Arquitectura cognitiva aplicada en procesos de clasificación para un robot colaborativo

Juan David Jamioy-Cabrera
Gabriel Alejandro Rodríguez-Téllez

***Resumen*— Se desarrollaron algoritmos que permiten interacciones gestuales, así como interacciones por voz y se implementaron a un modelo de arquitectura cognitiva SOAR la cual permite a los usuarios enseñarle y ejecutar tareas de clasificación de objetos cúbicos por color al robot colaborativo UR3 mediante interacción multimodal comandada por gestos y voz.**

Las pruebas cualitativas se realizaron mediante una encuesta de cinco preguntas evaluada con una escala Likert, consultando a los usuarios sobre su experiencia respecto al desempeño de las interacciones verbales, gestuales, multimodales, realimentación y clasificación. Las encuestas mostraron una alta satisfacción sobre la arquitectura propuesta durante la interacción del usuario con el robot.

I. INTRODUCCIÓN

Para abordar la problemática planteada en este trabajo investigativo, se llevó a cabo una investigación exhaustiva sobre el desarrollo de una arquitectura cognitiva que permita al robot colaborativo UR3 aprender la tarea de clasificación de objetos cúbicos por color. Sin embargo, para lograr una interacción efectiva entre el humano y el robot, se optó por utilizar una comunicación humano-máquina multimodal.

La arquitectura cognitiva multimodal desarrollada en este estudio permite a los usuarios establecer una comunicación fluida con el robot colaborativo UR3. Mediante la combinación de diferentes modalidades de entrada y salida, como voz, gestos, expresiones faciales y retroalimentación táctil, se logra una interacción más natural y eficiente. Esto reduce la experiencia requerida por parte del operario para enseñar la tarea de clasificación de objetos cúbicos por color al robot UR3.

En resumen, la investigación se centró en desarrollar una arquitectura cognitiva multimodal que facilitara la comunicación humano-robot en el contexto del aprendizaje de una tarea de clasificación de objetos cúbicos por color. La incorporación de la comunicación humano-máquina multimodal en esta arquitectura brinda una forma más intuitiva y efectiva de interactuar con el robot, mejorando así la experiencia y reduciendo la necesidad de experiencia por parte del operario.

II. FUNDAMENTACIÓN TEÓRICA

Los robots colaborativos se diseñan para trabajar junto a operarios humanos en tareas repetitivas o riesgosas. Gracias a su precio, adaptabilidad y elementos plug and play como cámaras y sensores, las pequeñas y medianas empresas comienzan a optar por esta tecnología [1]. Los entornos de trabajo colaborativos entre humanos y robots requieren métodos más eficientes de aprendizaje y comunicación [2]. Las arquitecturas cognitivas (AC) son una línea de investigación en inteligencia artificial que busca razonar, desarrollar conocimientos y adaptarse a nuevas situaciones [3], [4]. Hay diferentes arquitecturas basadas en el aprendizaje multimedia, procesos biológicos y distribución de memorias humanas [5], [6], [7]. La implementación de arquitecturas cognitivas en robots colaborativos simplifica la interacción y avanza hacia una interacción natural mediante gestos y comandos de voz [8], [9].

En la tabla 1 se presenta una comparativa con otros desarrollos de aplicaciones robóticas reales o simuladas, mediante interacciones comandadas por gestos, voz o sensores, las cuales utilizan arquitecturas cognitivas como ICARUS, SOAR o ACT-R y tienen una realimentación auditiva o visual.

Tabla 1. Tabla comparativa de desarrollos multimodales con arquitecturas cognitivas

.

Referencia	Aplicación robótica		Tipo de interacción			Arquitectura cognitiva	Realimentación	
	Real	Simulada	Voz	Gestos	Sensores		Audio	Visual
[10]	x			x	x	SOAR		x
[11]	x			x		SOAR		x
[13]		x			x	SOAR		x
[14]	x			x	x	ACT-R		x
[15]	x		x	x		ACT-R		x
[16]	x		x			ICARUS	x	
Arquitectura propuesto	**x**		**x**	**x**		**SOAR**	**x**	**x**

La revisión bibliográfica realizada muestra que la mayoría de sistemas fueron probados sobre aplicaciones robóticas reales, utilizando gestos como principal medio de interacción y arquitectura cognitiva SOAR. En los tipos de interacción se presentan casos de multimodalidad basada en gestos y sensores como [10] y [12], mientras que en [13] se presenta un sistema comandado por gestos y voz como el propuesto en este trabajo.

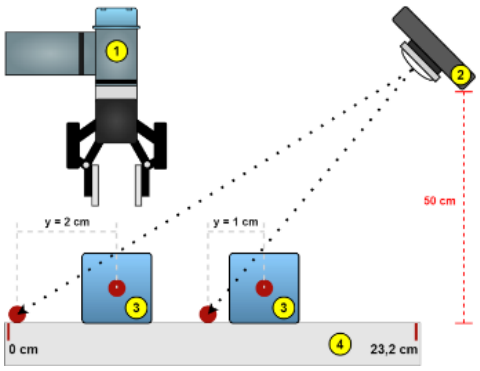

Figura 1.esquema del espacio de trabajo

Finalmente el desarrollo del proyecto propuesto se realizó en una locación la cual tenia un ruido lumínico y de sonido tipo oficina, y se utilizaron una camara y diadema de la marca logitech para realizar la captura de audio y video, en la Figura 1 se observa un esquema del espacio de trabajo. Cabe resaltar que la cámara se encuentra levemente inclinada para lograr una mejor visión del espacio de trabajo.

III. PRUEBAS Y RESULTADOS

Se realizaron pruebas con un grupo de siete personas, cuyas edades oscilaban entre los 22 y los 25 años. Todos los participantes contaban con conocimientos previos en computación y experiencia en robótica. A Continuación los resultados desempeño del sistema así como la realimentación auditiva y visual a las diferentes pruebas realizadas.

A. Pruebas cuantitativas

Las pruebas cuantitativas se dividieron en: diccionario de expresiones verbales, comandos de voz, gestos en zonas de interacción, comandos multimodales y desempeño de la arquitectura, los cuales son descritos a continuación:

1) Reconocimiento de Componentes verbales del Diccionario

Cada uno de los siete usuarios pronunció tres veces cada una de las 28 palabras del diccionario; por lo tanto, cada expresión fue evaluada 21 veces. Se obtuvo un desempeño global promedio del 94.56% con una desviación estándar del 7.41, siendo los usuarios 3 y 5 con quienes el sistema presentó el mayor y el menor desempeño en el reconocimiento con (96.43% ± 10.50%) y (91.67% ± 17.27%), respectivamente.

La figura 2 resume el porcentaje de reconocimiento de cada una de ellas. Se puede notar que 16 de las 28 expresiones alcanzaron un reconocimiento del 100%. La palabra con menor reconocimiento fue “coge ” con un 76.19% ± 25

20%, y siendo “toma” la opción mejor reconocida dentro de este tipo. mientras que las expresiones con un desempeño menor al 90% fueron: “coge” con (76.19% ± 25.20%), “dos” con (80.95% ± 17.82%), “cubo” con (80,95% ± 17,82%), finalmente “doce”, “sujeta” y “déjalo” con (85.71% ± 17.81%).

Figura 2. Resultados del reconocimiento del diccionario de palabras/componentes.

Experimentalmente se observaron errores en el reconocimiento de voz al confundir las palabras como “dos” con “doce”, “cubo” con “culo”, este tipo de nuevas expresiones podrían ser integradas al diccionario para evitar este tipo de errores que en ocasiones son dadas por la calidad del micrófono y la dicción o forma de pronunciar de los usuarios.

Tabla 2. Componentes de la estructura del comando de voz implementado

Componentes de la estructura del comando de voz implementado						
Tomar	**Objeto**	**Color**	**Índice**	**Dejar**	**Ubicación**	
['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']

También se evaluaron tres expresiones que podrían conformarse con los componentes del diccionario, cada una pronunciada tres veces por cada uno de los 7 usuarios. Se obtuvo un reconocimiento del 98.41% con una desviación estándar del 4.19% en las 63 ocasiones. La figura 3 presenta gráficamente el porcentaje de reconocimiento alcanzado por las tres expresiones evaluadas.

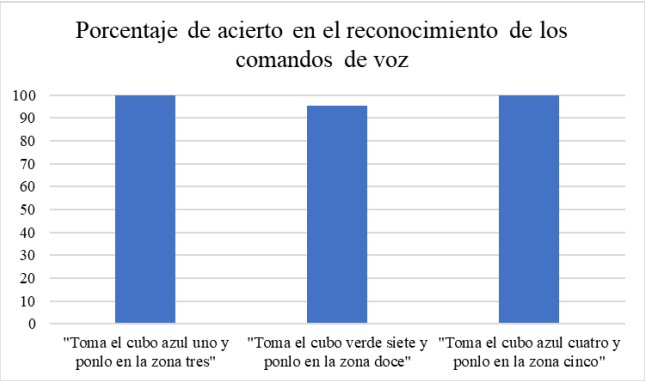

Figura 3. Resultados del reconocimiento de comandos de voz.

2) Interacción gestual

Se solicitó a cada uno de los siete usuario ejecutar el gesto “seleccionar” 3 veces, sobre cada una de las 12 zonas definidas en el espacio de trabajo. De este modo, en las 63 repeticiones se obtuvo un reconocimiento promedio del 81.35% con una desviación estándar del 7.98%. Las zonas 6, 7, 10 y 11 fueron las mejores , y ubicación está en la parte superior central de la imagen de la cámara. El porcentaje menor de reconocimiento ocurrió en las zonas 5 y 8 con (71.43% ± 23.00%) y (71.43% ± 12.60%) respectivamente, que se ubican en los bordes laterales centrales de la zona de trabajo.

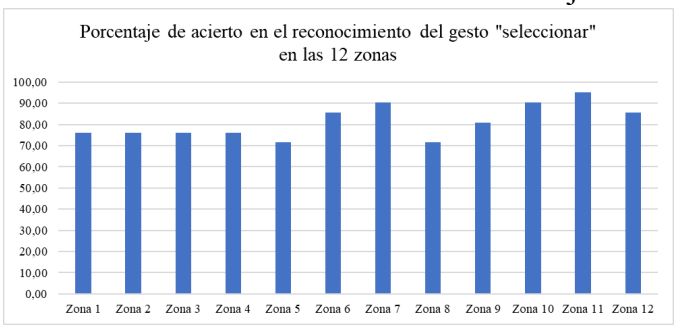

Figura 3. Resultados de interacciones gestuales.

3) Interacción multimodal

Cada uno de los 7 usuarios repitieron 3 veces la interacción multimodal que implica decir “Mira este cubo, ponlo aquí”, y al mismo tiempo, realizar el el gesto “seleccionar” durante la pronunciación de la parte “mira este cubo”, y posteriormente mover su mano hasta la zona a la que se desea llevar el cubo. Se eligieron 3 zonas como ubicación final del cubo, la 1, 5 y 9, que no fueron las de mejor reconocimiento en la ejecución del gesto “seleccionar” en la interacción gestual.Se logró un reconocimiento del 80.95% con una desviación estándar del 12.36% en las 63 interacciones.. En la figura 4 se muestran los resultados promedio obtenidos en el reconocimiento de las interacciones multimodales. Para más detalles sobre los resultados obtenidos.

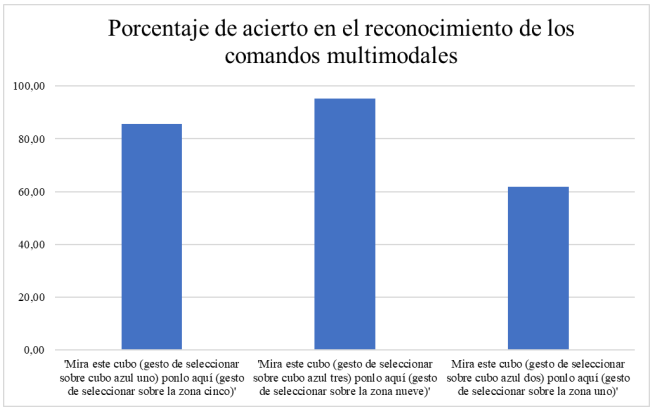

Figura 4. Resultados de interacciones multimodales.

Los resultados globales de esta prueba son muy similares a los presentados en las pruebas de la interacción gestual, lo que puede estar relacionado con el hecho de que la interacción gestual tiene mayor fuerza sobre los resultados de acierto en el

reconocimiento del comando, dado que esta es la que permite definir el cubo y la zona, además, para que el comando se cumpla correctamente ambas interacciones (verbal y gestual) deben ser bien reconocidas por la arquitectura.

4) Resultados de la Clasificación de los Objetos

Para evaluar el desempeño de la arquitectura se realizó una prueba con cada uno de los siete usuarios, en la que se le pide al usuario enseñarle una tarea al robot colaborativo UR3 con cada una de las posibles combinaciones de cubos.

Los resultados de esta prueba con un total de 63 interacciones en las que cada usuario ejecutó el ejercicio una sola vez por cada una de las nueve posibles combinaciones de cantidad de cubos, la arquitectura ejecutó correctamente el posicionamiento de todos los cubos de acuerdo a los recuerdos que almacenó en el 100% de los casos.

B. Pruebas cualitativas

Posterior a los tres tipos de interacciones que ejecutaron los siete usuarios, se les pidió diligenciar una encuesta de 5 preguntas, a través de las cuales indican su nivel de acuerdo o desacuerdo con el desempeño de la arquitectura, usando para ellos la escala *Likert de cinco niveles* [12]. Las preguntas se detallan en la Tabla 3.

Tabla 3. Cuestionario usado en la prueba cualitativa de la arquitectura.

N°	Pregunta
1	¿El sistema multimodal desarrollado en la arquitectura mejora la experiencia de usuario comparado con una interacción kinestésica o tele operada?
2	¿Los comandos de voz fueron identificados correctamente por la arquitectura?
3	En su opinión, ¿la arquitectura identificó correctamente los comandos gestuales?
4	¿Considera que la arquitectura identificó correctamente los comandos multimodales?
5	Desde su punto de vista, ¿considera que la realimentación gráfica/auditiva contribuyó a tener una mejor experiencia?

Los resultados obtenidos se presentan gráficamente en la Figura 17, donde sobresale la satisfacción de los usuarios con las interacciones verbales. Las preguntas 4 y 5 tuvieron un buen desempeño.La pregunta que hubo más disenso fue en la 1 sobre el aporte de la arquitectura en la interacción con el manipulador.

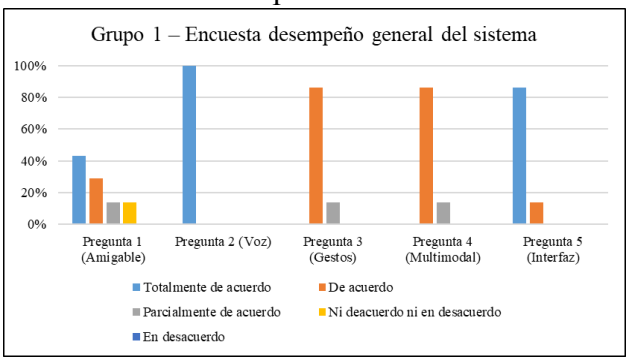

Figura 5. Resultados promedio de la encuesta del desempeño cualitativo de la arquitectura.

Finalmente, cinco de los siete usuarios encuestados manifestaron estar dispuestos a enseñar tareas a la arquitectura utilizando las interacciones multimodales como principal medio de enseñanza, a pesar de los errores presentados.

IV. DISCUSIÓN

El sistema propuesto tiene las siguientes características principales: (i) capacidad de interpretar comandos gestuales y verbales simultáneamente; (ii) uso de una arquitectura cognitiva SOAR para el aprendizaje y ejecución de tareas de clasificación; (iii) experimentación en un ambiente real con el robot UR3; (iv) realimentación gráfica y auditiva para mejorar la experiencia del usuario.

Las pruebas cuantitativas mostraron un alto desempeño del sistema en las interacciones verbales, con una tasa de acierto promedio del 98.41% y una baja desviación estándar del 4.19%. El análisis reveló que la distorsión de la lente y la inclinación de la cámara son factores importantes que introducen errores en las interacciones multimodales. Una solución sugerida es implementar un algoritmo de corrección de perspectiva más preciso y utilizar técnicas de calibración de la lente para mejorar la calidad de la imagen.

En las pruebas cualitativas, los usuarios expresaron satisfacción con la capacidad del sistema para identificar correctamente los comandos de voz, pero se identificaron oportunidades de mejora en la

identificación de los comandos gestuales y multimodales para las tareas enseñadas.

V. CONCLUSIONES

En este trabajo, se desarrolló y validó el desempeño de una arquitectura que utiliza interacción multimodal comandada por gestos y voz, para permitir a los usuarios enseñar y ejecutar tareas de clasificación de objetos cúbicos por color al robot colaborativo UR3. El enfoque utilizado fue una aproximación a la arquitectura cognitiva SOAR.

Aunque la arquitectura propuesta tiene un gran potencial, se han identificado problemas abiertos y limitaciones que requieren soluciones futuras. Estos incluyen la optimización de la cantidad de movimientos y la selección de recuerdos adecuados, así como la limitación del tamaño mínimo de las zonas para evitar colisiones. Se sugiere que para solucionar el último problema se podría agregar una funcionalidad adicional que permita a la pinza girar para evitar obstrucciones. Además, se observó que la corrección de coordenadas no fue perfecta en todos los casos, lo que sugiere la necesidad de fortalecer el módulo de visión con técnicas que permitan identificar la inclinación de la cámara y ajustar las correcciones en tiempo real.

VI. REFERENCIAS

[1] D. Tabuenca Alcusón, “Implantación de robots colaborativos en línea de producción,” *Trab. grado Ing. en Organ. Ind. Univ. valladolid*, p. 126, 2017, [Online]. Available: https://uvadoc.uva.es/bitstream/handle/10324/23076/TFG-584.pdf?sequence=1&isAllowed.

[2] I. Kotseruba and J. K. Tsotsos, *40 years of cognitive architectures : core cognitive abilities and practical applications*, vol. 53, no. 1. Springer Netherlands, 2018.

[3] R. E. Mayer, “Multimedia Learning (Second Edition), University of California, Cambridge University Press, ISBN 978-0-521-73535-3,” p. 304, 2009.

[4] R. E. Mayer, “Thirty years of research on online learning,” no. October 2018, pp. 152–159, 2019, doi: 10.1002/acp.3482.

[5] J. Sweller, J. J. G. Merriënboer, and F. Paas, “Cognitive Architecture and Instructional Design : 20 Years Later,” *Educ. Psychol. Rev.*, pp. 261–292, 2019, doi: 10.1007/s10648-019-09465-5.

[6] D. Choi and P. Langley, “Evolution of the ICARUS Cognitive Architecture,” *Cogn. Syst. Res.*, vol. 48, pp. 25–38, 2018, doi: 10.1016/j.cogsys.2017.05.005.

[7] J. E. Laird, *The SOAR of cognitive architecture*. 2013.

[8] J. H. Mosquera-DeLaCruz, H. Loaiza-Correa, S. E. Nope-Rodriguez, and A. D. Restrepo-Girón, “Disability and Rehabilitation : Assistive Technology Human-computer multimodal interface to internet navigation Human-computer multimodal interface to internet navigation,” *Disabil. Rehabil. Assist. Technol.*, vol. 0, no. 0, pp. 1–14, https://doi.org/10.1080/17483107.2020.179944, 2020.

[9] J. E. Laird, C. Lebiere, and P. S. Rosenbloom, “A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics,” *AI Mag.*, vol. 38, no. 4, pp. 13–26, 2017, doi: 10.1609/aimag.v38i4.2744.

[10] O. Janrathitikarn and L. N. Long, “Gait control of a six-legged robot on unlevel terrain using a cognitive architecture,” *IEEE Aerosp. Conf. Proc.*, 2008, doi: 10.1109/AERO.2008.4526240.

[11] N. M. Difilippo and M. K. Jouaneh, “Using the Soar Cognitive Architecture to Remove Screws From Different Laptop Models,” vol. 16, no. 2, pp. 767–780, 2019.

[12] S. Mcleod, “Likert Scale,” www.simplypsychology.org/likert-scale.html, pp. 1–4, 2008.

[13] C. Van Dang *et al.*, “Application of soar cognitive agent based on utilitarian ethics theory for home service robots,” *2017 14th Int. Conf. Ubiquitous Robot. Ambient Intell. URAI 2017*, pp. 155–158, 2017, doi: 10.1109/URAI.2017.7992698.

[14] A. D. Dubey and R. B. Mishra, “Cognition of a Robotic Manipulator Using the Q-Learning Based Situation-Operator Model,” *J. Inf. Technol. Res.*, vol. 11, no. 1, pp. 146–157, 2018, doi: 10.4018/JITR.2018010109.

[15] A. Bono, A. Augello, G. Pilato, F. Vella, and S. Gaglio, “An ACT-R based humanoid social robot to manage storytelling activities,” *Robotics*, vol. 9, no. 2, pp. 1–19, 2020, doi: 10.3390/ROBOTICS9020025.

[16] D. Choi, W. Shi, Y. S. Liang, K. H. Yeo, and J.-J. Kim, “Controlling Industrial Robots with High-Level Verbal Commands,” *Soc. Robot. 13th Int. Conf.*, pp. 216–226, 2021, doi: doi.org/10.1007/978-3-030-90525-5_19.