Ingeniería y Competitividad
Vol 25 (suplemento) 2023; doi: 0.25100/iyc.v25iSuplemento.13133

Ingeniería y Competitividad

Número Suplemento

ISSN
0123-3033
e- 2027-8284

Artículo de investigación

Edición especial 25 años del doctorado en ingeniería

Sistema de interacción humano-robot para la enseñanza-aprendizaje de una tarea de ordenamiento de objetos mediante comunicación verbal y gestual

Human-robot interaction system for teaching-learning of an object sorting task by means of verbal and gestural communication

Cómo citar: Mosquera-DeLaCruz, José-Hernándo., Martínez-Álvarez, A., Nope Rodríguez, S.P., Loaiza-Correa, H., Rodríguez-Tellez, G.A., Jamioy-Cabrera, J.D., Delgado-Giraldo, M.A., Penagos-Angrino, J.F. Ingeniería y Competitividad. 25(suplemento) ,e- 20613133. doi: 10.25100/iyc.v25iSuplemento.13133

Jose-Hernando Mosquera-DeLaCruz[1], Alexander Martínez-Álvarez[1], Sandra-Esperanza Nope-Rodríguez[§2], Humberto Loaiza-Correa[2], Gabriel-Alejandro Rodríguez-Téllez[1], Juan-David Jamioy-Cabrera[1], María-DeLosÁngeles Delgado-Giraldo[1], Juan-Felipe Penagos-Angrino

[1]Departamento de Electrónica y Ciencias de la Computación, Pontificia Universidad Javeriana Cali, Colombia.
[2]Escuela de Ingeniería Eléctrica y Electrónica, Universidad del Valle, Cali, Colombia.

hernando@javerianacali.edu.co, amartin@javerianacali.edu.co, §sandra.nope@correounivalle.edu.co, humberto.loaiza@correounivalle.edu.co, gabriel2000529@javerianacali.edu.co, juan2018@javerianacali.edu.co, mariadg19@javerianacali.edu.co, juanfepa0105@javerianacali.edu.co.

Resumen

Se desarrolló un sistema interacción humano-robot multimodal (gestos y voz) que permite a usuarios enseñarle tareas de clasificación de cubos por color a un robot. La evaluación del sistema fue realizada por siete usuarios de forma cuantitativa y cualitativa. En las pruebas cuantitativas se evaluó un total de 63 interacciones verbales, 252 interacciones gestuales, y 63 multimodales. El porcentaje de reconocimiento de las interacciones fue del 98.41% para los comandos de voz, 81.35 % para los gestuales, y 80.95% para las multimodales. Luego del aprendizaje, el robot fue capaz de realizar correctamente la tarea de clasificación de cubos por color en un 100%; siendo capaz de responder exitosamente ante condiciones iniciales (ubicaciones y cantidad de cubos) no enseñadas previamente. La evaluación cualitativa del sistema se realizó para conocer la percepción de los usuarios, arrojando resultados consistentes con los porcentajes de reconocimiento, favoreciendo la interacción verbal sobre la multimodal.

Palabras clave: Interacción multimodal humano robot, Aprendizaje por demostración, Clasificación de cubos, Robótica.

Abstract

A multimodal (gestures and voice) human-robot interaction system was developed that allows users to teach color-cube sorting tasks to a robot. The evaluation of the system was performed by seven users in a quantitative and qualitative way. In the quantitative tests, a total of 63 verbal interactions, 252 gestural interactions, and 63 multimodal interactions were evaluated. The recognition rate of the interactions was 98.41% for voice commands, 81.35% for gestural, and 80.95% for multimodal. After learning, the robot was able to correctly perform the task of classifying cubes by color in 100%; being able to respond successfully to initial conditions (locations and number of cubes) not previously taught. The qualitative evaluation of the system was carried out to know the perception of the users, yielding consistent results with the recognition percentages, favoring verbal interaction over multimodal interaction.

Keywords: Multimodal human-robot interaction, Learning by demonstration, Cube classification, Robotics.

doi: 10.25100/iyc.v25i4.13164

Introducción

Cada vez es más común encontrar entornos de trabajo en donde interactúan humanos y robots colaborativos para realizar una tarea, creando la necesidad de desarrollar maneras más naturales de interactuar así como la forma en las que los robots pueden aprender a ejecutar una tarea a partir de la demostración de un humano (1). El uso de un robot colaborativo por parte de un usuario depende en gran medida de la automatización en la programación de sus tareas (2), que generalmente requieren ajustes específicos, tiempos prolongados de programación y reprogramación, y que normalmente son realizadas por un ingeniero experimentado (3,4). Estos requerimientos pueden reducirse en gran medida si para la comunicación entre el usuario y el robot, involucra una interfaz intuitiva que permita la asignación de tareas al robot colaborativo por cualquier persona del común (5). En este sentido, la comunicación multimodal (voz y gestos) entre un robot y un humano es muy útil para permitir la utilización de diferentes canales para transmitir y recibir mensajes con una mayor naturalidad (6–8).

El desarrollo de estos sistemas de interacción multimodales humano-robot involucra muchos factores como por ejemplo: el paradigma de diseño, el escenario de aplicación, las estructuras de los datos y los algoritmos utilizados para su implementación (9). En la literatura se encuentran sistemas basados en el aprendizaje de máquina los cuales utilizan, entre otras, técnicas de visión por computador (10), reconocimiento de voz (11), algoritmos basados en tablas de decisión (12), modelamiento de políticas estocásticas (13), aprendizaje por reforzamiento (14) y por otro lado, sistemas basados en modelos del comportamiento humano como el aprendizaje multimedia (15) y sistemas basados en memorias ACT-R (16), ICARUS (17) y SOAR (18).

En la revisión bibliográfica realizada se encontraron diversos sistemas de interacción multimodal humano-robot aplicados a procesos de clasificación. En (19) se describe un sistema de interacción humano robot basado en un el corpus ELDERLY-AT-HOME (20) en el que interactúa una persona de edad avanzada con su enfermera. En este trabajo, la tarea consiste en encontrar un objeto específico previamente entrenado en un ambiente con tres cajas. El sistema está compuesto por tres módulos: el módulo de interpretación, que captura las señales gestuales y verbales del usuario utilizando un sensor Kinect y posteriormente realiza una fusión de estas señales enviando un vector de acción al siguiente módulo de mediación, el cual es el encargado de administrar el dialogo al inferir el estado actual de la tarea y planificar sus próximas acciones, para enviar las señales de control al módulo de ejecución encargado de controlar el robot. Un año después en (21) se desarrolló una modificación al sistema propuesto en (19), en el que se invierten los papeles en la interacción (Role-Switching) pasando el robot a hacer el papel del adulto mayor y el humano a ser la enfermera. El sistema conservó los mismos tres módulos: en la tarea estudiada, el robot escoge un objeto aleatoriamente e interactúa con el humano para que le ayude a buscar ese objeto en los cajones presentes en la escena. Otra tendencia identificada fue el aprendizaje mediante sistemas de memoria, (22) estudia el ensamble de una caja de madera guiado por los gestos del usuario capturados por sensores de medición inercial (IMU) y cámara para validar el ensamble. El sistema está compuesto por tres etapas: percepción, la cual determina el

doi: 10.25100/iyc.v25i4.13164

estado actual de la tarea, memoria, etapa que define el estado actual de la tarea basada los episodios pasados y realiza una predicción de las acciones de colaboración a realizar, para pasar la información a control, que realiza la planeación de los movimientos y el accionamiento del robot. Finalmente, en (5) se desarrolló un sistema basado en el principio que los humanos pueden inferir conceptos a partir de imágenes o diagramas, y aplicarlos en el mundo físico, en un entorno completamente diferente. Este estudio propone un sistema entrenado con imágenes sintéticas de prueba para que el sistema aprenda el objetivo de la tarea. Posteriormente, realiza una extracción del concepto para generalizar una situación con los mismos objetos en nuevas posiciones, luego se generaliza una situación con objetos nuevos del mundo real y finalmente, se aplican estas generalizaciones para resolver problemas con objetos físicos reales utilizando un robot manipulador, presentando un diagrama de bloques compuesto cinco módulos principales; módulo de visión jerárquica, módulo de fijación, módulo de atención, módulo de reconocimiento e indexado de objetos y módulo de control del robot manipulador.

Del análisis de los sistemas de interacción multimodal consultados en la literatura, se observa que presentan en común tres etapas principales: una primera etapa encargada de realizar la captura de las señales multimodales, una segunda etapa que realiza la planeación del siguiente movimiento del robot, y una tercera etapa que acciona el robot para realimentar al usuario. En los sistemas de aprendizaje con políticas estocásticas (stochastic policies), estas etapas se denominan interpretación, mediación y ejecución (19,21), y en los sistemas de aprendizaje por memorias, se denominan percepción, memoria y control (22). También se observa que en los sistemas consultados existen bloques internos encargados de combinar y desagregar la información de los sensores. Otro aspecto que se evidenció en estos sistemas de interacción humano-robot fue la ausencia de mecanismos que permitan el aprendizaje en línea de una tarea de clasificación a partir de instrucciones audiovisuales impartidas en un ambiente de experimentación real.

Los sistemas de interacción multimodal presentan diferentes formas de transmitir la información hacia el robot. Algunos sistemas presentan interfaces comandas por gestos, voz, interfaces hápticas y señales biológicas como las electromiográficas o electroencefalográficas. Las interfaces gestuales y verbales son tecnologías libres de contacto, las cuales se inspiran en la comunicación natural humano-humano por lo que reducen la curva de aprendizaje para utilizarlas. Además, permiten enseñar en línea una tarea a un robot mediante demostraciones enriquecidas con diálogos, y no precisa de conocimientos de programación por parte del instructor.

En este trabajo, se abordan tres de los problemas de los sistemas de interacción humano robot (1,23): la transmisión de información hacia el robot integrando señales gestuales y verbales, el aprendizaje en línea de una tarea de organización de objetos, y la generalización para realizar la tarea aprendida ante situaciones nuevas.

Con base en lo anterior, se presenta un sistema de interacción multimodal humano-robot que permite a un robot colaborativo UR3 aprender y ejecutar una tarea de

doi: 10.25100/iyc.v25i4.13164

clasificación de cubos por color, a partir de demostraciones comandas por gestos y voz, con capacidad de generalización. El sistema propuesto consta de cuatro módulos: de entrada, máquinas de estados, sistema de memorias, de salida.

Las secciones 2 y 3 presentan la metodología y la descripción del sistema, respectivamente. La sección 4 discute las pruebas y resultados, mientras que la sección 5 expone las conclusiones y trabajos futuros.

Metodología

La metodología utilizada es del tipo analítico-experimental, donde se parte de un estudio de antecedentes para comparar ventajas y limitaciones de las tecnologías existentes con el fin de seleccionar las más apropiadas a cada etapa, para posteriormente mejorarlas y adaptarlas al problema de enseñanza-aprendizaje de un robot mediante comandos multimodales en la realización de una tarea de organización de objetos.

Para la concepción del sistema se tuvieron en cuenta las relaciones existentes entre la tarea, el espacio de trabajo, los objetos y la interacción multimodal. Esto permitió establecer las especificaciones de cada una de las etapas que constituyen el sistema y su proceso de validación. Por lo anterior, en las secciones siguientes se describen la tarea de clasificación asociada al espacio de trabajo y objetos utilizados. Luego, se describen las interacciones gestuales, verbales y multimodales para la tarea, y finalmente las pruebas para verificar el funcionamiento del sistema.

Tarea de clasificación, espacio de trabajo y objetos

Se utilizó un robot industrial colaborativo (COBOT) UR3 CB de Universal Robots[1] para ejecutar la tarea de identificación de objetos cúbicos diferenciados por color y organizarlos. Este robot pequeño es ideal para aplicaciones sobre mesas de trabajo, debido a que su peso es de 11 kg y posee seis grados de libertad con rotaciones de ± 360° en todas las articulaciones. El UR3 CB cuenta con una morfología y apariencia que facilita la interacción con los usuarios. El robot se dispuso sobre una zona de trabajo de 33.2 cm por 23.2 cm demarcado con cinta roja sobre una superficie metálica. La zona de trabajo se subdividió en tres filas y cuatro columnas, conformando una retícula de doce zonas, cada una de aproximadamente 8.3 cm x 7.7 cm. La ubicación relativa entre la cámara y el robot se ilustra en la vista superior que se observa en la Figura 1. La cámara se ubicó aproximadamente a 50.0 cm de altura y 10.0 cm al frente de la zona de trabajo. Esta disposición facilita la observación de la cara superior de los cubos durante el desplazamiento del robot, y el agarre de éstos con la pinza.

1 https://www.universal-robots.com/es/productos/robot-ur3/

doi: 10.25100/iyc.v25i4.13164

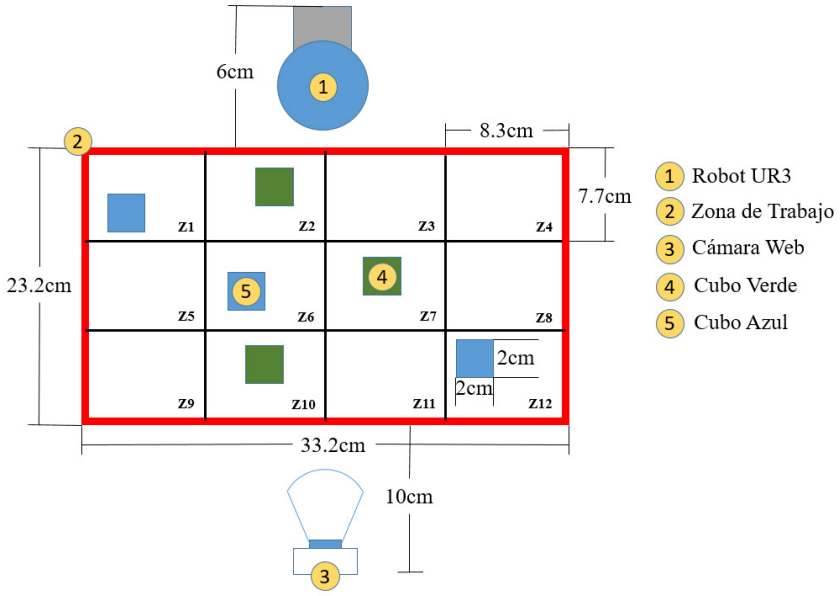

Figura 1. Ilustración del espacio de trabajo y zonas de la cuadrícula.

Se utilizaron seis cubos impresos en 3D con material PLA negro, de 2 cm. dándoles color en la cara superior con cintas de color azul (tres cubos) y verde (tres cubos). Cada cubo se ubica en una zona de la retícula, en cualquier posición y orientación. Sobre la zona de trabajo se pueden disponer entre uno y tres cubos de cada color. La tarea del robot consiste en detectar y clasificar por color los cubos ubicados en la posición inicial (Figura 2, izquierda) y disponerlos en columnas o filas organizados por color en la posición final (Figura 2, derecha).

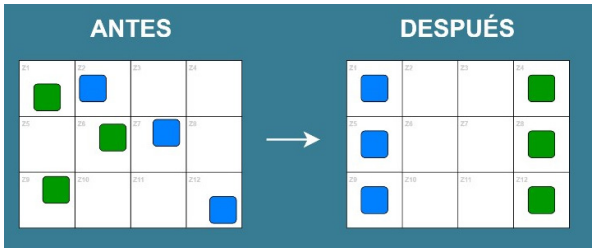

Figura 2. Disposición de los cubos antes y después de la ejecución de la tarea de clasificación.

Las líneas que delimitan la retícula presente en la Figura 2 no existen físicamente en la zona de trabajo y se adicionan en la figura para facilitar la ubicación relativa de los cubos por parte del usuario. Así mismo, en la parte superior izquierda de cada celda se incluye un número para identificar la zona usada en los comandos verbales y gestuales. El sistema cuenta con una cámara en espectro visible Logitech C920, con una resolución estándar de 640x480 pixeles y corrección de iluminación automática. La ubicación relativa entre el robot, la zona de trabajo, la posición y orientación de la cámara, y el demostrador humano, evitan problemas de oclusión. Adicionalmente, el ambiente donde se instaló el sistema es tipo oficina, por lo cual las condiciones de iluminación presentan pocas variaciones.

doi: 10.25100/iyc.v25i4.13164

Interacción gestual

Se definió un único gesto denominado "Seleccionar", que tiene como finalidad demostrar al robot la posición inicial del cubo con el que se va a interactuar y la posición final a la que debe trasladarse. El gesto inicia con la mano abierta (con una separación de al menos 1.5 cm entre la punta del dedo índice y la punta del dedo pulgar) y termina cuando los dedos índice y pulgar se unen mientras la mano permanece abierta (el resto de los dedos no presentan cambios significativos entre sus posiciones relativas). Esta acción se asemeja a la forma en que una persona intenta un agarre de pinza sobre un objeto pequeño. Al ejecutar este gesto sobre o cerca de un cubo por primera vez, se identifica el objeto sobre el cual se desea realizar el desplazamiento. Al repetir este gesto sobre la zona a la cual se desea desplazar el cubo, se determina la posición final.

Interacción verbal

Se definió un diccionario con 27 palabras en español (Tabla 1) asociadas a la definición del estado del robot (columna color amarillo), la descripción del cubo (columnas color verde), la ubicación final del cubo (columna color gris) y la acción a realizar (columnas color azul). El diccionario permite conformar diferentes frases para la enseñanza y ejecución de una tarea de clasificación, por ejemplo: "Toma el cubo verde dos y ponlo en la zona once". Las demás palabras como artículos y preposiciones no son tenidas en cuenta durante el reconocimiento verbal.

Tabla 1. Diccionario de comandos de voz

Definición Estado	**Tomar**	**Objeto**	**Color**	**Índice**	**Dejar**	**Ubicación**	
"Hola" "Aprender" "Ejecutar"	"Toma" "Sujeta" "Coge" "Agarra"	"Cubo"	"Azul" "Verde"	"Uno" "Dos" "Tres"	"Ponlo", "Colócalo" "Déjalo" "Muévelo"	"Zona"	"Uno", "Dos", "Tres, "Cuatro", "Cinco", "Seis", "Siete", "Ocho", "Nueve", Diez", "Once", "Doce"

Interacción multimodal

Se definió un solo comando multimodal compuesto de la oración "Este cubo ponlo aquí" y la ejecución coordinada del gesto "Seleccionar". Mientras se pronuncia "Este cubo" se ejecuta el gesto seleccionar por primera vez sobre el cubo que se desea mover. Y mientras se pronuncia "ponlo aquí" se ejecuta el gesto seleccionar sobre la zona a la que se desea mover.

doi: 10.25100/iyc.v25i4.13164

Pruebas del sistema

Se realizan dos tipos de evaluación para validar el sistema con la participación de siete usuarios. La primera evaluación cuantifica el porcentaje de reconocimiento de cada una de las formas de interacción (gestual, verbal, multimodal), y también la ejecución de la tarea completa de clasificación de cubos por color (tarea completa). La segunda es una evaluación cualitativa que, mediante encuestas, mide el nivel de satisfacción de los usuarios con las formas de interacción propuestas.

Descripción del sistema de interacción multimodal humano-robot

El sistema está constituido por cuatro módulos principales (Figura 3). La interacción con el usuario inicia en el módulo de entrada, que captura la información de comandos gestuales (video) y verbales (audio), y la fusiona. Dependiendo de los comandos, la máquina de estados gestiona la interacción humano-robot mediante una máquina de tres estados (reposo, aprendizaje y ejecución). La máquina de estados se apoya en las memorias semántica, episódica y procedural del módulo sistema de memorias, dependiendo del estado activo. El módulo de salida realimenta al usuario controlando las acciones del robot y emitiendo señales de audio y video.

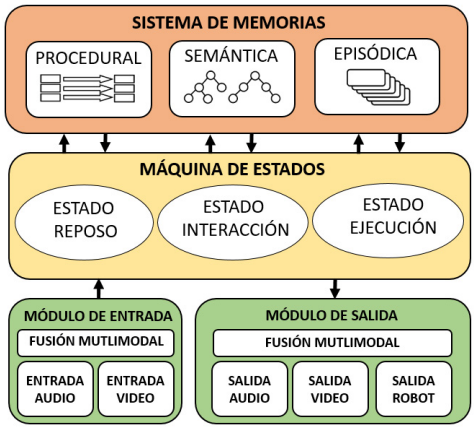

Figura 3. Diagrama de bloques del sistema.

El sistema fue implementado utilizando como IDE Visual Studio Code y lenguaje de programación Python 3.9.12, sobre un sistema operativo Windows 10 Pro de 64 bits. A continuación, se presenta una descripción detallada de cada uno de los módulos.

Módulo de Entrada

Compuesto por tres submódulos. El de entrada de audio que reconoce los comandos verbales; el de entrada de video que identifica el color, la ubicación de los cubos en la zona y el gesto "seleccionar"; mientras que, el de fusión multimodal prioriza e integra las señales verbales y gestuales.

Entrada de Audio: este submódulo reconoce diversas estructuras de comandos de voz combinando una de las palabras de las seis últimas columnas de la Tabla 1, para un total de 1152 estructuras diferentes. El audio captado se convierte en una cadena de texto, se

doi: 10.25100/iyc.v25i4.13164

separan las palabras y se comparan con las del diccionario (Tabla 1). Si en la estructura de comandos no se encuentra una componente de cada una de las seis categorías, el comando se considera inválido y el sistema le indicará al usuario la necesidad de repetirlo. Para la captura de audio, se utilizó una diadema inalámbrica Logitech G935, equipada con un micrófono de patrón de captación cardiode (unidireccional) y un ancho de banda entre 100 Hz y 10 kHz. Para el reconocimiento de los comandos de voz se utilizaron las librerías PyAudio 0.2.13 (24) y Speech Recognition 3.9.0 (25), que emplea el motor de reconocimiento de voz Google Speech Recognition (26).

Entrada de Video: a través de la entrada de video se reconocen la zona de trabajo, los cubos y el gesto "seleccionar". Para identificar la zona de trabajo y los cubos se utilizó la información de color en el espacio HSV. La zona de trabajo limitada con cinta roja (Figura 4.a), se detecta usando los umbrales definidos en la Ec. (1), dando como resultado la Figura 4.b.

$$\text{sí } ((0 \leq H \leq 5) o (175 \leq H \leq 179)) \& (100 \leq S \leq 255) \& (20 \leq V \leq 255) \rightarrow \in \text{color rojo} \quad \text{Ec. (1)}$$

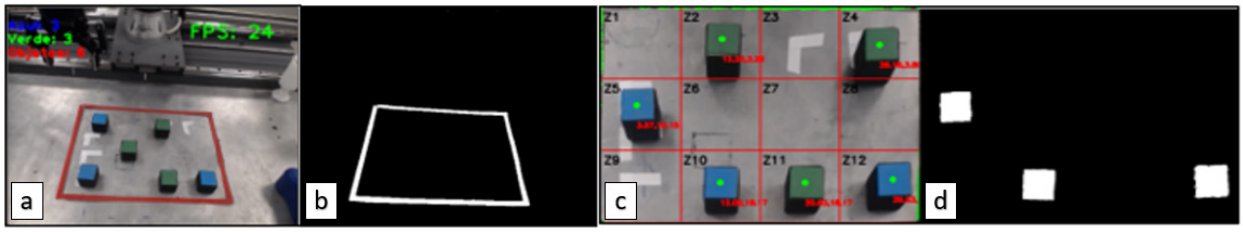

Figura 4. (a) Imagen capturada, (b) Delimitación zona de trabajo, (c) Transformación de perspectiva, (d) Filtrado cubos azules.

De forma análoga, los cubos de color azul y verde (Figura 4.c) se detectaron con los umbrales establecidos en las Ec. (2) y (3) respectivamente. La Figura 4.d es un ejemplo de la imagen de detección de los cubos azules.

$$\text{sí } (100 \leq H \leq 107) \& (150 \leq S \leq 255) \& (20 \leq V \leq 255) \rightarrow \text{color azul} \quad \text{Ec. (2)}$$

$$\text{sí } (47 \leq H \leq 99) \& (50 \leq S \leq 230) \& (20 \leq V \leq 230) \rightarrow \text{color verde} \quad \text{Ec. (3)}$$

Se aplicó una transformación de perspectiva a las imágenes para visualizarlas como si fueran captadas con la cámara perpendicular a la zona de trabajo. Lo anterior se logró aplicando la transformación dada por la matriz M definida por:

$$M = \begin{bmatrix} -2{,}6163 * 10^{0} & -1{,}8080 * 10^{-1} & 1{,}4692 * 10^{3} \\ -5{,}3953 * 10^{-1} & 2{,}5701 * 10^{0} & 5{,}5277 * 10^{1} \\ 2{,}0502 * 10^{-4} & 6{,}6274 * 10^{-4} & 1{,}0000 * 10^{0} \end{bmatrix}$$

La anterior matriz se obtuvo ubicando las esquinas de la imagen binarizada de la zona de trabajo (Figura 4.b), con el fin de transformar la zona de forma trapezoidal en un rectángulo. Adicionalmente, con base en el conocimiento de las dimensiones físicas de

doi: 10.25100/iyc.v25i4.13164

las retículas, a la zona transformada se le sobreponen líneas para localizar las 12 retículas y sus números de identificación (Figura 4.c). La imagen enriquecida es posteriormente presentada al usuario.

La ubicación de los cubos en cada zona se asume como el centroide de las regiones de color azul o verde detectadas. Esta ubicación se convierte de píxeles a centímetros con base en las Ec. (4).

$$x_{(cm)} = \frac{x_{(px)} * 21.57\ (cm)}{640\ (px)} \qquad y_{(cm)} = \frac{y_{(px)} * 19.87\ (cm)}{480\ (px)} \qquad \text{Ec. (4)}$$

El gesto "seleccionar" requiere la identificación de la posición relativa entre los dedos índice y pulgar, es decir, si están unidas o separadas las falanges distales. Para esta identificación, se parte de la esqueletización de la mano del usuario, que retorna un conjunto de 20 puntos (los puntos 4 y 8 corresponden a la punta de los dedos pulgar e índice, respectivamente). De este modo, si la distancia relativa entre ellos es mayor a 30 pixeles, se consideran separadas ("gesto abierto"), ver Figura 5.a, o, en caso contrario como "cerrado" (Figura 5.c). Para propósitos de realimentación al usuario, se sobrepone sobre la imagen una línea magenta entre las puntas de ambos dedos, cuyo punto medio se utiliza para trazar un círculo verde (Figuras 5.b y 5.d) que se denominó "cursor digital", y sirve para indicar el cubo o la zona de interés.

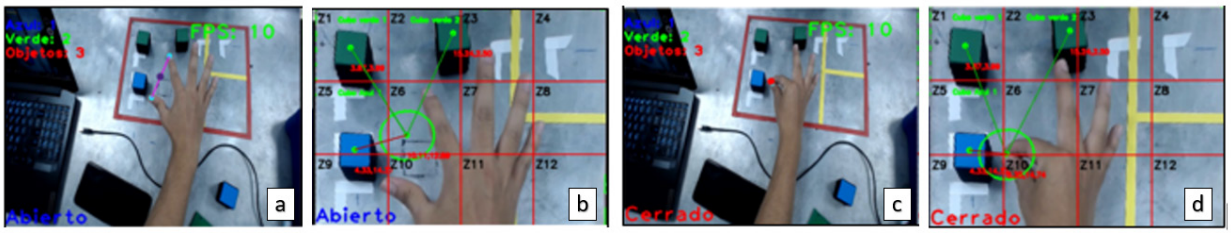

Figura 5. (a) Gesto abierto, (b) Cursor digital abierto, (c) Gesto cerrado, (d) Cursor digital cerrado.

Para la implementación de este submódulo se empleó la cámara Logitech C920 y las funciones proporcionadas por la librería OpenCV-Python 4.7.0.68 (27), en particular, la librería Mediapipe Hands (28) para la esqueletización de la mano.

Fusión multimodal: este submódulo tiene dos funciones. Por un lado, identifica cuando se está realizando una interacción multimodal; y por otro, prioriza e integra las interacciones verbales y gestuales que provienen de los submódulos de Entrada de Audio y Entrada de Video.

Se implementó una fusión jerárquica (29) en la cual prima el audio sobre el video, para el comando multimodal; "Mira este cubo (gesto seleccionar sobre un cubo) ponlo aquí (gesto seleccionar sobre una zona de la retícula)". Este comando multimodal permite realizar la selección y desplazamiento de un cubo hacia una zona específica. Se realizó una implementación por hilos para que los submódulos de entrada siempre se estén ejecutando al mismo tiempo, permitiendo recolectar y actualizar la información visual que se comparte a los demás módulos de manera constante, y al mismo tiempo mostrar al usuario la realimentación visual sin interrupción. Finalmente, la fusión multimodal

doi: 10.25100/iyc.v25i4.13164

evalúa en paralelo la interacción verbal y gestual, pero le da prioridad a la interacción verbal, de manera que solo al identificar el comando; "Mira, este cubo, ponlo aquí", es cuando se extrae la información visual correspondiente al cubo y zona de trabajo indicados por el gesto "seleccionar" completando el comando multimodal.

Para la ejecución simultanea de audio y video que requiere este submódulo, se realizó una implementación paralela en hilos mediante la librería threading de Python 3.9.12.

Máquina de estados

El diseño de la máquina de estados tomó como referencia el trabajo propuesto por (5), donde se propone una interacción basada en tres estados: reposo, aprendizaje y ejecución. Estos estados controlan el robot cuando no hay interacción y direccionan la interacción para aprender una nueva tarea o ejecutar las tareas aprendidas. La máquina de estados se implementó creando una función independiente en Python 3.9.12 para cada uno de los estados. La comunicación y sincronización de los tres estados se realizó mediante variables globales (flags). A continuación, se presenta una explicación de cada estado para el sistema propuesto.

Estado de reposo: en este estado el sistema se encuentra constantemente en escucha a la espera de la señal de activación ("hola"). Reconocida la palabra, se reproduce sintéticamente la frase "hola, ¿deseas que aprenda o ejecute una tarea?". El usuario debe responder con una de dos palabras, "Aprender" o "Ejecutar", que activan el estado correspondiente.

Estado de aprendizaje: para cada cubo presente en la zona de trabajo, debe establecerse un comando (verbal o multimodal) a través del submódulo de fusión multimodal. Una vez realizado el desplazamiento de cada uno de los cubos, el sistema identifica las condiciones de la zona de trabajo (cantidad, color y zona en la que están ubicados los cubos al inicio y al final del aprendizaje), para almacenar esta información en la memoria semántica, y la solución en la memoria episódica. Realizado esto, se realimenta auditivamente al usuario con la frase "Tarea aprendida", y el sistema vuelve al estado de reposo.

Estado de ejecución: el sistema solicita a la memoria semántica la cantidad, color y las retículas en las que están ubicados los cubos. Seguido a esto, solicita a la memoria episódica una solución aleatoria válida para la cantidad de cubos identificada. Posteriormente, la memoria procedural verifica que las retículas en las que deben ubicarse los cubos que estén libres. Cuando una zona está ocupada, la memoria procedural adiciona un paso a la secuencia de la solución provista por la memoria episódica, para desplazar el cubo a una zona libre y que no vaya a ser usada por otro cubo en la solución. Finalmente, se envía al módulo de salida los movimientos que deben realizarse de manera secuencial.

doi: 10.25100/iyc.v25i4.13164

Sistema de Memorias

El diseño del sistema de memorias propuesto está basado en la arquitectura cognitiva SOAR (16,18), la cual propone un sistema de memorias de largo plazo constituido por memorias semántica, procedural y episódica. Estas memorias se encargan de identificar escenas, almacenar procedimientos paso a paso y recordar experiencias pasadas para tomar acción el escenario actual. A continuación, se presenta una explicación de la función que cumple cada memoria para el sistema propuesto:

Memoria semántica: es la memoria empleada en los estados de aprendizaje y ejecución para interpretar y organizar el espacio de trabajo tomando como base la información multimodal proveniente del administrador del dialogo y arroja como salida un vector de posiciones que contiene la cantidad de cubos identificados, color y posición.

Memoria episódica: esta memoria debe contener al menos una solución para las posibles combinaciones del número de cubos por colores en el espacio de trabajo (9 combinaciones en total), las cuales se almacenan en un archivo de texto plano (.csv). Posterior al estado de aprendizaje se adiciona una nueva solución al archivo, si esta no existe previamente. Durante el estado de ejecución, se selecciona aleatoriamente una solución acorde con la cantidad y color de los cubos presentes en el espacio de trabajo, y la envía al administrador del dialogo.

Memoria procedural: esta memoria es utilizada sólo en el estado de ejecución para establecer la secuencia de movimientos que debe realizar el robot para clasificar por colores los cubos presentes en el área de trabajo. Una vez seleccionada una solución de la memoria episódica, la memoria procedural debe ajustar la secuencia de movimientos para ejecutar de manera correcta esta solución. Para ello, se desarrolló una funcionalidad que valida que las zonas en las que se van a ubicar los cubos se encuentren libres.
En caso de encontrar cubos en las zonas correspondientes a las posiciones finales de la solución, esta memoria adiciona movimientos que garanticen que previamente se desplaza el cubo que ocupa la zona del siguiente movimiento, en caso de ser necesario. La secuencia de movimientos a realizarse se envía al módulo de salida.

Durante el estado de aprendizaje, esta memoria se encarga de procesar las instrucciones dadas por el usuario en cada demostración y dar las órdenes de movimiento correspondientes al robot, para que este pueda desplazar los cubos siguiendo las indicaciones del usuario.

La implementación de los sistemas de memoria se basó en la lectura y escritura de archivos de texto plano mediante la librería csv de Python 3.9.12, con el fin de guardar y consultar las soluciones aprendidas durante la interacción multimodal.

Módulo de salida

Este módulo es el encargado de ejecutar las acciones en el robot y realimentar al usuario. Está compuesto por cuatro submódulos, que se describen a continuación.

doi: 10.25100/iyc.v25i4.13164

Fisión Multimodal: es el encargado de distribuir las órdenes de la máquina de estados a los submódulos de salida correspondientes (audio, visual y robot).

Salida de audio: realimenta auditivamente al usuario, a través de un diccionario de síntesis de voz con 15 expresiones (Tabla 2), este diccionario genera una interactividad verbal compuesta por las respuestas verbales del robot, ya que van guiando la interacción mediante múltiples saludos, frases de selección de tarea (aprender o ejecutar), frases de afirmación durante y al finalizar el aprendizaje. Para su implementación se utilizó la librería de conversión de texto a voz Pyttsx3 2.90 (30), con intérprete en español.

Tabla 2. Diccionario de síntesis de voz

Función	Síntesis de Voz
	"¿Quieres que aprenda una tarea o que la ejecute?"
	"Error, comando de voz no identificado"
	"Hola", "Hola como estás", "Hola, que tal", "Es bueno volver a verte", "Regresaste", "¿Qué tal?"
	"La tarea se ejecutó con éxito"
	"Enséñame"
	"Tarea aprendida"
	"La tarea no fue reconocida, por favor repita nuevamente"
	"Siguiente instrucción"
	"Error, comando de voz no identificado"

Salida de video: brinda al usuario una realimentación gráfica en la pantalla del equipo de cómputo utilizado. Mediante la visualización de la imagen enriquecida que incluye líneas rojas para las retículas, numerales identificadores de las retículas, el centro de cada cubo, el cursor digital y la leyenda (abierto/cerrado) según la posición relativa de los dedos pulgar-índice (ver Figura 5.b y 5.d).

Salida del Robot: convierte la secuencia de movimientos provista por la memoria procedural, en instrucciones para el desplazamiento y accionamiento de la pinza del robot colaborativo UR3. Este submódulo se apoyó en (31,32) y utilizó la librería comunicacionUR3.py para activar las funciones Gripper.activate(), Gripper.close(), Gripper.half(), Gripper.open() que establecen comunicación con el robot UR3 y accionan la pinza. Adicionalmente, utilizó la librería controlUR3.py para el desplazamiento del brazo robótico con la función move (). Se utilizó una comunicación TCP/IP en una red LAN entre el robot y el equipo de cómputo que corre los algoritmos, estableciendo una comunicación cliente/servidor tipo socket con IP y puerto estáticos. Los algoritmos de este módulo pueden ser consultados en (33).

doi: 10.25100/iyc.v25i4.13164

Resultados

Se realizaron pruebas con un grupo de siete personas, con conocimientos previos y experiencia en el uso de robots, cuyas edades oscilaban entre los 22 y los 25 años. El desempeño del sistema fue evaluado mediante un protocolo de pruebas que incluyó evaluaciones cualitativas y cuantitativas.

Pruebas con evaluación cuantitativa

Se evaluó de manera aislada el desempeño en el reconocimiento de los comandos verbales, gestual y multimodal. Finalmente, la ejecución de la tarea de clasificación de cubos por color.

Reconocimiento de comandos verbales: cada uno de los siete usuarios pronunció tres veces cada una de las 28 palabras del diccionario (para un total de 7x28x3=588). La Figura 6 presenta los porcentajes de reconocimiento por palabra, donde se observa que en 16 de las 28 palabras se alcanzó un reconocimiento del 100 %, mientras que el menor porcentaje fue para la palabra "coge" con un 76.19%. Se alcanzó un porcentaje promedio de reconocimiento del 94.56% ± 7.41%.

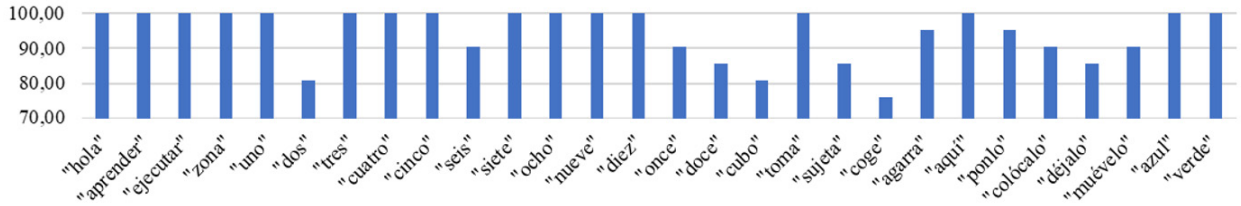

Figura 6. Resultados del reconocimiento del diccionario de palabras.

Posteriormente, cada usuario pronunció tres veces, cada uno de los siguientes comandos compuestos: a) "Toma el cubo azul 1 y ponlo en la zona 3", b) "Coge el cubo verde 2 y colócalo en la zona 12" y c) "Sujeta el cubo verde 3 y déjalo en la zona 7". Para las frases a y c se alcanzó un porcentaje de reconocimiento del 100%; mientras que para la frase b, se obtuvo un 95.2%. Este último resultado se debe a la dicción de la palabra "coge" por el usuario 5 presente en la frase b. En consecuencia, el porcentaje de reconocimiento promedio de las frases fue del 98.41% ± 4.19%.

Reconocimiento gestual: cada uno de los siete usuarios realizó el gesto "Seleccionar" tres veces, en cada una de las 12 retículas del espacio de trabajo. La tabla 3 presenta los porcentajes de reconocimiento del gesto "seleccionar" en cada celda de la zona de trabajo. Se obtuvo un promedio de reconocimiento del 81.35% ± 7.98% para las 252 (7x3x12) repeticiones del gesto. Se observa que el desempeño es mayor en las retículas centrales y cercanas a la mano del usuario (Z6, Z7, Z9 a Z12); mientras que el menor porcentaje de reconocimiento se presentó en las retículas Z5 y Z8; en tanto que las retículas más cercanas a la cámara (Z1 a Z4) presentan porcentajes de reconocimiento similar entre ellas (76.19%).

doi: 10.25100/iyc.v25i4.13164

Tabla 3. Desempeños de reconocimiento del gesto "seleccionar"

ZONA	U1 (%)	U2 (%)	U3 (%)	U4 (%)	U5 (%)	U6 (%)	U7 (%)	Promedio (%)
Z1	66,67	100	100	100	66,67	66,67	33,33	**76,19**
Z2	66,67	100	100	66,67	66,67	66,67	66,67	**76,19**
Z3	66,67	66,67	100	100	66,67	66,67	66,67	**76,19**
Z4	33,33	66,67	66,67	66,67	100	100	100	**76,19**
Z5	66,67	66,67	66,67	100	100	66,67	33,33	**71,43**
Z6	66,67	100	100	66,67	100	66,67	100	**85,72**
Z7	100	100	100	100	66,67	100	66,67	**90,48**
Z8	66,67	100	66,67	66,67	66,67	66,67	66,67	**71,43**
Z9	66,67	100	66,67	66,67	100	100	66,67	**80,95**
Z10	100	100	100	100	66,67	100	66,67	**90,48**
Z11	66,67	100	100	100	100	100	100	**95,24**
Z12	100	100	66,67	66,67	100	100	66,67	**85,72**
Promedio	**72,22**	**91,67**	**86,11**	**83,34**	**83,34**	**83,34**	**69,45**	**81,35**

Para los Usuarios 1 y 7 que presentaron el menor desempeño con (72.22% ± 19.24%) y (69.45% ± 22.28%) respectivamente. Los errores en el reconocimiento del gesto "seleccionar" ocurrieron en casos donde el usuario acercaba el dedo índice hacia el pulgar, en lugar de mover los dos dedos simultáneamente. Esto puede ocasionar un error en la ubicación del cursor del cubo o de la zona planeados por el usuario.

Reconocimiento multimodal: cada uno de los siete usuarios repitió tres veces la interacción multimodal compuesta por el gesto "Seleccionar" combinado con una de las tres frases indicadas en la Tabla 4. Las frases presentan una estructura similar donde solo se cambian los verbos asociados a las categorías 'Tomar' y 'Dejar' (ver Tabla 1). En las pruebas se utilizaron las retículas (Z1, Z5 y Z9) como ubicación final del cubo, ya que presentaron el menor desempeño en las pruebas gestuales. Se obtuvo un reconocimiento promedio general del 80.95% ± 12.36% para las 63 (7x3x3) interacciones. El desempeño por frases presenta una correspondencia con el desempeño de reconocimiento de cada verbo, por lo que la mejor combinación es 'Agarra'+'Colócalo' (oración 2); seguida por 'Sujeta'+ 'Ponlo' (oración 1). El menor desempeño lo obtuvo la oración 3 con la combinación 'Coge'+ 'Muévelo'. También se observa una correspondencia entre el desempeño del reconocimiento del gesto "Seleccionar" y el reconocimiento del comando multimodal, donde los usuarios 2, 3 y 5 alcanzaron los mayores desempeños, al contrario de los usuarios 1 y 7 con los menores.

doi: 10.25100/iyc.v25i4.13164

Tabla 4. Desempeños de reconocimiento del comando multimodal.

USUARIO	Comando Multimodal (%)			Promedio (%)
	"Sujeta este cubo (gesto de seleccionar sobre cubo azul uno) y ponlo aquí (gesto de seleccionar sobre la zona cinco)"	"Agarra este cubo (gesto de seleccionar sobre cubo azul tres) y colócalo aquí (gesto de seleccionar sobre la zona nueve)"	"Coge este cubo (gesto de seleccionar sobre cubo azul dos) y muévelo aquí (gesto de seleccionar sobre la zona uno)"	
U1	66,67	100,00	66,67	**77,78**
U2	100,00	100,00	66,67	**88,89**
U3	100,00	100,00	66,67	**88,89**
U4	66,67	100,00	66,67	**77,78**
U5	100,00	100,00	66,67	**88,89**
U6	100,00	100,00	66,67	**88,89**
U7	66,67	66,67	33,33	**55,55**
Promedio	**85,71**	**95,24**	**61,90**	**80,95**

Ejecución de la tarea de clasificación de cubos por color: para la evaluación de la clasificación de los cubos, cada usuario realiza los siguientes dos pasos: a) activar el estado de aprendizaje y enseñarle desde cero una tarea por cada una de las posibles combinaciones que pueden tener los cubos en el espacio de trabajo, b) ordenar al robot que ejecute una tarea con cada una de las posibles 63 combinaciones de cantidad y color de los cubos, pero variando las posiciones iniciales de dichos cubos. Los resultados de esta prueba en las que cada usuario ejecutó el ejercicio una sola vez por cada una de las nueve posibles combinaciones. El sistema ejecutó correctamente la clasificación de todos los cubos, obteniendo un desempeño del 100%.

Discusión

Los resultados indican que el sistema tiene un alto desempeño en el reconocimiento de comandos verbales (98.41% ± 4.19%). Este modo de interacción depende fuertemente de la correcta dicción del usuario, mientras que el porcentaje de reconocimiento del gesto "seleccionar" fue menor (81.35% ± 7.98%) observando que es muy dependiente de la ubicación de la zona en la retícula, ya que las zonas con menor desempeño corresponden a aquellas en donde la transformación de perspectiva modificó mayormente la posición de los pixeles. De otra parte, se observa que el porcentaje de reconocimiento mejora cuando el usuario realiza el gesto cerrando los dedos índice y pulgar de forma simultánea, evitando un desplazamiento del cursor. En la interacción

doi: 10.25100/iyc.v25i4.13164

multimodal se obtuvo un reconocimiento promedio del 80.95% ± 12.36%; siendo menor que las dos pruebas anteriores debido a que un error en el reconocimiento de cualquiera de los comandos (verbal o gestual) ocasiona un fallo en el reconocimiento multimodal. El sistema ejecutó correctamente la clasificación de todos los cubos, obteniendo un desempeño del 100%, independientemente de las posiciones y orientaciones iniciales de los cubos dentro de las retículas elegidas por el usuario.

Evaluación cualitativa de las pruebas

Finalizadas las pruebas, los usuarios diligenciaron una encuesta sobre el sistema de interacción multimodal (Tabla 3), cuyas opciones de respuesta se ajustan a la escala Likert de cinco niveles (34).

Tabla 5. Afirmaciones para evaluación cualitativa

Pregunta	Afirmación
1	El sistema multimodal mejora la interacción de usuario comparado con una interacción kinestésica o tele operada.
2	Los comandos de voz fueron identificados correctamente por el sistema multimodal.
3	El sistema multimodal identifica correctamente los comandos gestuales.
4	El sistema multimodal identifica correctamente los comandos multimodales.
5	La realimentación gráfica/auditiva contribuye a tener una mejor experiencia con el sistema multimodal

La Figura 7 resume los resultados de la encuesta de satisfacción de los usuarios a las afirmaciones de la Tabla 5, donde se utilizan barras de colores para ilustrar la elección ante cada opción de respuesta.

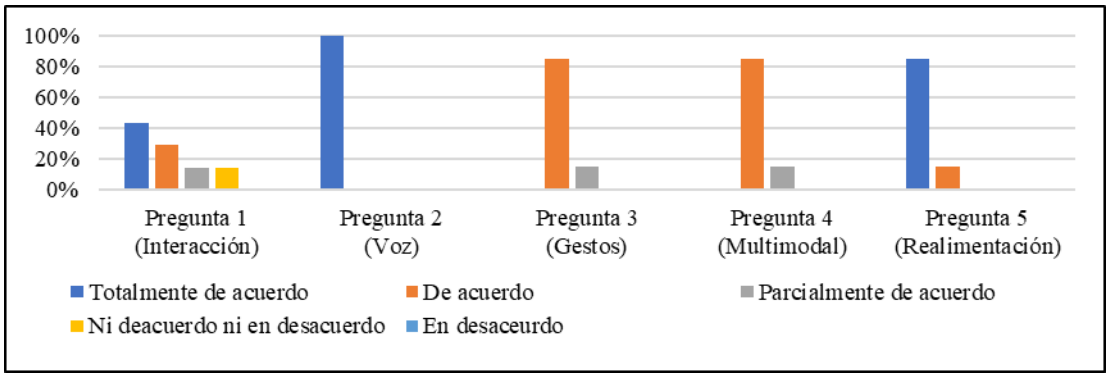

Figura 7. Respuestas de satisfacción de los usuarios.

doi: 10.25100/iyc.v25i4.13164

Se observa que la mayoría de los usuarios eligieron las opciones Totalmente de acuerdo (azul) y De acuerdo (Naranja) para todas las afirmaciones. Las interacciones con mejor satisfacción por parte de los usuarios fueron la interacción verbal y la realimentación entregada por el sistema. Le siguieron las interacciones gestual y multimodal. La satisfacción respecto a la mejoría de la experiencia del sistema multimodal frente a la interacción kinestésica o tele operada presentó mayor variabilidad en las respuestas de los usuarios. Esto puede ser debido a los errores del sistema en la identificación del gesto "Seleccionar" y a que el usuario debe girar su cabeza para observar en la pantalla del computador las retículas, pues no se visualizan físicamente en la zona de trabajo.

Conclusiones

Se implementó un sistema para enseñarle a un robot una tarea de clasificación de objetos cúbicos que debe ejecutar, mediante una interfaz multimodal que permite que un usuario le indique las acciones a realizar mediante comandos gestuales y de voz. A diferencia de los sistemas de interacción humano-robot tradicionales, este desarrollo permite el aprendizaje en línea de una tarea de clasificación a partir de instrucciones audiovisuales impartidas en un ambiente de experimentación real. La interacción multimodal permitió que la tarea de clasificación sea enseñada a un robot de manera natural (gesto y voz) por una persona sin conocimientos de programación, reduciendo los tiempos de ajustes y de reprogramación que normalmente son realizadas por un ingeniero experimentado en robótica.

El sistema fue diseñado para interpretar un comando multimodal compuesto de un gesto (seleccionar) y la composición de frases a partir de un diccionario de 27 palabras. También interpreta comandos solamente verbales.

La evaluación del sistema fue realizada por 7 usuarios mediante pruebas cuantitativas y cualitativas. En la prueba cuantifica se evaluó cada modo de interacción (gestual, verbal, multimodal) en el estado de aprendizaje, y posteriormente la clasificación de cubos por color en el estado de ejecución. El porcentaje promedio de reconocimiento de las interacciones fue de 81.35% ± 7.98% para la gestual, de 98.41% ± 4.19% para la verbal, y de 80.95% ± 12.36% para multimodal. Las pruebas cualitativas arrojaron resultados consistentes con los porcentajes de reconocimiento, puesto que valoraron más la interacción verbal sobre la multimodal.

El desempeño del reconocimiento de la interacción gestual se ve afectada por la manera de abrir y cerrar los dedos de cada usuario, siendo mejor cuando se desplazan simultáneamente los dedos pulgar e índice. El reconocimiento de la interacción verbal se ve influenciada por la dicción del usuario y las palabras seleccionadas para componer las frases. Las limitaciones anteriores se ven reflejadas en el reconocimiento de la interacción multimodal sumado a los errores de la estimación de la posición inicial y final de los cubos.

Para mejorar el desempeño del sistema se podría limitar el diccionario, escogiendo un conjunto de palabras con mayor facilidad de dicción por parte de los usuarios y así disminuir los errores en reconocimiento verbal. De igual manera, se podría dibujar la retícula en la zona de trabajo para que los usuarios tenga una mejor visualización de la ubicación sobre la cual realiza el gesto de 'seleccionar' y disminuir las repeticiones durante el estado de aprendizaje. También se podría incluir una etapa de calibración de cámara para disminuir los errores de la estimación de la posición inicial y final de los cubos, causada

doi: 10.25100/iyc.v25i4.13164

por la transformación de perspectiva. Finalmente, se propone evaluar la adaptabilidad del sistema propuesto en tareas robóticas más complejas como el ensamble de piezas.

Agradecimientos

Este trabajo fue financiado con recursos del proyecto de investigación interna Aproximación a una arquitectura cognitiva para el aprendizaje y generalización de un proceso de clasificación aplicado en un robot colaborativo el cual pertenece a la convocatoria interna de proyectos: Por una universidad transformadora: Horizonte 2021-2025 de la Pontificia Universidad Javeriana Cali.

Referencias bibliográficas

1. Billard A, Ravichandar H, Polydoros AS, Chernova S. Recent Advances in Robot Learning from Demonstration. Annu Rev Control Robot Auton Syst. 2020;3(1):297–330.

2. Drolshagen S, Pfingsthorn M, Gliesche P, Hein A. Acceptance of Industrial Collaborative Robots by People With Disabilities in Sheltered Workshops. 2021;7(January).

3. Haage M, Piperagkas G, Papadopoulos C, Mariolis I, Malec J, Bekiroglu Y, et al. Teaching Assembly by Demonstration Using Advanced Human Robot Interaction and a Knowledge Integration Framework. Procedia Manuf. 2017;11(June):164–73.

4. So W, Wong MK, Lam CK, Lam W, Chui AT, Lee T, et al. Using a social robot to teach gestural recognition and production in children with autism spectrum disorders. Disability and Rehabilitation: Assistive Technology. 2017;

5. Lázaro-Gredilla M, Lin D, Swaroop Guntupalli J, George D. Beyond imitation: Zero-shot task transfer on robots by learning concepts as cognitive programs. Sci Robot. 2019;4(26):1–16.

6. Mukherjee D, Gupta K, Chang LH, Najjaran H. A Survey of Robot Learning Strategies for Human-Robot Collaboration in Industrial Settings. Robot Comput Integr Manuf [Internet]. 2022;73(July 2021):102231. Available from: https://doi.org/10.1016/j.rcim.2021.102231

7. Li S, Zheng P, Fan J, Wang L. Toward Proactive Human – Robot Collaborative Assembly : A Multimodal Transfer-Learning-Enabled Action Prediction Approach. 2022;69(8):8579–88.

8. Mosquera-DeLaCruz J-H, Nope-Rodríguez S-E, Restrepo-Girón A-D, Martínez-Álvarez A, Loaiza-Correa H. Disability and Rehabilitation : Assistive Technology Human-computer multimodal interface to internet navigation. Disabil Rehabil Assist Technol. 2020;0(0):1–14, https://doi.org/10.1080/17483107.2020.179944.

9. Kotseruba I, Tsotsos JK. 40 years of cognitive architectures : core cognitive abilities and practical applications [Internet]. Vol. 53, Artificial Intelligence Review. Springer Netherlands; 2018. 17–94 p. Available from: https://doi.org/10.1007/s10462-018-

9646-y

10. Das N, Prakash R, Behera L. Learning object manipulation from demonstration through vision for the 7-DOF barrett WAM. 2016 IEEE 1st Int Conf Control Meas Instrumentation, C 2016. 2016;(Cmi):391–6.

11. Du G, Chen M, Liu C, Zhang B, Zhang P. Online robot teaching with natural human-robot interaction. IEEE Trans Ind Electron. 2018;65(12):9571–81.

12. Argall BD, Chernova S, Veloso M, Browning B. A survey of robot learning from demonstration. Rob Auton Syst. 2009;57(5):469–83.

13. Hausman K, Chebotar Y, Schaal S, Sukhatme G, Lim JJ. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. Adv Neural Inf Process Syst. 2017;2017-Decem:1236–46.

14. Gonzalez-Fierro M, Balaguer C, Swann N, Nanayakkara T. A humanoid robot standing up through learning from demonstration using a multimodal reward function. In: 2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids). IEEE; 2013. p. 74–9.

15. Mayer RE. Thirty years of research on online learning. 2019;(October 2018):152–9.

16. Laird JE, Lebiere C, Rosenbloom PS. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. AI Mag. 2017;38(4):13–26.

17. Choi D, Langley P. Evolution of the ICARUS Cognitive Architecture. Cogn Syst Res [Internet]. 2018;48:25–38. Available from: https://doi.org/10.1016/j.cogsys.2017.05.005

18. Laird JE. The SOAR of cognitive architecture. Proceedings of the 2013 International Conference on Current Trends in Information Technology, CTIT 2013. 2013. 135–142 p.

19. Abbasi B, Monaikul N, Rysbek Z, Eugenio B Di. A Multimodal Human-Robot Interaction Manager for Assistive Robots. 2019;6756–62.

20. Chen L, Javaid M, Eugenio B Di. The roles and recognition of Haptic-Ostensive actions in collaborative multimodal human – human dialogues &. 2015;34:201–31.

21. Monaikul N, Abbasi B, Rysbek Z, Eugenio B Di. Role Switching in Task-Oriented Multimodal Human-Robot Collaboration. 2020;1150–6.

22. Male J, Martinez-hernandez U. Collaborative architecture for human-robot assembly tasks using multimodal sensors *. 2021;1024–9.

23. Billard AG, Calinon S, Dillmann R. Learning from Humans. Springer Handb Robot. 2016;Pages 1995-2014.

doi: 10.25100/iyc.v25i4.13164

24. Pypi.org. PyAudio 0.2.13 [Internet]. 2022 [cited 2023 Jan 18]. Available from: https://pypi.org/project/PyAudio/

25. Pypi.org. Python Speech Recognition 3.9.0 [Internet]. 2022 [cited 2023 Jan 18]. Available from: https://pypi.org/project/SpeechRecognition/

26. Google LLC. Language model selection for speech-to-text conversion [Internet]. 2023 [cited 2023 Mar 29]. Available from: https://patents.google.com/patent/US9495127B2/en

27. Pypi.org. OpenCV-Python 4.7.0.68 [Internet]. 2022 [cited 2023 Jan 18]. Available from: https://pypi.org/project/opencv-python/

28. Google LLC. Mediapipe Hands [Internet]. 2022 [cited 2023 Jan 18]. Available from: https://google.github.io/mediapipe/solutions/hands

29. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Syst [Internet]. 2018;161:124–33. Available from: https://doi.org/10.1016/j.knosys.2018.07.041

30. Pypi.org. Pyttsx3 2.90 [Internet]. 2022 [cited 2023 Jan 18]. Available from: https://pypi.org/project/pyttsx3/

31. Blandon JS. Interfaz de voz humano-robot para controlar un brazo robótico UR3. Trabajo de Grado en Ingeniería Electrónica, Pontificia Universidad Javeriana Cali; 2021.

32. Holguin JD. Algoritmo de fusión de señales de audio y vídeo para el manejo de un UR3. Trabajo de Grado en Ingeniería Electrónica, Pontificia Universidad Javeriana Cali; 2021.

33. Mosquera-DeLaCruz J-H, Martínez-Álvarez A, Nope-Rodríguez S-E, Loaiza-Correa H, Rodríguez-Téllez G-A, Jamioy-Cabrera J-D, et al. UR3 Multimodal Interaction Color Classification [Internet]. 2023 [cited 2023 Aug 10]. Available from: https://github.com/nandostiwar/UR3_Multimodal_Interaction_Color_Classification

34. SimplyPsicology. Likert Scale [Internet]. 2023 [cited 2023 Aug 10]. Available from: www.simplypsychology.org/likert-scale.html

doi: 10.25100/iyc.v25i4.13164