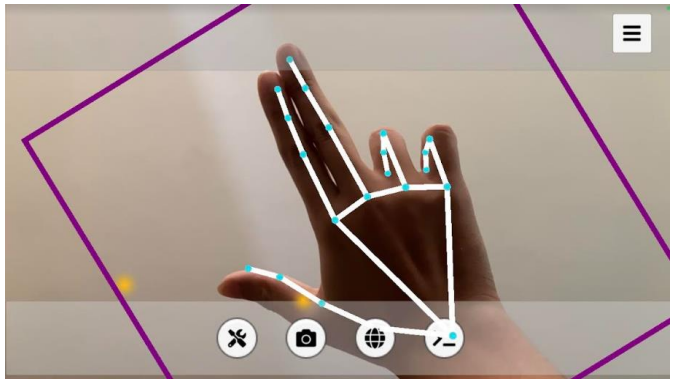

Figura 9. Imagen referencial del gesto 1

Gesto 2

Similar al gesto anterior, pero con la inclinación hacia la derecha (ver Figura 10). Los tres dedos se extienden, y la orientación de la mano determina la dirección del comando. Este gesto se traduce en un movimiento lateral hacia la derecha (eje X positivo), facilitando maniobras de posicionamiento horizontal del efector final.

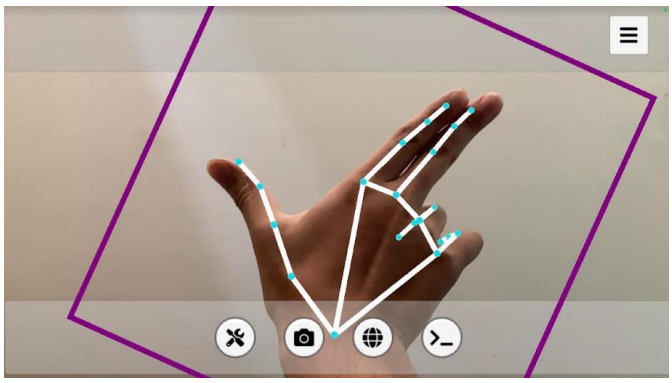

Figura 10. Imagen referencial del gesto 2

Gesto 3

Este gesto consiste en extender el dedo índice hacia arriba y el pulgar hacia un costado, mientras los demás dedos permanecen doblados (ver Figura 11). Se interpreta como un comando de movimiento vertical ascendente (eje Y positivo), indicando que el efector final debe elevarse. Es un gesto intuitivo y ampliamente reconocido por su claridad visual.

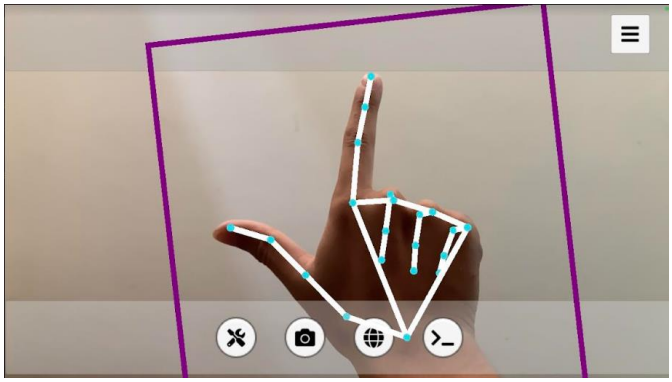

Figura 11. Imagen referencial del gesto 3

Gesto 4

En este gesto, el dedo índice apunta hacia abajo mientras el pulgar permanece extendido lateralmente (ver Figura 12). De igual manera, los demás dedos siguen permaneciendo doblados. Se asocia con un movimiento vertical descendente (eje Y negativo), útil para accionar el descenso controlado de la pala.

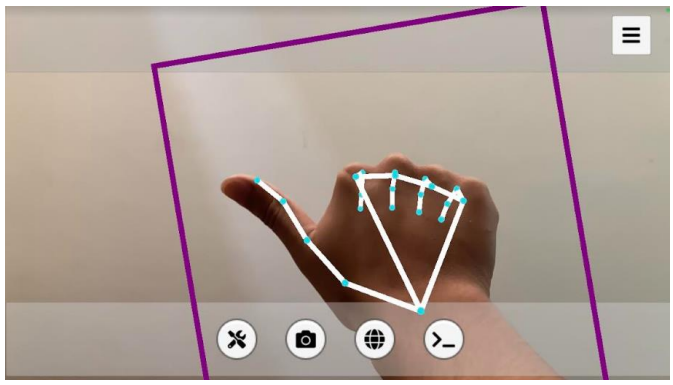

Figura 12. Imagen referencial del gesto 4

Gesto 5

Cuando únicamente el dedo índice se encuentra extendido en una posición recta, apuntando hacia adelante, se interpreta como un comando de avance (eje Z positivo, ver Figura 13). Este gesto es simple, natural y fácil de reconocer, ideal para indicar desplazamientos frontales del efector en la dirección del operador.

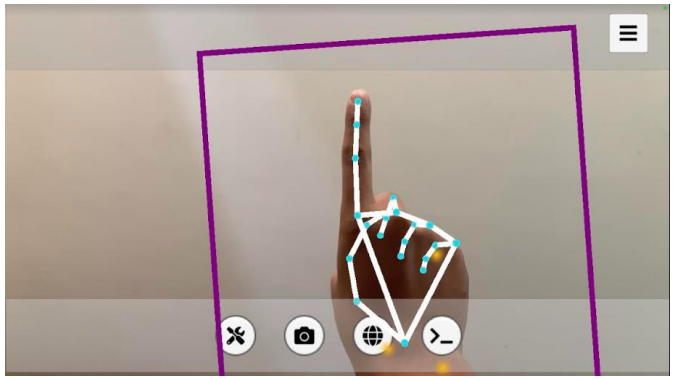

Figura 13. Imagen referencial del gesto 5

Gesto 6

En este gesto, el pulgar apunta hacia un lado y el meñique se extiende, mientras el resto de los dedos permanece recogido (ver Figura 14). Esta configuración, poco común en gestos involuntarios, se asocia con un movimiento de retroceso (eje Z negativo), permitiendo desplazar la pala hacia atrás con precisión.

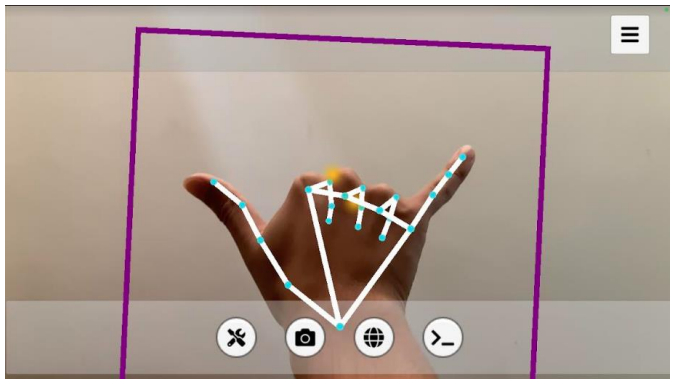

Figura 14. Imagen referencial del gesto 6

Estos gestos fueron seleccionados considerando criterios de claridad visual, facilidad de ejecución, baja ambigüedad y robustez frente a variaciones individuales. El sistema interpreta los gestos en tiempo real, y los comandos resultantes se transmiten a través del cliente gRPC al servidor de control, permitiendo un control seguro e intuitivo de la pala minera sin contacto físico con el equipo.

3.6. Estudio de usabilidad

Con el fin de validar la efectividad e intuición del sistema propuesto, se realizó un estudio de usabilidad con 10 participantes, incluyendo tanto usuarios vinculados al sector minero como individuos sin experiencia previa. Diversos estudios han demostrado que una muestra de cinco usuarios es suficiente para identificar aproximadamente el 85 % de los problemas de usabilidad (Nielsen & Landauer, 1993). No obstante, se optó por ampliar el número de participantes para aumentar la fiabilidad de los hallazgos y abarcar una mayor diversidad de perfiles y comportamientos. Este enfoque permitió una evaluación más representativa de la interacción entre los gestos manuales y el sistema de control de palas mineras en un entorno simulado.

Durante las sesiones de prueba, cada participante interactuó con la aplicación en un entorno simulado, utilizando la cámara para emitir gestos y ejecutar comandos en el efector final de la pala. Se registraron diferentes métricas cualitativas y cuantitativas para evaluar la interacción:

- *Comodidad en el uso de gestos*. Calificación de los usuarios en el esfuerzo físico y la naturalidad al mantener y ejecutar cada uno de los seis gestos definidos, considerando factores como fatiga, repetitividad y facilidad postural.

- *Aprendizaje gestual*. Evaluación del tiempo que tomó a los participantes comprender y recordar la asociación entre cada gesto y el comando correspondiente del sistema, así como su evolución durante la prueba.

- *Manipulación efectiva*. Capacidad de los usuarios para completar tareas específicas como mover la pala en los tres ejes (XYZ), detenerla en posiciones precisas y cambiar rápidamente de dirección mediante gestos, valorando tanto la precisión como el control intuitivo.

- *Precisión gestual*. Medición en porcentaje de los gestos de la mano correctamente identificados en tiempo real por el sistema.

- *Tiempo de respuesta*. Intervalo entre la ejecución del gesto y la respuesta de la pala minera, verificando la fluidez de la interacción.

Además, se aplicó un cuestionario de satisfacción y usabilidad al finalizar la prueba, basado en la escala SUS (System Usability Scale), complementado con preguntas abiertas para recoger observaciones cualitativas. Los resultados indicaron una percepción mayoritariamente positiva en cuanto a la facilidad de uso, la utilidad del sistema y el potencial de este tipo de interfaz gestual para reducir el contacto físico y mejorar la seguridad operativa en entornos mineros.

4. Presentación y discusión de resultados

Durante la evaluación de usabilidad, se aplicó una escala cualitativa de 1 a 5 para tres métricas clave: comodidad en la ejecución de gestos, aprendizaje de la relación gesto-comando y manipulación efectiva de la pala minera. A continuación, se presentan los resultados obtenidos y su respectiva discusión.

4.1. Comodidad en el uso de gestos

En cuanto a la comodidad al ejecutar los gestos, 7 de los 10 participantes calificaron la experiencia con un 3 (nivel intermedio), mientras que los 3 restantes otorgaron una puntuación de 4, lo que indica una percepción levemente positiva. Aunque no se reportaron molestias físicas relevantes, varios usuarios señalaron que algunos gestos requerían mantener posturas poco habituales. Este resultado sugiere que el sistema es moderadamente cómodo, pero que podría beneficiarse de ajustes en el diseño gestual para reducir la fatiga en usos prolongados.

4.2. Aprendizaje gestual

Respecto a la facilidad de aprendizaje de los gestos y su asociación con los comandos de la pala, los resultados fueron más favorables: 5 usuarios puntuaron con un 3, 3 con un 4, y 2 con un 5. Esto indica que, aunque la mayoría tuvo una experiencia de aprendizaje neutra o positiva, un grupo considerable necesitó cierto tiempo para interiorizar la lógica de la interfaz gestual. No obstante, se observó que todos los participantes mostraron mejoras notables durante la sesión, lo que

evidencia una curva de aprendizaje accesible y eficaz en contextos breves de entrenamiento.

4.3. Manipulación efectiva

En la categoría de manipulación del sistema, los resultados fueron notablemente positivos: 7 usuarios calificaron con un 4, 1 con un 5, y solo 2 con un 3. Esto refleja que la gran mayoría logró controlar la pala minera con precisión, realizar movimientos en los tres ejes y responder correctamente a los gestos definidos. La alta puntuación en esta métrica respalda la viabilidad funcional del sistema y su aplicabilidad en entornos reales, donde el control intuitivo y sin contacto físico puede mejorar tanto la seguridad como la eficiencia operativa.

4.4. Precisión gestual

Durante las pruebas, se registró una precisión promedio del 92 % en el reconocimiento de los seis gestos definidos, medida como el porcentaje de gestos correctamente interpretados por el sistema en comparación con el total de intentos realizados por los usuarios. La mayoría de los errores se produjeron en condiciones de iluminación baja o cuando los gestos eran realizados de manera ambigua o incompleta. A pesar de estos casos, la precisión alcanzada es suficiente para aplicaciones prácticas, especialmente considerando que los usuarios mejoraron su ejecución a medida que se familiarizaron con el sistema. Esto demuestra que el modelo de IA empleado, junto con la lógica de interpretación gestual, presenta un nivel alto de fiabilidad operativa.

4.5. Tiempo de respuesta

El tiempo de respuesta del sistema, medido como el intervalo entre la ejecución del gesto y la respuesta visual o mecánica del efector final, fue de aproximadamente 30 milisegundos (ms) en promedio. Este valor abarca la captura de la imagen (~5 ms), la inferencia del modelo YOLO Pose (~15 ms), la interpretación del gesto (~5 ms) y el envío del comando mediante gRPC (~5 ms). El sistema se mantuvo dentro de un rango de respuesta compatible con aplicaciones en tiempo real a 30 fps, sin generar demoras perceptibles para el usuario. Esto evidencia que la integración técnica entre Unity Sentis, el modelo ONNX y el canal de comunicación gRPC fue eficiente y adecuada para entornos interactivos, incluso en condiciones de carga moderada.

Por un lado, los resultados cualitativos muestran que el sistema presenta un nivel aceptable a positivo en usabilidad, siendo especialmente fuerte en la capacidad de manipulación, lo cual es crucial para su uso en operaciones mineras. Las métricas indican que con mejoras menores en la comodidad gestual, el sistema podría alcanzar un nivel de aceptación aún mayor (ver Tabla 3).

Métrica	1	2	3	4	5	Promedio	Observación
Comodidad	0	0	7	3	0	3.3	Aceptable, con oportunidad de mejora
Aprendizaje	0	0	5	3	2	3.7	Positivo, con curva de aprendizaje breve
Manipulación	0	0	2	7	1	3.9	Muy positiva, control preciso e intuitivo

Tabla 4. Métricas cualitativas

Por otro lado, los resultados cuantitativos respaldan la solidez técnica de la aplicación, permitiendo su uso eficiente y confiable en tareas de control remoto de maquinaria pesada mediante gestos, donde tanto la rapidez como la exactitud son factores críticos para la seguridad y la eficacia operativa (ver Tabla 4).

Métrica	Valor medio	Observación
Captura de imagen	~0–5 ms	Ideal en capturas a 30 fps.
Inferencia del modelo	≤ 15 ms	Buen tiempo optimizado en Unity Sentis.
Interpretación del gesto	≤ 5–8 ms	Lógica implementada viable.
Comunicación gRPC	≤ 5–8 ms	Optimizado en red local

Tabla 5. Métricas cuantitativas

5. Conclusiones

La integración de la arquitectura YOLO-Pose junto con una lógica heurística para la interpretación de gestos logra resultados altamente satisfactorios en términos de precisión (92 %) y tiempo de respuesta (30 ms), lo cual confirma su idoneidad para aplicaciones en tiempo real en entornos exigentes como el sector minero. Estos indicadores reflejan

que el sistema propuesto no solo es técnicamente sólido, sino también lo suficientemente eficiente como para operar en contextos donde la velocidad de respuesta y la precisión son factores críticos.

Además, el uso de gestos manuales como método de interacción representa un avance significativo en términos de ergonomía y seguridad laboral. Al eliminar la necesidad de interfaces físicas tradicionales, se reduce la carga física y cognitiva sobre el operador, lo que potencialmente disminuye la probabilidad de errores humanos y contribuye a una operación más segura y eficiente de maquinaria pesada. Este enfoque se alinea con los objetivos de automatización industrial centrada en el usuario.

Finalmente, las pruebas de usabilidad demostraron que el sistema presenta una curva de aprendizaje accesible, permitiendo que los usuarios comprendieran e implementarán los gestos en un corto periodo de tiempo. Esta facilidad de adopción, combinada con un desempeño técnico robusto, evidencia que la propuesta no solo es viable a nivel experimental, sino también aplicable en escenarios reales de operación, lo que abre camino hacia futuras implementaciones en campo y escalabilidad del sistema.

6. Referencias bibliográficas

Afifi, M. (2019). *11K Hands: Gender recognition and biometric identification using a large dataset of hand images*. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-019-7424-8

Chacón Soto, V. (2024, abril 3). Minería en Latinoamérica sigue causando peligrosos accidentes. Semanario Universidad. https://semanariouniversidad.com/pais/mineria-en-latinoamerica-sigue-causando-peligrosos-accidentes/

Giridhar, R. (2022). *Hand gestures dataset* [Conjunto de datos]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/2206851

Nielsen, J., & Landauer, T. K. (1993, 24-29 de abril). A mathematical model of the finding of usability problems. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93). Association for Computing Machinery, 206-213*. https://doi.org/10.1145/169059.169166

Sparsh, I. (s. f.). *Gesture Recognition* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/imsparsh/gesture-recognition

Ultralytics. (s. f.). *YOLO-Pose: Detección de poses con YOLO*. En *Ultralytics YOLO Docs*. https://docs.ultralytics.com/es/tasks/pose/