PI173-4-multimodal

Reconocimiento Multimodal del Estado Emocional de un Niño en un Contexto Educativo

Javier Alcalá Vásquez

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
2017

PI173-4-multimodal

Reconocimiento Multimodal del Estado Emocional de un Niño en un Contexto Educativo

Autor:

Javier Alcalá Vásquez

MEMORIA DEL TRABAJO DE GRADO REALIZADO PARA CUMPLIR UNO DE LOS REQUISITOS PARA OPTAR AL TITULO DE

MAGÍSTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Directora

Ing. Enrique González PhD

Comité de Evaluación del Trabajo de Grado

Carlos Alberto Parra Rodríguez

Cesar Julio Bustacara Medina

Página web del Trabajo de Grado

http://pegasus.javeriana.edu.co/~PI173-4-multimodal

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERIA DE SISTEMAS Y COMPUTACIÓN
BOGOTÁ, D.C.
Noviembre, 2017

PONTIFICIA UNIVERSIDAD JAVERIANA
FACULTAD DE INGENIERIA
MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Rector Magnífico

Jorge Humberto Peláez, S.J.

Decano Facultad de Ingeniería

Ingeniero Jorge Luis Sánchez Téllez

Director Maestría en Ingeniería de Sistemas y Computación

Ingeniera Ángela Carrillo Ramos

Director Departamento de Ingeniería de Sistemas

Ingeniero Efraín Ortíz Pabón

Artículo 23 de la Resolución No. 1 de Junio de 1946

"La Universidad no se hace responsable de los conceptos emitidos por sus alumnos en sus proyectos de grado. Sólo velará porque no se publique nada contrario al dogma y la moral católica y porque no contengan ataques o polémicas puramente personales. Antes bien, que se vean en ellos el anhelo de buscar la verdad y la Justicia"

AGRADECIMIENTOS

Este trabajo únicamente es posible gracias a la contribución definitiva del Ing. Enrique González PhD, quien con su dedicación, proceso metódico y riguroso y, en especial, con sus aportes definitivos, logró dar rumbo a esta idea que al principio se veía difícil y retadora y que ahora veo convertida en realidad. De igual forma le estoy profundamente agradecido a mi compañero John Jairo Páez Rodríguez. Sus aportes e ideas ayudaron a resolver problemas clave. En especial le agradezco su dedicación para lograr que las pruebas experimentales se llevaran a cabo de forma exitosa. Le deseo lo mejor en su proceso de doctorado.

A mi esposa, Ingrid, le debo todo. Ella me ha brindado su apoyo incondicional durante todo este tiempo. En los momentos más difíciles ha estado a mi lado y me ha animado a continuar hasta el final. Finalmente, le debo agradecer a mis hijas Catalina y Julieta, ustedes son la razón por la que todos los días quiero ser una mejor persona y un mejor padre.

Javier Alcalá Vásquez
Noviembre de 2017

A Dios, creador de todo
y esencia misma del conocimiento

Contenido

ABSTRACT

This work proposes an architecture for the multimodal recognition of a child's emotions relevant to the learning process in an educational context. The proposed architecture incorporates verbal, corporal, task and physiological modalities. In the development of the research, concepts of education and robotics, mental states, models of emotions and investigations regarding the recognition of emotions are explored, in particular with techniques and difficulties related to multimodal recognition. Some possibilities of existing software are explored and applied in the construction of an integrated solution that implements the proposed architecture for the multimodal recognition of emotions.

RESUMEN

Este trabajo propone una arquitectura para el reconocimiento multimodal de emociones de un niño relevantes al proceso de aprendizaje en un contexto educativo. La arquitectura propuesta incorpora las modalidades verbal, corporal, de tarea y fisiológica. En el desarrollo de la investigación, se exploran conceptos de educación y robótica, estados mentales, modelos de emociones e investigaciones efectuadas respecto al reconocimiento de emociones, en particular con técnicas y dificultades relativas al reconocimiento multimodal. Se exploran posibilidades de software existente y se aplican en la construcción de una solución integrada que implementa la arquitectura propuesta para el reconocimiento multimodal de emociones.

RESUMEN EJECUTIVO

Este trabajo de grado elabora una arquitectura de reconocimiento multimodal del estado emocional de un niño en un contexto educativo. Diversas investigaciones que se detallan en el marco teórico, muestran que dicho estado emocional es relevante para el proceso de aprendizaje y que esta información puede ser utilizada de manera efectiva en la implementación de un sistema robótico que sirva como herramienta al maestro para el mejor desempeño del estudiante durante el proceso.

La ejecución del trabajo, se distribuyó en tres fases. Durante la primera fase se profundizó en el estado del arte respecto a la relación entre educación y robótica, los conceptos de estado mental, estados emocionales y computación afectiva. Se investigaron diversos modelos de emociones especialmente diseñados para el contexto educativo. Adicionalmente, se investigaron técnicas específicas para el reconocimiento de emociones a través de señales de voz, de video y otros dispositivos sensoriales.

Se investigó también software de reconocimiento de emociones existente que haya sido utilizado en trabajos de investigación o que se ofrezca de manera comercial o como producto de código abierto. Como parte de la primera fase, se llevó a cabo un proceso de selección mediante criterios que llevó a un conjunto de componentes que fueron aplicados durante la implementación de la prueba de concepto a través de un montaje experimental. Para el proceso de selección, se efectuó una prueba funcional de laboratorio. Los resultados de esta prueba fueron utilizados para la selección final de los componentes y para mejorar la identificación de requerimientos, arquitectura y diseño del módulo.

La prueba funcional en laboratorio presentó un problema de análisis a los niños, de tal forma que resolvieran el reto, bajo la tutoría de un maestro guía. Durante la ejecución de la prueba se utilizó un dispositivo Kinect, una cámara de video y el wearable Xiaomi Mi Band 2. Los videos fueron presentados al panel de expertos, de tal forma que identificaran las emociones comunes y aquellas del modelo educativo. Esta información se comparó con el resultado del reconocimiento de emociones por parte de los componentes y se analizaron los datos para identificar aquellos que lograron un mejor desempeño durante la ejecución de la prueba funcional en laboratorio.

Posteriormente, en la fase 2, se identificaron los requerimientos del módulo y se escogió el modelo de emociones. Con esta información, se elaboró la arquitectura del módulo de reconocimiento y se realizó el diseño del agente de reconocimiento multimodal, de tal forma que fuera flexible, orientado por eventos y con bajo acoplamiento, permitiendo la incorporación de nuevos agentes modales y utilizando una estrategia de fusión de decisión. Para el diseño del Sistema Multiagente (SMA), se utilizó la metodología AOPOA. Los agentes se instancian mediante el framework BESA, esta es una plataforma de desarrollo que facilita la implementación SMA.

Con esta información, durante la fase 3, se elaboró una prueba de concepto a través de un montaje experimental en laboratorio. Se creó una prueba de concepto, en la que se implementó un agente modal que percibe las emociones a partir de las expresiones del rostro utili-

zando el software Affdex SDK. Se implementó también un agente modal de lenguaje utilizando el software Synesketch. Adicionalmente, se implementó un agente modal para el reconocimiento de emociones de la postura a partir de los datos registrados por un dispositivo Kinect durante la ejecución de la prueba. También se implementó un agente modal de interacción / tarea que percibe emociones a partir de la interacción del niño con los bloques que hacen parte del experimento. Para esta implementación se utilizaron los sensores de la cámara especializada Intel RealSense. Adicionalmente, se implementó el agente de reconocimiento multimodal, el cual recibe la información de emociones registrada por los agentes modales y las consolida utilizando la estrategia de fusión multimodal. La prueba de concepto se ejecutó en el montaje experimental de laboratorio, contando con la participación de niños de entre 10 y 13 años.

Para este experimento se incorporó el robot Baxter, de tal forma que interactuara con los niños durante la ejecución de la actividad de aprendizaje. La actividad incluyó dispositivos sensores que capturaran información de la postura del niño, su interacción con los elementos de la tarea, los gestos del rostro y el lenguaje utilizado. Esta información se presentó a un panel de expertos para determinar la percepción de las emociones comunes y las del modelo, durante la ejecución del experimento. Con estos datos, se comparó el resultado del agente multimodal frente a los resultados del panel de expertos. Se obtuvo un 8% de mejora en la detección de mejores comunes y un 5% de mejora frente a la detección de las emociones propias del modelo seleccionado. Estos resultados se calcularon al comparar la precisión del reconocimiento por parte del agente multimodal frente a la precisión del agente de reconocimiento que obtuvo el mejor desempeño en cada una de las emociones. De esta forma se logró demostrar que la aplicación de la estrategia multimodal diseñada mejora el desempeño de los módulos de reconocimiento por separado.

En resumen, las principales contribuciones de este trabajo de grado radican en elaborar una arquitectura flexible para el reconocimiento multimodal de emociones, la incorporación de emociones especializadas en el proceso educativo más allá de las emociones comunes previamente estudiadas, y el diseño de agentes especializados en el reconocimiento de emociones a partir de la postura y la interacción con los elementos que hacen parte de la actividad educativa. También se integran una serie de dispositivos, el robot Baxter, Kinect, Intel RealSense, wearables, mediante los cuales se logra incorporar información de postura y de interacción con los bloques. Esta puesta en funcionamiento es novedosa, puesto que en los trabajos de investigación identificados durante la profundización del estado del arte, no se encontró un montaje tan completo en el que también hiciera parte la interacción con el robot humanoide y que formara parte de un proceso educativo.

El área de reconocimiento de emociones en general, admite mayores esfuerzos por parte de los investigadores para encontrar métodos más adecuados a las necesidades de evaluación en línea y bajo consumo de recursos de procesamiento. Nuevas técnicas de inteligencia artificial pueden aplicarse para desarrollar agentes modales que se integren con la arquitectura propuesta.

INTRODUCCIÓN

La interacción entre robots y humanos se desarrolla en situaciones que incluyen el uso de comunicación verbal y no verbal en diversos contextos: educativos, situaciones de entretenimiento, cuidado de personas de la tercera edad, ambientes de hogares inteligentes, entre otros [1], [2]. Con el propósito de brindar una retroalimentación efectiva al sistema robótico, es necesario hacer seguimiento al estado emocional y a la intencionalidad por parte del usuario [3]. A la combinación de estos dos factores, el estado emocional del usuario y su intención, se le denomina "estado mental" [4], también se le conoce como "estado afectivo" [5].

La utilización de robots en educación presenta beneficios tanto para alumnos como para docentes. La interacción entre robots y humanos mejora la efectividad del aprendizaje y además la motivación de los estudiantes en tanto se encuentran en un ambiente que incentiva el interés del alumno, la aplicación práctica del conocimiento y un entorno propicio para el aprendizaje [2], [6]–[8]. Al interactuar con un robot, la experiencia se hace tangible y se logra mayor compromiso por parte del alumno que lo que se alcanzaría con la intermediación de video o mecanismos de realidad virtual [2]. El uso de robots en este contexto no implica el remplazo del maestro sino que el propósito es facilitarle herramientas que le permitan llevar un rol de organizador, guía, tutor y orientador del proceso de aprendizaje [9], [10].

En este trabajo se diseñan mecanismos para el reconocimiento multimodal del estado emocional de un niño en un contexto educativo. El resultado contribuye al desarrollo de la tesis de doctorado en curso del estudiante John Jairo Páez Rodríguez: Aprendizaje de la estrategia de Análisis de Medios-Fines a través de un robot antropomórfico que da soporte metacognitivo y emocional, cuya arquitectura general se denomina HRS-BDIBESA [13]. Este trabajo de grado se enmarca en el contexto de dicha tesis. En especial, se profundiza en el desarrollo de una arquitectura flexible que permita agregar diferentes modalidades de reconocimiento y que sus resultados sean fusionados a través de una estrategia de decisión multimodal. Esta información es utilizada para retroalimentar la toma de decisiones de un sistema educativo que soporta el proceso de aprendizaje mediante la interacción con un robot humanoide.

Durante el desarrollo del trabajo se identifican las diferentes modalidades significativas para la implementación del módulo de reconocimiento en el contexto educativo definido. Se evalúa y selecciona software disponible que facilite la implementación del módulo, a la vez que se utilizan componentes especializados en la identificación de emociones. Adicionalmente, se diseñan mecanismos especiales para el reconocimiento a través de la interacción del niño con los elementos de la tarea y se implementa el reconocimiento a través de información de postura obtenida con un dispositivo Kinect.

El documento a continuación, presenta la metodología utilizada y profundiza en el desarrollo de los diferentes aspectos haciendo énfasis en la arquitectura de reconocimiento multimodal. En el capítulo 1 se realiza una descripción general de la problemática que aborda el trabajo de investigación. En el capítulo 2 se detallan el objetivo general y objetivos específicos, al igual que se presentan las fases de la metodología utilizada para la ejecución del proyecto. El capítulo 3 aborda el estado del arte presentando la relación entre educación y robótica, los conceptos de estado mental, estados emocionales y computación afectiva, al igual que presenta

modelos de emociones relacionados al contexto educativo. Más adelante, en el mismo capítulo se relacionan técnicas específicas para el reconocimiento de emociones a través de señales de voz y de video mediante la utilización de diferentes dispositivos sensoriales.

Posteriormente, se presentan estudios relacionados con técnicas para el reconocimiento multimodal de emociones. A partir de estos trabajos, se destacan también componentes de software que podrían ser utilizados en la construcción de una solución integrada. En el capítulo 4 se detalla la arquitectura y diseño del módulo de reconocimiento propuesto utilizando un enfoque de Sistemas Multiagente (SMA). Posteriormente, en el capítulo 5 se muestra el detalle de la prueba experimental realizada. En el capítulo 6 se realiza el análisis de datos recolectados durante la prueba experimental. Finalmente, se presentan las conclusiones del trabajo y las opciones de trabajo futuro.

1 DESCRIPCIÓN GENERAL

Oportunidad y problemática

El estudio de emociones más relacionadas con el aprendizaje, tales como el estado de confusión, frustración, aburrimiento, afluencia, curiosidad o ansiedad son tan importantes a los ambientes de aprendizaje como otros estados emocionales básicos tales como la ira, el miedo, la alegría o la tristeza [11]. Este enfoque más amplio que incluye la recopilación de sentimientos, estados de ánimo, actitudes, estilos afectivos, y el temperamento ha llevado al desarrollo del campo de la "Computación Afectiva" [11], [12].

La Figura 1 presenta el modelo de cuatro cuadrantes propuesto por Kort y Reilly [5], especialmente diseñado para el uso en Interacción Hombre – Máquina (HCI) y Computación Afectiva. El eje horizontal muestra las emociones positivas a la derecha y las negativas a la izquierda, mientras que el eje vertical simboliza la construcción de conocimiento o el descarte de ideas en la parte inferior.

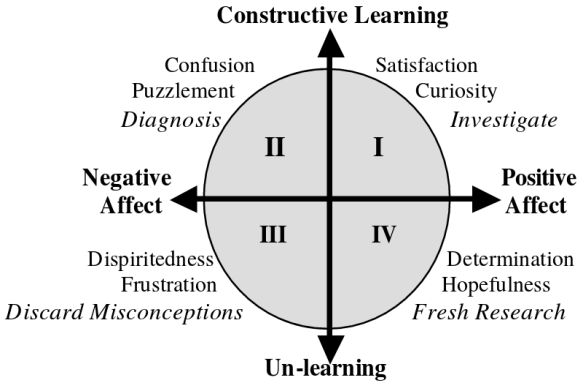


Figura 1 Modelo que relaciona fases de aprendizaje y emociones. Tomado de [5].

La Figura 2 presenta el contexto para el módulo de reconocimiento multimodal, en el que un niño interactúa con un sistema educativo durante el desarrollo de una tarea para resolver un problema particular. El propósito del desarrollo del modelo es reconocer que el estado emocional del estudiante impacta positiva o negativamente el proceso de aprendizaje y que la intervención apropiada en el estado afectivo facilitará el aprendizaje, para lo cual es necesario identificar con precisión el estado cognitivo-emotivo del estudiante.

El reconocimiento de estados emocionales puede incluir datos de la interacción [13], reconocimiento de lenguaje natural, de la entonación [4], [14], expresión facial o postura corporal [15], [16]. Múltiples fuentes de información pueden ser tenidas en cuenta para lograr un modelo del estado emocional del usuario. A esta estrategia que integra varias fuentes se le denomina reconocimiento multimodal [17], [18]. La información así obtenida es necesaria para brindar retroalimentación al sistema robótico [4], de esta forma, otros módulos del sistema podrán brindar soporte metacognitivo o emocional al estudiante con el propósito de alcanzar los objetivos de aprendizaje [7].

Trabajos previos de reconocimiento de emociones multimodal durante la interacción humano-robot [17], [19], [20] utilizan dos modos, análisis de voz y de expresión facial, para aplicarlos en un robot social. Barros et al [18] emplean redes neuronales con estímulos de expresión facial y movimiento corporal para el reconocimiento en escenarios sociales. La investigación de D'Mello y Graesser [21] utiliza datos preexistentes de diálogo, postura y expresión facial tomados de un sistema tutor. En el trabajo de Saneiro et al [22] se construye una base de datos de video y sonido en un escenario educativo de tal forma que pueda ser usada en software de reconocimiento de emociones multimodal. Estos trabajos demuestran la utilidad de la integración multimodal para el análisis de emociones, sin embargo, los modelos utilizados están orientados a emociones genéricas (felicidad, tristeza, ira). En este trabajo de grado, el modelo emocional que se trabaja está ligado al contexto educativo, y en particular a reconocer el estado emocional del niño durante el proceso de aprendizaje; se reconocerán emociones relevantes como confusión, frustración, aburrimiento, curiosidad o ansiedad. En efecto, gracias a esta información se podrá apoyar en forma adecuada la toma de decisiones de soporte metacognitivo y emocional.

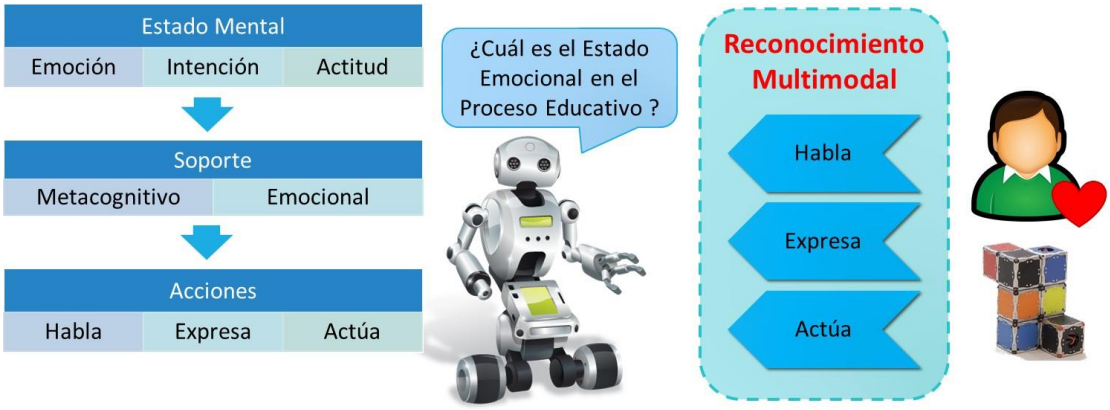


Figura 2 Contexto del sistema de reconocimiento multimodal de emociones. Elaboración propia.

La pregunta de investigación que se busca responder con este trabajo es ¿Cómo reconocer el estado emocional de un niño en un contexto educativo para dar retroalimentación efectiva a un robot humanoide con el que interactúa? Tal como se mostró anteriormente, existen varias posibilidades en una estrategia de reconocimiento multimodal [17], [23].

La investigación aporta en la integración multimodal de componentes de reconocimiento auditivo o visual en una solución de reconocimiento del estado emocional. El diseño del módulo está acompañado por la prueba de concepto a través de un montaje experimental de laboratorio, de tal forma que se logre una implementación que incluya al robot Baxter disponible en la Facultad de Ingeniería de la Pontificia Universidad Javeriana. Baxter es un robot de tipo humanoide con funciones como gravedad cero y múltiples sensores [24], [25] que lo hacen ideal para su uso en interacción con humanos y, en este caso, con niños en un entorno educativo [26], [27].

2 DESCRIPCIÓN DEL PROYECTO

2.1 Objetivo general

Diseñar mecanismos de reconocimiento de indicadores del estado emocional de un niño de entre 10 y 13 años, a partir de información multimodal, durante la interacción con un robot humanoide en un contexto educativo no formal en la ciudad de Bogotá.

2.2 Objetivos específicos

1. Identificar los requisitos del módulo a partir del estado del arte y las condiciones sensoriales aplicables a un robot humanoide en el marco de la arquitectura de referencia HRS-BDIBESA.
2. Seleccionar componentes disponibles en el mercado que permitan la integración de una solución de reconocimiento multimodal del estado emocional del niño en un proceso de aprendizaje.
3. Diseñar el módulo de reconocimiento multimodal del estado emocional del niño a partir de los requisitos y componentes identificados, de acuerdo a las condiciones del entorno.
4. Realizar una prueba de concepto del módulo de reconocimiento multimodal a través de un montaje experimental de laboratorio utilizando el robot Baxter.

2.3 Fases de desarrollo

El diseño y prueba experimental de los mecanismos para el reconocimiento multimodal del estado emocional del niño se distribuyen en tres fases:

1. Identificación de requisitos y selección de componentes.
2. Arquitectura y diseño del módulo de reconocimiento.
3. Prueba experimental.

La Figura 3 muestra la articulación de las fases de tal forma que contribuyen simultáneamente al logro de los objetivos del proyecto. La fase 3 se desarrolla en paralelo a las fases 1 y 2 debido a que el caso de estudio de laboratorio y el diseño del protocolo experimental están relacionados con los requisitos y la arquitectura del sistema. La identificación de requisitos inicial proporciona información necesaria para la arquitectura de agentes que a su vez brinda la estructura técnica relevante para la prueba de concepto. Una relación similar ocurre entre el diseño del módulo y la prueba experimental en tanto que el diseño del caso de estudio se realiza en paralelo a las primeras iteraciones de la fase de diseño del módulo.

En la fase 1 se completaron los objetivos 1 y 2 (Identificación de requisitos y selección de componentes). El contexto del sistema de reconocimiento multimodal de emociones para la arquitectura HRS-BDIBESA a la que se hace referencia en el objetivo 1, fue presentado en la Figura 2. En la fase 2 se elabora el diseño detallado del módulo de reconocimiento cumpliendo con el objetivo 3. Finalmente, el objetivo específico 4 se completa en la fase 3, de forma transversal, con la prueba de concepto y el montaje experimental.

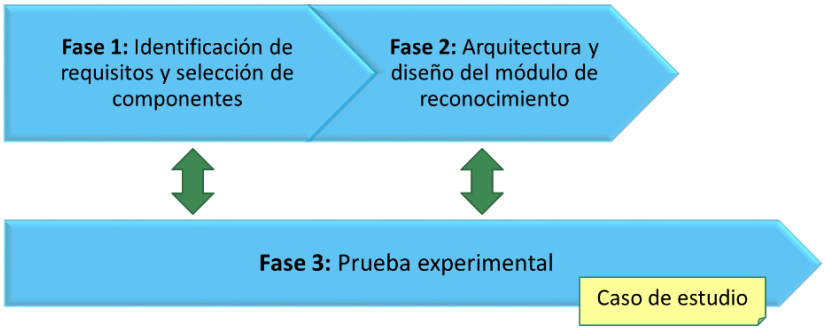


Figura 3 Secuencia de las fases.

2.3.1 Fase 1 Identificación de requisitos y selección de componentes

La Fase 1 prepara los elementos necesarios para la arquitectura y el diseño. Durante la ejecución de esta fase, se aplica una metodología de investigación mixta de tipo exploratorio, en la que primero existe una etapa cualitativa enfocada en la profundización del estado del arte, la identificación de requerimientos funcionales y la búsqueda de componentes existentes que proporcionen funciones requeridas por la solución integrada de reconocimiento multimodal. Posteriormente se realiza una selección basada en criterios de tal forma que, mediante la aplicación de una prueba de laboratorio. Los criterios de selección fueron definidos conforme al análisis cualitativo del estado del arte. Algunos criterios aplicados se relacionan a continuación:

- Tipo de licencia.
- Disponibilidad durante el tiempo de la investigación.
- Compatibilidad con los sensores del robot Baxter.
- Velocidad de procesamiento.
- Precisión en tiempo real.
- Posibilidad de procesar idioma español para los casos de procesamiento de voz.

Los componentes se preseleccionan a partir de filtros de primer nivel que conducen a una lista depurada para la prueba de laboratorio. Una vez se ejecuta dicha prueba, se analizan los resultados, y a partir de un análisis comparativo se seleccionan aquellos que reciban el mayor puntaje.

Para el desarrollo de esta fase se realizan las siguientes actividades:

a) Profundizar en el estado del arte.
b) Identificación de requerimientos.
c) Definición de criterios de selección.
d) Identificación de componentes aplicables.
e) Ejecutar la prueba funcional en laboratorio.
f) Seleccionar componentes a partir de los resultados de la prueba.
g) Análisis y selección de componentes.

2.3.2 Fase 2 Arquitectura y diseño del módulo de reconocimiento

En esta fase se construyó un sistema de agentes racionales que capaz de reconocer emociones relevantes al contexto de aprendizaje (confusión, frustración, aburrimiento, curiosidad o ansiedad). A partir de los requisitos identificados y el modelo de emociones, se elaboró la arquitectura y diseño del módulo de reconocimiento. En primera instancia, se utilizó una metodología de Sistema Multiagente (SMA) debido a que el módulo integra un conjunto de componentes especializados concurrentes y cooperativos [22]. Gracias a este análisis SMA se determinó no solo la funcionalidad de los módulos sino también la forma en que estos interactúan. Luego, en segunda instancia, se procedió a diseñar la inteligencia de los agentes que lo requerían. En efecto, los agentes pueden en sí mismos implementarse usando componentes existentes o mediante algoritmos de clasificación de fuentes heterogéneas como Deep Learning, redes neuronales o Support Vector Machines (SVM).

Es claro que la arquitectura SMA no impone restricciones a la racionalidad de los agentes, por cuanto es totalmente compatible con el uso de diversas herramientas de Inteligencia Artificial. La decisión sobre la herramienta específica justamente es el principal resultado de las actividades de esta fase. Por ejemplo, Prado et al [19] utilizan redes neuronales en la implementación del reconocimiento de emociones, al igual que Zhang et al [28]. Por su parte, Dobrisek et al [20] emplean mecanismos de Support Vector Machines (SVM) para la clasificación, al igual que las investigaciones [29]–[31].

Específicamente, el modelo de emociones fue seleccionado a partir del estado del arte especializado en el contexto de aprendizaje; en particular, este modelo está definido en la arquitectura HRS-BDIBESA de la tesis doctoral que enmarca este trabajo de grado. Para el diseño de la arquitectura SMA, se utilizó la metodología AOPOA [32]. A diferencia de otras metodologías de diseño SMA como MAS-CommonKADs [33] o Tropos [34], AOPOA llega a un nivel de diseño detallado, lo que resulta adecuado al alcance de este trabajo de investigación. Para el diseño del sistema inteligente se aplicó la metodología de desarrollo de Sistemas Inteligentes para Aprendizaje Inductivo, utilizada en el curso de Sistemas Inteligentes de la maestría de la Pontificia Universidad Javeriana [35]. En esta metodología los pasos que se siguen son los siguientes: planteamiento del problema, identificación de requerimientos del sistema, revisión bibliográfica, selección de la técnica de IA, caracterización de variables, elección de la base de ejemplos, técnicas de pre-procesamiento, entrenamiento experimental y validación cruzada.

Para el desarrollo de esta fase se realizan las siguientes actividades:

a) Identificación de requerimientos funcionales y no funcionales.
b) Caracterización del modelo de emociones de aprendizaje.
c) Descomposición de metas AOPOA y diseño de arquitectura de alto nivel.
d) Diseño detallado de agentes e interacciones.
e) Diseño inteligente para clasificación con fuentes heterogéneas.
f) Arquitectura y diseño.

2.3.3 Fase 3 Prueba experimental

En paralelo a las fases 1 y 2 se realizó la fase 3 de prueba experimental, el caso de estudio de laboratorio y el diseño del protocolo experimental están relacionados con los requisitos y la arquitectura del sistema. El grupo experimental estuvo compuesto por niños con edades entre 10 y 13 años. Este rango de edad se escogió con base a la teoría del desarrollo cognitivo de Jean Piaget, coincidiendo con la frontera en que se han desarrollado habilidades analíticas concretas e inicia el desarrollo del análisis abstracto en el niño [36], [37]. La Figura 4 muestra las etapas del modelo de Piaget y la ubicación del rango de edad seleccionado para los niños que hacen parte del grupo experimental.

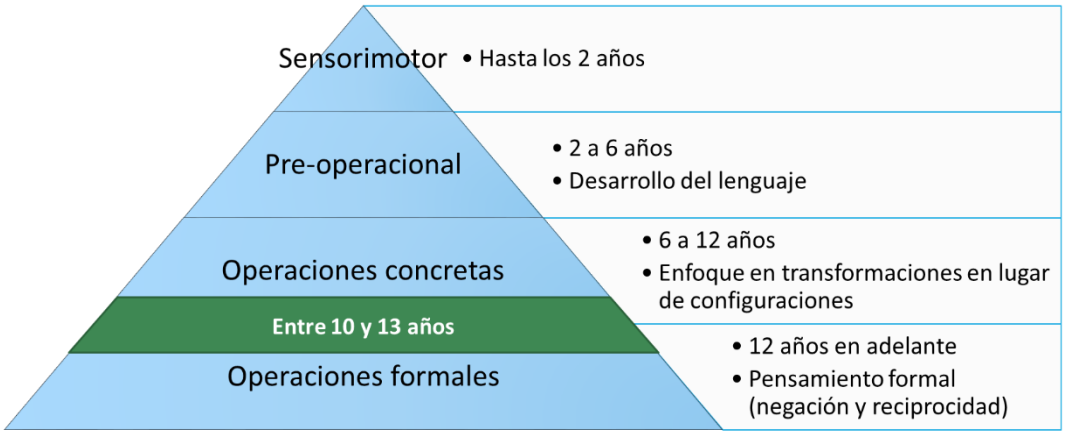


Figura 4 Etapas del desarrollo cognitivo. Elaboración propia adaptado de [37].

Se aplicó una metodología de investigación experimental, mediante un caso de estudio en un ambiente controlado de laboratorio. El diseño del protocolo de experimental tuvo en cuenta la situación especial del grupo experimental. Puesto que se trata de interacción con menores de edad, se contó con autorización escrita por parte de los padres o tutores. El montaje experimental fue aplicado de acuerdo al diseño previo y los datos así recolectados fueron analizados con el propósito de elaborar las conclusiones del proyecto.

A continuación se plantea el esquema del experimento:

- Se diseña una actividad en la que una persona, el instructor, presenta un problema con bloques al niño invitándolo a resolverlo.
- En cada sesión del experimento participó un solo niño.
- En esta actividad se le pidió al niño que hable sobre las ideas que tiene para resolver el problema mientras lo ejecuta.
- Se diseñó un ambiente de grabación con cámaras y audio en el que se registraron los datos necesarios para el módulo de reconocimiento.
- Se contó con un panel de expertos en educación de tal forma que evaluaron el grado de concordancia entre el resultado del módulo y las imágenes del experimento.

- Se presentó a cada experto en forma individual el resultado generado por el módulo de tal forma que pueda indicar en cada caso si está de acuerdo o en desacuerdo en una escala de Likert de cinco niveles.
- Se analizaron los datos de acuerdo a los resultados en los diferentes puntos de control respecto al grado de concordancia y si existían variaciones en el grupo experimental.
- Finalmente, se elaboraron conclusiones y recomendaciones.

Para el desarrollo de esta fase se realizan las siguientes actividades:

a) Definir el caso de estudio de laboratorio.
b) Diseño del protocolo experimental.
c) Reclutamiento del grupo experimental.
d) Construcción del prototipo para el experimento.
e) Aplicación del experimento.
f) Análisis de datos y conclusiones.
g) Escritura final del documento de tesis.

3 MARCO TEÓRICO / ESTADO DEL ARTE

Este capítulo aborda el estado del arte presentando la relación entre educación y robótica, los conceptos de estado mental, estados emocionales y computación afectiva, al igual que presenta modelos de emociones relacionados al contexto educativo. Más adelante, se relacionan técnicas específicas para el reconocimiento de emociones a través de señales de voz y de video mediante la utilización de diferentes dispositivos sensoriales. Posteriormente, se presentan estudios relacionados con técnicas para el reconocimiento multimodal de emociones. A partir de estos trabajos, se identifican también componentes de software que podrían ser utilizados en la construcción de una solución integrada.

3.1 Educación y robótica

La Asistencia Social Robótica (SAR) es aquella que proporciona asistencia a los usuarios humanos a través de la interacción social más que física [10]. La aplicación de los robots sociales incluye su uso en procesos educativos como es el caso de enseñanza de lenguaje en diversos idiomas [6], [38], tecnología y computación [39], ciencias experimentales [40] o educación preescolar [7], [41]. Se han elaborado también estudios especializados en educación de niños con síndrome de Down [8], autismo [25], [42], o generales como es el caso de la solución de problemas [11], entre otras aplicaciones en contextos educativos [1], [2], [43].

Al proporcionar al sistema robótico la capacidad de detectar las expresiones emocionales del estudiante, se promueve el desarrollo de una integración cognitiva que favorece la adaptabilidad del sistema [22]. En este sentido se pueden utilizar datos fisiológicos, provenientes de la interacción con dispositivos como ratón o teclado, información subjetiva proporcionada por el instructor, registros de las expresiones faciales o movimiento corporal por parte del estudiante. Con las fuentes de información mencionadas y su integración, es posible identificar expresiones durante una tarea de aprendizaje que logren un impacto positivo en la efectividad del proceso dependiendo de la duración de la tarea o su nivel de dificultad [22].

En el contexto presentado en la Figura 2, el robot humanoide actúa como una herramienta para el educador, de tal forma que las tareas que se asignan al estudiante cuenten con un seguimiento adecuado al plan de trabajo y retroalimentación conforme al desempeño de la actividad (soporte metacognitivo) [44]. De igual forma resalta la necesidad de brindar al alumno soporte emocional según las expresiones del niño en lenguaje verbal y no verbal con el propósito de encausar el proceso de aprendizaje y alcanzar el objetivo del proceso educativo.

La arquitectura HRS-BDIBESA [44], a la que se hace referencia en el objetivo 1, se detalla en la Figura 5. En este diagrama los sensores de un sistema robótico reciben una serie de información resultado de la ejecución de la tarea por parte del niño. El resultado del módulo de reconocimiento del estado emocional, se transfiere a un módulo que procesa las creencias del sistema, de tal forma que se tomen decisiones orientadas a contribuir en el desarrollo del proceso de educativo. Utilizando la arquitectura BDI [45], el sistema ejecuta un conjunto de acciones encaminadas a brindar soporte al niño tanto desde el punto de vista cognitivo como emocional.

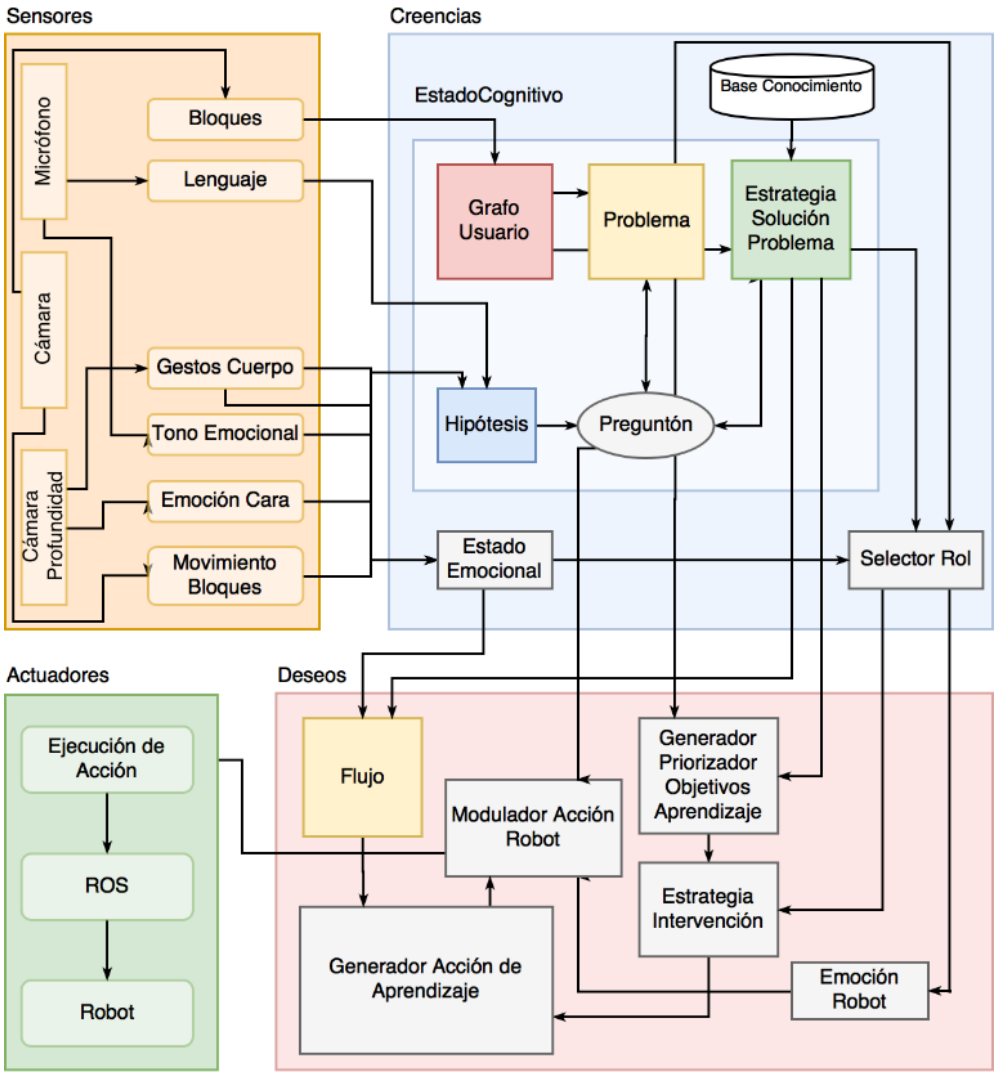


Figura 5 Arquitectura HRS-BDIBESA. Tomado de [44].

Este trabajo se enfoca en la propuesta de una arquitectura que permita el reconocimiento de emociones para este contexto educativo, en el marco de una situación general en la que el sistema educativo usaría esta información para tomar acciones adecuadas a los objetivos del proceso de aprendizaje.

3.2 Estado mental y estados emocionales

En la introducción fue presentado el término “estado mental” como la combinación de estado emocional y la intencionalidad por parte del usuario [4]. Desde la perspectiva de Sistemas Multiagente (SMA), se define estado mental como la información interna de un agente empleada para tomar acciones en las situaciones presentadas por su entorno [36]. Dichas situaciones pueden ser problemas que requieren solución por parte del agente. A su vez, diferentes

agentes pueden asignar diferentes prioridades a variables similares en sus estados mentales lo que conlleva cierto grado de personalidad. El estado mental puede incluir sentidos de valores, intenciones, obligaciones, capacidades, entre otros.

Los seres humanos cuentan con la capacidad de manifestar estados mentales complejos en tanto que expresan simultáneamente variadas emociones como ira, miedo, felicidad, sorpresa, tristeza o desagrado, a las que se suman otros estados neutrales como confusión, sorpresa o concentración [18]. Tales estados mentales involucran actitudes, estados cognitivos e intenciones. Otros autores han propuesto la detección de acciones dinámicas y también los movimientos físicos como la principal fuente de información para reconocer los estados mentales [4].

No existe consenso al momento de definir estado emocional [18], algunos autores consideran equivalentes los estados mentales y los estados emocionales. En tanto que el afecto es un mecanismo evolutivo que ocupa un papel fundamental en la interacción humana, se ha agregado el término "estado afectivo" en referencia a emociones, actitudes, creencias, intenciones, deseos, entre otros [4]. Aquellos sistemas, que incluyen información emocional del usuario, han propiciado una nueva área de investigación conocida como "computación afectiva" [23].

3.3 Modelos de emociones

Puesto que la clasificación de las emociones humanas puede llevar a un alto nivel de complejidad, se han desarrollado modelos que buscan representar dicha complejidad en un subconjunto de emociones. El psicólogo Robert Thayer desarrolló un modelo, fuertemente relacionado con componentes psicofisiológicos y bioquímicos en el que las actividades cognitivas personales y los eventos causales juegan un papel crítico [46]. El modelo de emociones de Thayer se muestra gráficamente en la Figura 6.

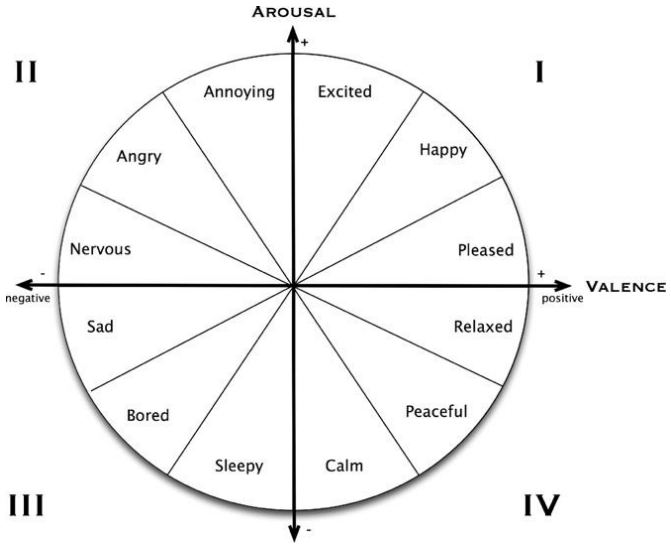


Figura 6 Modelo de emociones de Thayer. Tomado de [46].

El modelo de Thayer está compuesto por cuatro cuadrantes en los que el eje horizontal representa la valencia positiva o negativa de las emociones, mientras que el eje vertical representa el nivel de excitación, es decir, qué tan emocionado o en calma se encuentra la persona. Dicho modelo incluye un total de doce emociones reconocibles, entre más cercano se está al origen de la gráfica menos intensas son las emociones, lo contrario ocurre al alejarse del origen.

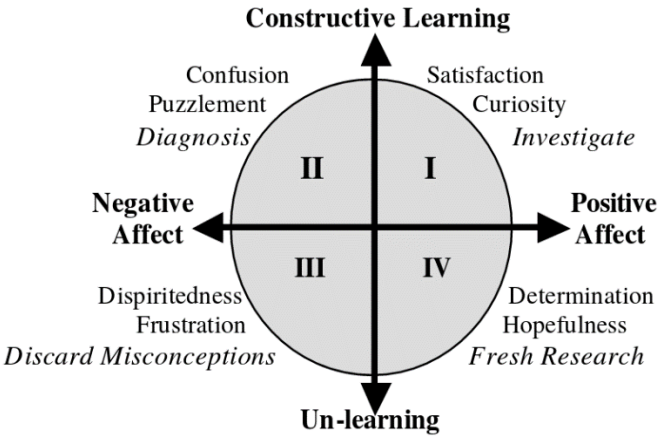


Figura 7 Modelo que relaciona fases de aprendizaje y emociones. Tomado de [5].

La Figura 7 presenta el modelo de cuatro cuadrantes propuesto por Kort y Reilly [5], especialmente diseñado para el uso en Interacción Hombre - Máquina (HCI) y Computación Afectiva. El eje horizontal muestra las emociones positivas a la derecha y las negativas a la izquierda, mientras que el eje vertical representa la construcción de conocimiento en la parte superior y el descarte de ideas en la parte inferior. En los cuatro cuadrantes ubicados en la gráfica se genera un ciclo en el que el estudiante puede iniciar en el primer cuadrante en una situación de inquietud por adquirir nuevo conocimiento, pudiendo pasar a otros cuadrantes en los que se produzca frustración debido al descarte de ideas preconcebidas para luego avanzar a un estado de expectativa por la adquisición de nuevos modelos de conocimiento y el inicio de un nuevo ciclo a partir del primer cuadrante ante nuevas preguntas y situaciones que promuevan un nuevo ciclo de aprendizaje.

El propósito del desarrollo del modelo es reconocer que el estado emocional del estudiante impacta positiva o negativamente el proceso de aprendizaje y que la intervención apropiada en el estado afectivo facilitará el aprendizaje, para lo cual es necesario identificar con precisión el estado cognitivo-emotivo del estudiante [5]. Esta es la principal fortaleza de este modelo frente al modelo genérico de Thayer, al estar diseñado específicamente para contextos educativos, se encuentra en capacidad de reconocer emociones de mayor relevancia al aprendizaje como confusión, frustración, aburrimiento, curiosidad o ansiedad. Estas emociones se agrupan en ejes emocionales, tal como lo muestra la Figura 8. Por tal motivo, la arquitectura que se presentará más adelante, para el reconocimiento multimodal de emociones, utiliza este modelo y lo aplica en la identificación de este tipo de emociones.

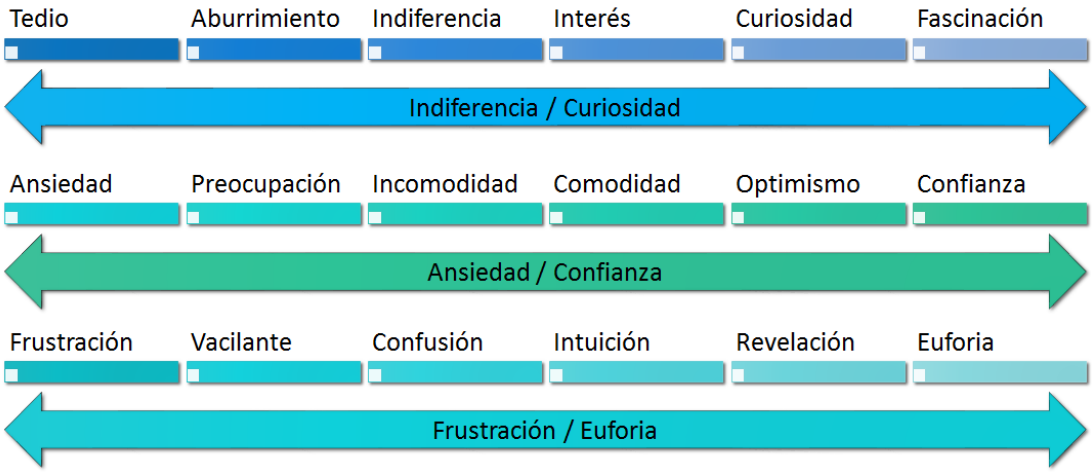


Figura 8 Ejes emocionales del modelo de Kort y Reilly [5]. Elaboración propia.

3.4 Técnicas de reconocimiento

En tanto que se ha mostrado la importancia de las emociones para el proceso educativo, en esta sección se presentan técnicas que permiten su reconocimiento. Estas técnicas puede incluir datos de la interacción [13], reconocimiento de lenguaje natural, de la entonación [4], [14], expresión facial, postura corporal [15], [16] o fisiológicos. La información así obtenida es necesaria para brindar retroalimentación al sistema robótico [4], [44].

3.4.1 Audio

La interpretación de la lengua hablada por parte de un computador es un proceso complejo realizado por sistemas de comprensión del lenguaje hablado (SLU) [30]. Una de las técnicas de modelado discriminativo más utilizadas, es la de campos aleatorios condicionales (CRF). En esta técnica, a partir de una expresión como “quiero volar de Seattle a Miami mañana en la mañana”, se busca reducir la comprensión a la ubicación de variables específicas como ciudad de destino, ciudad y fecha de partida en una intención específica relacionada con la búsqueda de un vuelo.

El reconocimiento de emociones con esta técnica requiere: una tarea de reconocimiento de voz automática (ASR) para obtener una secuencia o un conjunto de hipótesis de palabras, y una tarea de interpretación que transforma dichas hipótesis de palabras en hipótesis de estructura semántica descritas por un lenguaje de representación de significado (MRL) [29]. Basándose en teorías lingüísticas, las estructuras semánticas se obtienen por la composición de las estructuras semánticas que son fragmentos de la ontología aplicada por el sistema SLU.

El reconocimiento de emociones a través de señales de audio involucra también el uso de la tonalidad de la voz. El rendimiento del tono acústico y de las características lingüísticas es comparable; una combinación de ambos conjuntos mejora el rendimiento del reconocimiento [47]. Otros modelos incorporan la tonalidad como entrada directa para el reconocimiento

emocional, mientras que las técnicas de lenguaje natural se emplean para el reconocimiento de la intencionalidad del usuario [4].

3.4.2 Video

Las técnicas de reconocimiento de emociones basadas en video se pueden clasificar en dos grupos [20]:

1. Técnicas basadas en características que detectan y hacen seguimiento de aspectos faciales, tales como las esquinas de la boca o las cejas, y utilizan la información obtenida para llevar a cabo el reconocimiento de emociones.
2. Enfoques basados en la región, donde el movimiento facial se mide en ciertas regiones de la cara y luego son explotados para el reconocimiento de emociones. Patrones de movimiento del globo ocular y la variación de tamaño de la pupila pueden considerarse como posibles factores de reconocimiento de intención [3].

El trabajo de Littlewort [48] utiliza video para realizar reconocimiento de emociones a partir de la expresión del rostro y la posición relativa de la cabeza, logrando clasificar cinco emociones en tiempo real: ira, miedo, alegría, alivio y tristeza. En otros trabajos se propone una arquitectura neuronal jerárquica capaz de reconocer las acciones humanas observadas [49]. Cada capa en la arquitectura representa características cada vez más complejas de la actividad humana.

Otra característica que puede ser extraída de dispositivos especiales son los movimientos del cuerpo y la expresión de emociones a partir de lenguaje corporal [15]. Esta técnica resulta especialmente efectiva cuando el contacto visual con el usuario se realiza a una distancia apropiada, de tal forma que se logre la extracción de emociones a partir del movimiento. Es necesario tener en cuenta variables como las diferencias de género, culturales y las dificultades inherentes a la recolección de los datos de postura en tiempo real y la diferenciación frente al entorno [50]. Es estos casos se utiliza un dispositivo especializado como Kinect para el registro del movimiento corporal [22], [27].

3.4.3 Signos vitales

Otros enfoques incorporan medidas electrofisiológicas como electromiografía (EMG), electroencefalografía (EEG) o sensores mecánicos colocados sobre una parte del cuerpo [51]. Mediante la medición continua de datos fisiológicos como la frecuencia cardíaca, la tensión muscular, la conductancia de la piel, la frecuencia respiratoria entre otros, al combinarse con información de contexto de la actividad realizada, estos datos se pueden utilizar para inferir emociones [52].

3.5 Técnicas de reconocimiento multimodal

Debido a que cada aproximación presenta fortalezas para el reconocimiento de ciertos tipos de emociones, a continuación se presentan soluciones que incorporan diversos tipos de señales en una metodología integrada.

Múltiples fuentes de información pueden ser tenidas en cuenta para lograr un modelo del estado emocional del usuario. A esta estrategia que integra varias fuentes se le denomina reconocimiento multimodal [17], [18], [51]. Algunas investigaciones en el campo de reconocimiento de emociones multimodal durante la interacción humano-robot, utilizan dos modos: análisis de voz y de expresión facial, para aplicarlos en un robot social [19], [20], [23]. En otros casos se emplean redes neuronales con estímulos de expresión facial y movimiento corporal para el reconocimiento en escenarios sociales [18], [22], el sistema es capaz de aprender y extraer características espaciales y temporales profundas y utilizarlas en la clasificación de emociones. Estudios más amplios incluyen señales provenientes de conversación, lenguaje corporal y características faciales [53].

Las implementaciones multimodales son poco comunes [11], debido a la complejidad resultante al unir percepciones sensoriales disímiles. La integración de señales se puede efectuar mediante los siguientes métodos [54]:

1. Fusión de datos: se efectúa en los datos sin procesar y únicamente es posible cuando las señales tienen la misma resolución en el espacio temporal, es decir, no es aplicable para unir una señal de video con el texto transcrito por un componente de reconocimiento de voz (ASR).
2. Fusión de características: se efectúa sobre características interpretadas a partir de las señales, como es el caso de la media, mediana, desviación estándar, máximos y mínimos, junto con algunas características únicas de cada sensor.
3. Fusión de decisión: se realiza uniendo la salida del clasificador de cada señal. Los estados afectivos primero se clasifican por cada sensor y luego se integran para obtener una estimación general a partir del resultado encontrado en los diversos sensores. Este es el enfoque más comúnmente utilizado.

Es recomendable incorporar periodos de tiempo más largos para el reconocimiento de emociones; esto es, considerar información de contexto como la evolución de las emociones en la persona [55]. Al incluir el contexto temporal se tiende a mejorar el rendimiento de la clasificación de la emoción.

Experimentos con usuarios reales muestran una alta tasa de éxito en el reconocimiento automático de la emoción del usuario, al utilizar dos canales de información: auditiva y visual [17]. Al utilizar datos de diálogo, postura y expresión facial tomados de un sistema tutor, se alcanza una exactitud de 78.3%, mejorando en 17% el desempeño frente al mejor modelo de reconocimiento basado en un único canal de entrada [11].

En la investigación de Poria et al [56], se extrajeron Puntos de Características Faciales (FCPs), y emplearon las distancias entre esos FCPs como características, además utilizaron GAVAM [56] para extraer el movimiento de la cabeza y otras características de rotación. Para la extracción de características de audio, se empleó openEAR. El software SenticNet se utilizó para extraer características textuales, concatenaron los vectores de características de las tres modalidades, para formar un único vector. Este vector se utilizó para clasificar cada segmento de video en clases de sentimientos. En este caso, se emplearon estrategias de fusión por Características y de Decisión, alcanzando una precisión de 77%. Una vez comparados los métodos SVM, ANN y Extreme Learning Machine (ELM), finalmente se escogió

ELM por velocidad. Se utilizó un clasificador independiente para cada modalidad. La salida de cada clasificador se trató como una puntuación por emoción (3 emociones). Para el calculó final utilizaron la ecuación (1).

$$l' = \text{argmax}_i(q_1 s_i^a + q_2 s_i^v + q_3 s_i^t), \quad i = 1, 2, 3, \ldots, C \qquad \textbf{(1)}$$

En donde q_1 q_2 q_3 son pesos por modalidad, todos iguales a 1/3. Los S_i son vectores de salida de los clasificadores. En este caso, C=3, y corresponde al número de emociones. Finalmente, promedian el reconocimiento y escogen el de mayor valor. Ésta investigación indica que los estudios previos favorecen la fusión a nivel de decisión como el método preferido de fusión de datos porque los errores de los diferentes clasificadores tienden a no estar correlacionados y la metodología es independiente de las características [56].

En la investigación de Ringeval et al [57], se utilizan redes neuronales de tres tipos: LSTM-RNN, BLSTM-RNN y FF-NN. Estudiaron la influencia de diferentes tamaños de ventana de tiempo para predecir la emoción de cuatro modalidades (audio, video, ECG y EDA). Dichas ventanas varían de 0.48 s to 6.24 s con anchos de 0.48s. Desde el punto de vista del tipo de fusión, aplicaron la estrategia multimodal en dos niveles: características o decisión. Como resultado, evaluaron el interés de usar probabilidad basada en eventos como una característica para determinar cuál modalidad es más confiable en el tiempo. Para la fusión de características, todas las variables fueron concatenadas por cada marco, conservando el tamaño de ventana que proporciona el mejor rendimiento para cada modalidad. Para la fusión de decisión, utilizaron un SVR lineal. Como conclusión de este estudio, se encontró que la Fusión de Decisión generó mejor resultado tanto para Valencia como para Excitación.

El estudio de Alonso-Martin et al [17] concuerda con los trabajos antes mencionados en que la estrategia de fusión de decisión es la más apropiada para el componente de fusión multimodal. La ecuación (2), tomada de Alonso-Martin et al [17], indica el cálculo que debe hacer el agente de fusión multimodal para determinar la probabilidad condicional de que la emoción s_i sea la real, dado que el agente modal *C* identificó la emoción s_j. La matriz *M* se refiere a la matriz de confusión encontrada a partir del proceso de calibración. Como resultado, se obtiene el grado de confianza para cada emoción detectada por los diferentes agentes. Finalmente, se escoge la emoción que cuente con el mayor valor calculado según la ecuación (2).

$$p(S = s_i | S_C = s_j) = \frac{M_{ij}}{\sum_{k \in [1,n]} M_{kj}} \qquad \textbf{(2)}$$

En vista que la aproximación multimodal tiene como beneficio un mayor rango de emociones posibles para el reconocimiento y a que permite contrastar resultados a partir de diversos tipos de señal, la arquitectura para el reconocimiento multimodal de emociones que se presentará más adelante, incorpora esta técnica de reconocimiento integrada.

3.6 Software existente

Con el propósito de incluir software existente y que pueda ser utilizado en la implementación de la arquitectura, a continuación se presentan algunos componentes aplicados en trabajos de investigación y sobre los que se demostró su efectividad en el reconocimiento de emociones. Estos componentes pueden ser utilizados en la construcción de una solución integrada para el reconocimiento multimodal de emociones.

Gender and Emotion Voice and Facial Analysis (GEVA y GEFA) [17]**:** desarrollado en la Universidad Carlos III, GEVA analiza señales de audio. Utilizando árboles y reglas de decisión, genera como resultado la emoción y género identificados junto con un intervalo de confianza. La implementación de este componente se encuentra integrada con el sistema operativo ROS utilizado en el robot Baxter. Baxter es un robot de tipo humanoide con funciones como gravedad cero y múltiples sensores [24], [25], que lo hacen ideal para su uso en interacción con humanos [26], [27].

Affdex SDK: software desarrollado por Affectiva, reconoce siete emociones a partir de video del rostro, permite medir también por separado los ejes de valencia y nivel de excitación.

Computer Expression Recognition Toolbox (CERT) [48]**:** permite la clasificación de cinco emociones en tiempo real: ira, miedo, alegría, alivio y tristeza. Para el reconocimiento facial, utiliza aproximación por puntos de interés y una Support Vector Machine (SVM) para la clasificación de emociones. Adicionalmente, incluye la posición relativa de la cabeza como factor para la detección de emociones.

Synesketch [58]**:** software de código abierto para el reconocimiento de emociones a partir de texto, escrito en lenguaje Java. Si bien su base de datos se encuentra en idioma inglés, a través de un archivo de texto puede ser migrada a español.

Verbio ASR: software de reconocimiento de voz desarrollado en España y que soporta el idioma español. El costo de la licencia es inferior a US$ 1000. Cuenta con una efectividad entre el 85% y el 91%.

CMU Sphinx: software de reconocimiento de voz con capacidad de identificar el idioma español. Disponible con licencia de código abierto.

HARK Open Source Library [59]**:** software de audición especializado en robots. Con licencia de código abierto desde 2008 y utilizado en robots como el Honda ASIMO, SIG2 y Robovie R2. Implementa capacidades de reconocimiento de voz (ASR) con un gran nivel de compatibilidad.

OpenCV [60]**:** librería de código abierto en C++ con interfaces para Java que incluye múltiples algoritmos para visión. Está diseñada para la captura de imágenes en tiempo real, el procesamiento y el reconocimiento de color. Ha sido utilizado para el reconocimiento de objetos en un espacio de interacción HCI.

Con el propósito de determinar cuáles componentes de software son adecuados a los diversos tipos de modalidades de reconocimiento, estas herramientas fueron evaluadas e integradas a la arquitectura propuesta en razón a su utilidad y desempeño.

4 ARQUITECTURA Y DISEÑO

En el capítulo anterior se profundizó en el estado del arte respecto a estados mentales, modelos de emociones e investigaciones efectuadas en relación con el reconocimiento de emociones. Fueron presentadas también, técnicas y dificultades relativas al reconocimiento multimodal. A partir de los conceptos explorados y la metodología presentada en capítulos anteriores, a continuación se presenta la arquitectura propuesta que incluye estos conceptos para lograr una solución integrada.

4.1 Requerimientos funcionales y no funcionales

En el contexto presentado en la Figura 5, el robot humanoide actúa como una herramienta para el educador, es por este motivo que se requiere que el módulo de reconocimiento se integre en el ambiente educativo sin afectar la habilidad del niño para desenvolverse en las actividades del proceso, esto es, que se eviten dispositivos que limiten los movimientos del niño, o lo incomoden. Por otra parte, es necesario que el módulo de reconocimiento cuente con facilidades de integración en general para que su resultado pueda ser aprovechado por módulos de decisión y acción. En particular, que se integre en el contexto de la Figura 2, es decir, la arquitectura HRS-BDIBESA [44], en la que el estudiante interactúa con el sistema educativo para resolver un problema particular.

Adicionalmente, la respuesta del módulo de reconocimiento debe ser continua frente a los estímulos presentados por los sensores que serán utilizados para la detección de emociones. Estos sensores deben capturar un conjunto de fuentes de información que permitan llegar a desarrollar la capacidad de reconocimiento multimodal de emociones. Esto es importante para la arquitectura puesto que, de acuerdo a la investigación de Kort & Reilly [5], identificar con precisión el estado cognitivo-emotivo de un estudiante permite a los maestros proporcionar a los estudiantes una experiencia de aprendizaje eficiente y placentera.

A partir de la revisión del estado del arte, se encontró que el módulo de reconocimiento debe permitir agregar modalidades de manera flexible, dichas modalidades estarán clasificadas como: tonalidad, lenguaje, rostro, postura, de tarea o bloques y de actividad corporal o fisiológica, tal como se muestra en la Figura 9.

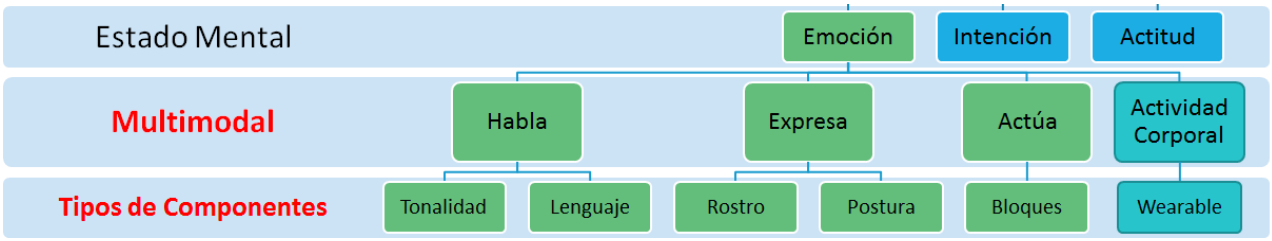


Figura 9 Identificación de requerimientos funcionales. Elaboración propia.

La definición de la arquitectura está guiada por los criterios a continuación:

1. Respuesta en línea: con el propósito de obtener una retroalimentación efectiva para el proceso educativo; es necesario que la arquitectura de reconocimiento proporcione información permanente sobre el estado emocional del niño durante la ejecución de las tareas.
2. Evitar elementos invasivos: este tipo de dispositivos resulta molesto para la ejecución de la tarea en la medida en que incomoden al niño o limiten sus movimientos.
3. Aplicabilidad en el contexto escolar: teniendo en cuenta limitaciones de espacio físico y costo limitado en los elementos hardware y software requeridos para la implementación.
4. Multimodal: como se ha mostrado en las secciones anteriores, el reconocimiento emocional involucra múltiples dimensiones; se deben tener en cuenta respuestas simultáneas e incluso contradictorias en las modalidades.

4.2 Selección de componentes

Con el propósito de evaluar el software existente que podría ser utilizado para la implementación y, que se obtenga información relevante para enfocar la arquitectura y diseño, se definieron un conjunto de criterios que permitieran filtrar el software existente. Se realizó también una prueba funcional en laboratorio que permitiera escoger los componentes disponibles en el mercado a ser usados durante la implementación del prototipo para la prueba experimental.

En las tablas a continuación, la columna descripción incluye la explicación del criterio y la forma en que se asignan los rangos para los casos diferentes a respuestas Sí o No. Para cada valor posible se le asigna un número, el valor deseado es aquel que tenga el mayor número asignado.

Los criterios de primer nivel definidos en la Tabla 1 presentan valores excluyentes sí / no con el propósito de servir como primer filtro para requisitos de mayor prioridad. A continuación se aplicaron los criterios de segundo nivel detallados en la Tabla 2, los cuales profundizan en las características de los componentes y su conveniencia de aplicación en las condiciones experimentales. Finalmente, se aplicaron los criterios de selección de tercer nivel detallados en la Tabla 3 para llegar a los componentes candidatos a la prueba funcional de laboratorio.

Tabla 1 Criterios de selección de primer nivel excluyente. Elaboración propia.

Código	Criterio	Descripción	Valores	Peso
N1C1	Respuesta en línea	Indica si es posible obtener el resultado del componente a partir de los datos de entrada en un tiempo menor a diez segundos.	(1) Sí (0) No	1/6
N1C2	Invasivo	Indica si el componente requiere elementos invasivos para medir los datos de entrada necesarios, como es el caso de electrodos que pudieran restringir la libertad de acciones del niño o causarle incomodidad.	(1) No (0) Sí	2/6
N1C3	Interés Experimental	Indica si existe interés especial en la evaluación del componente específico	(1) Sí (0) No	1/6

Código	Criterio	Descripción	Valores	Peso
N1C4	Disponibilidad de Recursos	Los recursos necesarios para evaluar el componente están o no disponibles en la Universidad Javeriana	(1) Sí (0) No	2/6

Tabla 2 Criterios de selección de segundo nivel. Elaboración propia.

Código	Criterio	Descripción	Valores	Peso
N2C1	Tipo de licencia de software	Indica el tipo de licencia bajo la que está disponible el software del componente [61]. • Código abierto permisiva: Apache, MIT, BSD, entre otras. • Código abierto menos restrictiva: LGPL. • Código abierto restrictiva: GPL. • Sin licencia. • Código cerrado.	(3) Permisiva (2) Menos (1) Restrictiva (0) Cerrada	1/7
N2C2	Velocidad de procesamiento	Indica la velocidad en que es posible obtener el resultado del componente a partir de los datos de entrada. • Alto: la respuesta se obtiene en un tiempo menor a un segundo. • Medio: la respuesta se obtiene en un tiempo menor a cuatro segundos. • Bajo: La respuesta tarda más de cuatro segundos.	(3) Alto (2) Medio (1) Bajo	1/7
N2C3	Comparativa del número de emociones reconocidas	Indica el número de emociones reconocibles por el componente evaluado frente a las posibilidades de los demás componentes del mismo tipo. Se calcula como el porcentaje entre el número de emociones reconocido por el componente frente al número máximo de emociones reconocido por los demás componentes del mismo tipo.	Entre 0 y 100%	2/7
N2C4	Posibilidad de integración con el robot Baxter	Indica si es posible o no integrar el componente con el robot Baxter, bien sea en forma directa o si es necesaria la construcción de una capa intermedia de integración.	(3) Alto (2) Medio (1) Bajo (0) Ninguno	2/7

Código	Criterio	Descripción	Valores	Peso
		• Alto: indica que el componente ha sido implementado previamente en el robot Baxter. • Medio: indica que el componente no ha sido implementado previamente en el robot Baxter pero que se considera viable su implementación directa. • Bajo: indicaría que es necesario desarrollar una capa intermedia de integración. • Ninguno: indica que el componente no se puede integrar con el robot y que se deben usar elementos externos.		
N2C5	Tiempo de implementación	Es posible construirlo en el tiempo disponible para el trabajo de grado.	(1) Sí (0) No	1/7

Tabla 3 Criterios de selección de tercer nivel. Elaboración propia.

Código	Criterio	Descripción	Valores	Peso
N3C1	Nivel de precisión de reconocimiento	Indica, en forma de porcentaje, qué tan fiel fue el reconocimiento de las emociones frente al esperado en el experimento.	Entre 0 y 100%	1/5
N3C2	Capacidad de integración multimodal	Indica si el resultado del componente permite ser utilizado para la integración multimodal. • Alto: indica que el componente resuelve directamente la probabilidad de ocurrencia de un conjunto de emociones. • Medio: indica que el resultado del componente debe ser transformado para integrarse con el reconocimiento multimodal. • Bajo: indica que no existe una forma clara en que se pueda integrar el resultado del componente.	(3) Alto (2) Medio (1) Bajo	2/5

Código	Criterio	Descripción	Valores	Peso
N3C3	Potencial educativo	Indica que las emociones reconocidas por el componente están especialmente relacionadas con el contexto educativo. • Alto: indica que el componente reconoce tres emociones o más, relevantes al contexto educativo. • Medio: indica que el componente reconoce entre dos y tres emociones relevantes al contexto educativo. • Bajo: indica que el componente reconoce una emoción relevante al contexto educativo. • Ninguno: indica que las emociones reconocidas no son relevantes para el contexto educativo.	(3) Alto (2) Medio (1) Bajo (0) Ninguno	2/5

Estos criterios fueron aplicados sobre los componentes de software listados en la Tabla 4, los cuales fueron identificados a partir del software utilizado en las investigaciones encontradas en la profundización del estado del arte y a partir de búsquedas en internet de software disponible el mercado.

Tabla 4 Componentes de software para el reconocimiento identificados. Elaboración propia.

Tipo	Componente
Voz	Gender and Emotion Voice Analysis (GEVA)
Video	Gender and Emotion Facial Analysis (GEFA)
Video	Machine Perception Toolbox
Imagen	Facial Expression Recognition
Lenguaje	Synesketch
Voz	Verbio ASR
Voz	CMU Sphinx
Voz	HARK
Voz / Emoción	UAH System
Lenguaje	SENNA
Video	Microsoft Cognitive Services
Video	SHORE
Lenguaje	Watson
Imagen	Google cognitive services
Video	Intel Realsense SDK
Video	Affectiva developer portal Affdex SDK
Video	Kairo

Como resultado de la aplicación de los criterios de selección, se efectuó la prueba funcional en laboratorio. Dicha prueba se desarrolló de la siguiente forma: con el permiso de los padres de familia, se invitaron cinco niños, con edades entre 10 y 13 años. Cada niño participaba de forma separada, un instructor le presentaba un ejercicio con cubos a resolver y las siguientes reglas:

- Un cubo puede moverse al espacio consecutivo vacío.
- Un cubo puede saltar a otro de su mismo color.
- Los movimientos no se deben regresar. Si es necesario, debe empezar de nuevo.
- El problema termina cuando todos los cubos hayan pasado al lado contrario de la escalera.

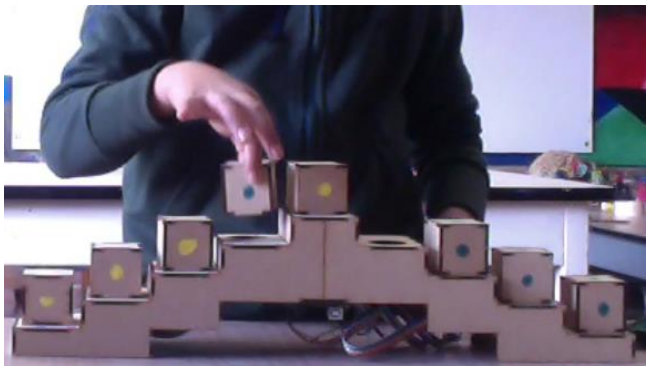


Figura 10 Imagen del ejercicio en la prueba funcional en laboratorio. Elaboración propia.

La ejecución del ejercicio se grabó con una cámara de un computador portátil y un sensor Kinect. La Figura 10 muestra una captura de imagen durante la ejecución del ejercicio. Esta información se presentó al software seleccionado a partir de los criterios de selección, de tal forma que ejecutar el reconocimiento de emociones.

Se solicitó a dos expertos en educación con niños, la evaluación de los videos de tal forma que identificaran en qué momentos (minuto y segundo) se reconocían: emociones comunes, ejes de emociones según el modelo de Kort & Reilly [5]. Cada experto detenía el video en el momento en que identificara un cambio en el estado emocional, y diligenciaba una encuesta en la que se registraban las emociones reconocibles. El contenido de la encuesta se detalla más adelante en la sección 5.3 de este documento.

Los resultados se compararon con los valores identificados por el software evaluado. La Tabla 5 muestra los valores encontrados por cada componente y emoción, al igual que el promedio global de reconocimiento para cada uno. Finalmente se escogieron los siguientes componentes a utilizar en la implementación del módulo de reconocimiento multimodal:

- Affdex SDK: Obtuvo un 81% de precisión en la prueba funcional, comparado con un resultado de 34% para Machine Perception Toolbox.
- CMU Sphinx: Obtuvo un 60% de precisión en la prueba funcional, comparado con un resultado de 51% para Hark.

- Synesketch: Obtuvo un 66% de precisión en la prueba funcional, comparado con un resultado de 37% para SENNA.

Tabla 5 Resultados de la prueba funcional en laboratorio. Elaboración propia.

Emoción / Componente	MPT	Affdex SDK	HARK	CMU Sphinx	Synesketch	SENNA
Alegría	55%	82%	71%	82%	78%	40%
Tristeza	65%	85%	56%	65%	73%	45%
Ira	60%	89%	51%	65%	82%	65%
Sorpresa	0%	75%	59%	75%	65%	30%
Miedo	10%	70%	20%	20%	20%	15%
Desagrado	12%	87%	49%	55%	75%	25%
Promedio	34%	81%	51%	60%	66%	37%

4.3 Arquitectura de alto nivel

A partir de la profundización en el estado del arte y el análisis de requerimientos, se identificaron cuatro modalidades o dimensiones para el reconocimiento multimodal de emociones:

1. Verbal: está compuesta por la tonalidad de la voz del niño y el lenguaje que utiliza al expresar su situación frente al problema que se le presenta.
2. Corporal: incluye el lenguaje no verbal que se puede captar a partir de las expresiones del niño. En esta modalidad se incluyen los gestos específicos del rostro y expresiones de postura corporal.
3. Tarea: se refiere a la forma en que el niño interactúa con el entorno en el marco del proceso educativo. Si el problema que se le está presentando incluye un conjunto de bloques a desplazar, esta modalidad contiene información sobre la forma en que el niño toma estos elementos, los mueve o juega con ellos mientras resuelve el problema.
4. Fisiológica: incluye signos vitales que pudieran medirse durante el proceso educativo. Es posible utilizar dispositivos modernos (wearables) como manillas para medir el ritmo cardiaco u otros signos que pudieran resultar relevantes.

La Figura 11 presenta la arquitectura propuesta teniendo en cuenta los criterios antes mencionados. Los agentes SMA están marcados en color azul claro. Inicialmente, existe un conjunto

de sensores necesarios para el procesamiento de las modalidades verbal, corporal, de tarea y fisiológica. Estos sensores transfieren información a los agentes representados por un rectángulo con líneas verticales en cada lado. Estos agentes están especializados por responsabilidades de acuerdo a las modalidades identificadas y contienen la inteligencia necesaria para reconocer emociones a partir de los sensores de su modalidad. El agente de Fusión de decisión implementa la técnica de cooperación de tal forma que se coordinen las salidas de los agentes especializados y se resuelvan los conflictos producto de las habilidades especiales de cada modalidad en la detección de emociones.

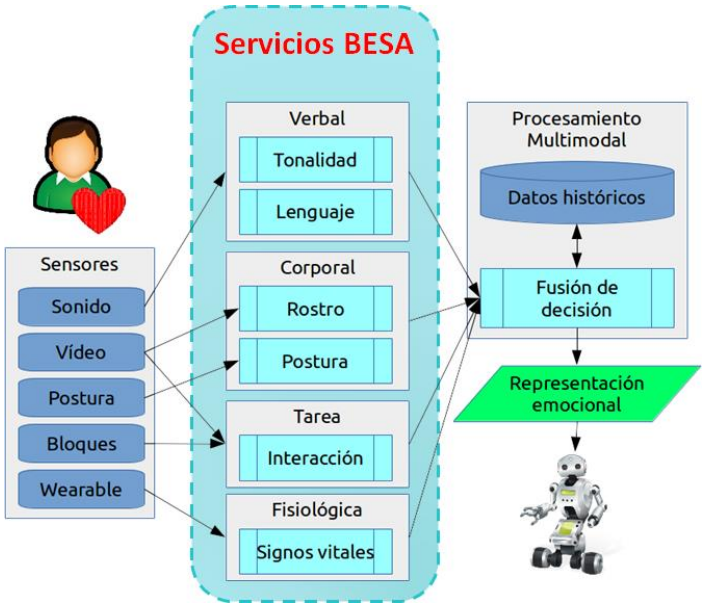


Figura 11 Arquitectura de reconocimiento multimodal. Elaboración propia.

La modalidad verbal específicamente requiere un sensor de sonido, con esta información es posible utilizar componentes existentes para el reconocimiento de emociones a partir de la tonalidad de la voz. Mediante la información de los mismos sensores de sonido y en forma paralela, se utilizan técnicas de SLU para el reconocimiento a partir del lenguaje utilizado por el estudiante durante la ejecución de las actividades.

La modalidad corporal utiliza sensores de video con el propósito de lograr un reconocimiento a partir del rostro del estudiante, por este motivo, es recomendable que existan sensores que puedan hacer seguimiento a la ubicación del niño en el entorno. Para solucionar este inconveniente, se puede utilizar más de una cámara de tal forma que el sistema logre obtener los datos faciales con la mayor continuidad posible. En la medida en que esto no fuera posible por alguna situación durante la ejecución de la actividad, el sistema es robusto, puesto que se puede apoyar en otros sensores para el reconocimiento. En relación con la postura es posible utilizar un sensor de tipo Kinect que calcule a intervalos regulares la posición relativa de la cabeza y extremidades del niño.

La modalidad de tarea es la única que requiere sensores especiales dependientes de la actividad que se está realizando. Si el problema involucra bloques para su ejecución, dichos blo-

ques podrían contener sensores de movimiento que transmitan información de velocidad, aceleración o posición relativa con respecto a los demás bloques y elementos de tal forma que el sistema educativo pueda determinar que la tarea se está adelantando en forma exitosa o que existe necesidad de intervención. En el contexto del reconocimiento de emociones, esta misma información puede ser utilizada para reflejar comportamientos relacionados con la confianza, comodidad o ansiedad con las que el niño está afrontando el problema. En la medida en que no se cuente con sensores especiales en los elementos, la información de los sensores de video podría ser utilizada para hacer el seguimiento necesario a la actividad.

La modalidad fisiológica utiliza sensores específicos que proporcionan datos de signos vitales como pulso o tensión [52], es posible utilizar dispositivos de tipo wearable que permitan al sensor realizar las medidas sin incomodar al niño en la ejecución de la tarea o afectar su concentración.

Cada una de las modalidades se procesa mediante componentes especializados para el reconocimiento de emociones a partir de información específica como es el caso de la tonalidad de la voz o la expresión facial. Las salidas de estos componentes son transmitidas en línea a otro agente especializado en realizar el procesamiento multimodal a partir de la fusión de las salidas de los componentes previos. Este agente incluye un mecanismo de memoria de tal forma que la decisión sobre el estado emocional del niño incluya información del contexto e información histórica como sería el caso de los últimos minutos de ejecución de la actividad.

El resultado del reconocimiento multimodal es la representación emocional del niño en la que, para cada una de las emociones objetivo, se produce una salida a escala con la fortaleza de la emoción detectada. Esta información es clave para que sea utilizada por el sistema educativo en el soporte metacognitivo, relacionado con el tema de estudio, y soporte emocional, relacionado con el estado emocional del niño. De esta forma, el sistema educativo contará con información relevante para tomar las acciones necesarias que logren el objetivo final del sistema, esto es, el aprendizaje por parte del niño [44].

La arquitectura utiliza una metodología de Sistema Multiagente (SMA) debido a que integra un conjunto de componentes especializados concurrentes y cooperativos [32]. Los componentes de las diversas modalidades son dinámicos y pueden incorporar software existente o el desarrollo de otros agentes específicos como es el caso de la modalidad de tarea. Incluso admite la posibilidad de incorporar nuevas modalidades en el futuro, en caso que fueran identificadas.

4.4 Diseño SMA

Para el diseño del Sistema Multiagente (SMA), se utilizó la metodología AOPOA [32]. Los agentes se instancian mediante el framework BESA [62], esta es una plataforma de desarrollo que facilita la implementación SMA. En la plataforma BESA, cada agente cuenta con guardas que le permiten actuar continuamente para alcanzar sus metas. Cuenta también con un estado particular con el que mantiene la información necesaria para desempeñar sus funciones. El agente obtiene acceso a los recursos que necesita, por ejemplo acceso al audio, cámara de video o Kinect. Con esta información y de forma continua, ejecuta la función de reco-

nocimiento que le corresponde. Las emociones reconocidas son publicadas mediante eventos de la plataforma. Estos eventos son recibidos por los agentes interesados, en particular, el agente de fusión multimodal.

La Tabla 6 lista los agentes identificados y describe los recursos requeridos para cumplir con sus metas. Estos agentes intercambiarán mensajes a través de eventos en la plataforma BESA. Los eventos están compuestos por información de las emociones identificadas, cada agente publica el resultado de reconocimiento de una emoción, una vez el estímulo de los sensores logre superar un umbral configurado. Estos eventos son recibidos por el agente de reconocimiento multimodal mediante un patrón de publicador / suscriptor, por cuanto se logra bajo acoplamiento de los componentes. De esta forma, es posible agregar nuevos agentes de reconocimiento modal a la arquitectura, sin modificar la lógica de los demás agentes ni la de fusión multimodal.

Tabla 6 Caracterización de agentes. Elaboración propia.

Agente	**Habilidades**	**Recursos**
Tonalidad	Detecta emociones a partir del tono de la voz	Sensor de audio
Lenguaje	Detecta emociones a partir de lo que el estudiante dice	Sensor de audio
Rostro	Detecta emociones a partir de la expresión del rostro del estudiante	Sensor de video
Postura	Detecta emociones a partir de la posición de los brazos y cabeza del estudiante	Sensor Kinect
Interacción	Detecta emociones a partir de la manipulación de los bloques o elementos específicos de la tarea	Sensor con funcionalidad de tracking (Intel RealSense)
Signos vitales	Detecta emociones a partir de signos vitales especializados. En este caso se utilizó la medición de pulso a partir del dispositivo Mi Band 2, sin embargo, este agente estará especializado conforme a las características del dispositivo.	Sensor de pulso Mi Band 2
Fusión de decisión	Encargado de implementar la fusión de decisión a partir de las emociones de los demás agentes. Genera la representación emocional como salida del módulo de reconocimiento.	Eventos de reconocimiento de emociones generados por los demás agentes. Datos históricos de eventos de reconocimiento proporcionados por los agentes especializados.

La Figura 12 muestra la secuencia de interacciones en el sistema. Los sensores envían la información recibida a través de streams a los agentes modales de acuerdo a los recursos definidos en la Tabla 6, cada agente modal procesa el flujo de datos de acuerdo a su lógica de implementación, por ejemplo, mediante la ejecución de Affdex SDK para el agente de rostro, CMU Sphinx y Synesketch para el agente de lenguaje o la lógica específica de los agentes de interacción y postura. Las emociones detectadas se transmiten mediante eventos con marcas

de tiempo a la plataforma BESA. El agente de procesamiento multimodal recibe los eventos de manera asíncrona y los registra en su almacén de datos históricos, mediante la lógica de fusión, se genera un evento de emoción detectada que puede ser leído e interpretado por el sistema externo de scaffolding para tomar las acciones programadas en el proceso educativo.

Cada agente modal es responsable de la administración de las ventanas de tiempo requeridas por la lógica interna de implementación, el agente de reconocimiento multimodal recibe los eventos generados y los registra en su almacén de datos, con el fin de ejecutar su lógica interna para la toma de decisión. Con cada evento que llega, el agente de reconocimiento multimodal recalcula el estado del sistema y genera un nuevo evento de reconocimiento dirigido al sistema externo.

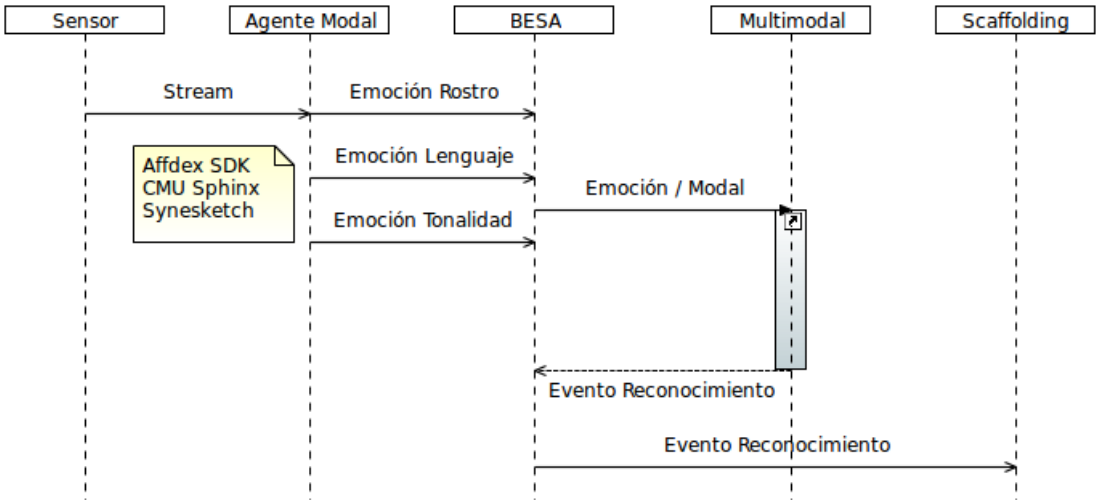


Figura 12 Diseño de interacciones. Elaboración propia.

4.5 Diseño del agente de fusión multimodal

A continuación, se presenta el diseño del agente de decisión multimodal. Para este diseño se consideraron los trabajos de Poria et al [56], Ringeval et al [57] y Alonso-Martin et al [17] , que fueron descritos anteriormente en la sección 3.5 de este documento. El diseño de nuestro componente se basa en la estrategia propuesta por Alonso-Martin et al [17]. El proceso de cálculo consta de dos etapas: calibración y ejecución. La Figura 13 muestra los pasos que componen ambas etapas. En la primera, utilizando la matriz de confusión calculada a partir de pruebas previas, se calculan los factores de confianza de cada agente modal en relación con las diferentes emociones. En este caso se utilizaron los datos de la prueba de laboratorio para calcular las matrices de confusión y los factores de confianza. De esta forma, se obtiene una matriz que será utilizada durante la ejecución del módulo.

Durante la etapa de ejecución, a partir de los eventos de reconocimiento detectados por los agentes modales, se envían mensajes al agente multimodal indicando las emociones reconocidas y la valencia con que fueron identificadas. Una valencia es un número entre cero y uno que representa la fortaleza con la que el agente determina que está ocurriendo una emoción.

Con esta información, se realiza un proceso de selección ponderando las salidas de los agentes mediante la matriz de factores de confianza previamente calculada. El resultado de dicho cálculo, se filtra mediante un criterio de umbral para generar una decisión final del agente multimodal. Esta decisión es un vector que incluyen las diferentes emociones y una valencia con un número entre cero y uno. Debido al umbral aplicado, se reduce a cero la valencia de aquellas emociones que no fueron reconocidas con la fortaleza suficiente, esto es, se separan los eventos de reconocimiento y aquellos en que no aplican las diferentes emociones. Esto permite que el módulo de reconocimiento genere un resultado que será utilizado por el sistema educativo para tomar acciones en relación con las emociones que hayan sido identificadas.

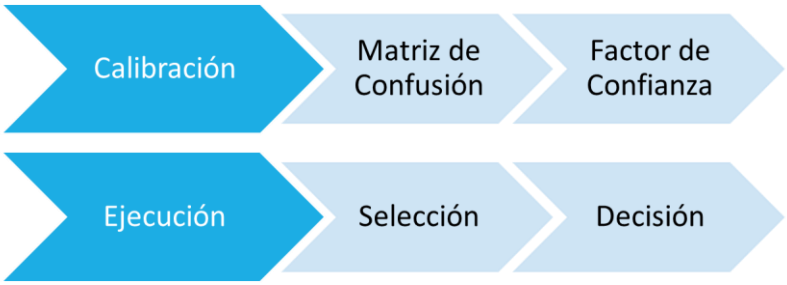


Figura 13 Etapas de cálculo para la fusión multimodal. Elaboración propia.

El diseño del mecanismo de fusión multimodal se describe a continuación:

- Calibración: a partir de un experimento previo, se calculan las matrices de confusión para cada agente modal. Las matrices de confusión se calculan comparando el resultado generado por el sistema frente a lo indicado por el panel de expertos en los diferentes puntos de control. Para este experimento se utilizaron los resultados de la prueba funcional en laboratorio para calcular las matrices de confusión de cada agente modal, la Tabla 7 muestra un ejemplo de matriz de confusión para el agente modal de rostro.

Tabla 7 Matriz de confusión en la calibración del agente modal de rostro. Elaboración propia.

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
Agente Modal Rostro	Alegría	82	0	0	13	0	0
	Tristeza	0	85	0	0	8	0
	Ira	0	0	89	0	0	10
	Sorpresa	18	0	0	75	15	0

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
	Miedo	0	10	6	12	70	3
	Desagrado	0	5	5	0	7	87

- Fusión de decisión multimodal: se escogió esta estrategia en consecuencia con los resultados de las investigaciones previamente mencionadas. El factor de confianza de la salida de cada agente de reconocimiento modal se calcula utilizando el teorema de Bayes y las matrices de confusión calculadas en el proceso de calibración. La ecuación (3), adaptada de Alonso-Martin et al [17], indica el cálculo que debe hacer el agente de fusión multimodal para determinar el factor de confianza f_{ij} que se debe utilizar para cada agente *j* ante la emoción *i*. En donde *n* es el número total de emociones reconocibles y la matriz *M* se refiere a la matriz de confusión encontrada a partir del proceso de calibración para el agente en particular. Como resultado, se obtiene el factor de confianza para cada emoción detectada por los diferentes agentes.

$$f_{ij} = \frac{M_{ii}}{\sum_{k\in[1,n]} M_{ik}} \quad \textbf{(3)}$$

Esta ecuación se obtiene al aplicar el teorema de Bayes y la definición de Kolmogorov para determinar la probabilidad de ocurrencia de una emoción s_i dado que el agente *i* haya reconocido la emoción s_j tal como se muestra en las ecuaciones (4) a la (7) originalmente presentadas por Alonso-Martin et al [17].

$$p(S = s_i | S_C = s_j) = p(S_C = s_j | S = s_i) \frac{p(S = s_i)}{p(S_C = s_j)} \quad \textbf{(4)}$$

$$p(S = s_i | S_C = s_j) = \frac{p(S = s_i) p(S_C = s_j | S = s_i)}{\sum_{k\in[1,n]} p(S_C = s_j \cap S = s_k)} \quad \textbf{(5)}$$

$$p(S = s_i | S_C = s_j) = \frac{p(S = s_i) p(S_C = s_j | S = s_i)}{\sum_{k\in[1,n]} p(S_C = s_j | S = s_k) p(S = s_k)} \quad \textbf{(6)}$$

$$p(S = s_i | S_C = s_j) = \frac{p(S = s_i) M_{ij}}{\sum_{k\in[1,n]} M_{kj} p(S = s_k)} \quad \textbf{(7)}$$

En tanto que todos los estados son igualmente probables, la ecuación (7) se simplifica a la presentación de la ecuación (8), de donde se obtiene la base para el cálculo de los factores de confianza presentado en la ecuación (3).

$$p(S = s_i | S_C = s_j) = \frac{M_{ij}}{\sum_{k\in[1,n]} M_{kj}} \quad \textbf{(8)}$$

- Selección: durante la ejecución de los eventos de decisión, se realiza una suma ponderada de la valencia de las emociones resultantes de los diferentes agentes. La ecuación (9) muestra el cálculo para obtener los pesos w_{ij} que serán aplicados como ponderadores en la ecuación (10). El cálculo se realiza para cada emoción *i* a partir de los factores de confianza f_{ij} identificados durante la calibración. La cantidad *m* corresponde al número de agentes disponibles para la emoción y los valores v_{ij} son las valencias resultantes para la emoción *i* por parte del agente *j*. En caso que el cálculo resultante s_i sea inferior a un umbral configurable, por ejemplo 0.3, la emoción no se incluye como resultado de la identificación del módulo.

$$w_{ij} = \frac{f_{ij}}{\sum_{l\in[1,m]} f_{il}} \quad \textbf{(9)}$$

$$S_i = \sum_{k\in[1,m]} w_{ik} v_{ik} \quad \textbf{(10)}$$

A continuación se presenta un ejemplo aplicando el algoritmo de decisión multimodal. A partir del resultado obtenido en la prueba de laboratorio y utilizando las matrices de confusión de la Tabla 7 para el agente modal de rostro y la Tabla 8 para el agente modal de lenguaje, se calculan los factores de confianza utilizando la ecuación (3), el resultado se muestra en la Tabla 9. Estos resultados indican que el agente de rostro tendrá un peso mucho mayor para la emoción de tristeza, miedo y desagrado, mientras que la diferencia es pequeña para ira y sorpresa. Para la emoción de alegría, el agente modal de lenguaje tiene un peso ligeramente superior.

Tabla 8 Matriz de confusión en la calibración del agente modal de lenguaje. Elaboración propia.

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
Agente Modal Lenguaje	Alegría	78	0	0	10	0	0
	Tristeza	8	73	5	0	19	6

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
	Ira	7	7	82	0	0	0
	Sorpresa	7	8	0	65	15	12
	Miedo	0	12	3	10	20	7
	Desagrado	0	0	10	15	46	75

Tabla 9 Cálculo de factores de confianza para el ejemplo. Elaboración propia.

	Rostro	Lenguaje
Alegría	86	89
Tristeza	91	66
Ira	90	85
Sorpresa	69	61
Miedo	69	38
Desagrado	84	51

Durante la ejecución del módulo de reconocimiento, los agentes generan como resultado un vector con las valencias para cada emoción *i*. Esta información la recibe el agente multimodal que calcula la valencia s_i ponderada a partir de los factores de confianza identificados en el paso anterior. La Tabla 10 presenta un ejemplo en el que el agente de rostro identificó la emoción alegría con valencia 0.7 y la emoción sorpresa con valencia 0.4. A su vez, el agente de lenguaje identificó la emoción alegría con valencia 0.55 y la emoción miedo con valencia 0.5. La columna s_i muestra el resultado de la ponderación aplicando la ecuación (10). Al aplicar un umbral de 0.3 sobre las valencias resultantes, el agente multimodal identifica para este ejemplo, la emoción Alegría con una valencia de 0.62.

Tabla 10 Cálculo multimodal para el ejemplo. Elaboración propia.

	Rostro	Lenguaje	S_i	Resultado Multimodal
Alegría	0.7	0.55	0.62	0.62
Tristeza	0	0	0	0
Ira	0	0	0	0
Sorpresa	0.4	0	0.21	0
Miedo	0	0.5	0.18	0
Desagrado	0	0	0	0

4.6 Diseño del agente modal de interacción / tarea

Dado que el experimento presentado utiliza un problema que requiere manipular un conjunto de bloques, se diseñó de manera especial un agente que pueda reconocer emociones correspondientes al modelo de Kort & Reilly [5] utilizando técnicas de lógica difusa. La Figura 14 presenta los pasos para la implementación de este agente específico.

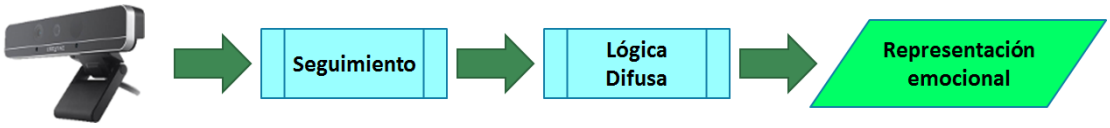


Figura 14 Secuencia del agente modal de tarea

El agente mantiene un ciclo de seguimiento permanente a los bloques que representan en el problema. Cada bloque es identificado y se le hace seguimiento a su estado de reposo / movimiento. Esto se logra utilizando una cámara especial Intel RealSense que cuenta con funcionalidades de tracking de objetos. El seguimiento genera, para cada bloque identificado, variables de posición, velocidad, aceleración y tiempo en movimiento. Esta información se pasa a un conjunto de reglas de lógica difusa de tal forma que se obtenga como resultado una representación del estado emocional del estudiante.

Tabla 11 Reglas de juicio de experto para el agente modal de interacción

Regla	**Emoción**
El bloque presenta una trayectoria continua con velocidad constante	Comodidad / Confianza

Regla	**Emoción**
Existen cambios de trayectoria	Duda
No existe movimiento durante un tiempo mayor a 4 segundos	Confusión
Bloque fuera del estado de reposo durante más de 4 segundos	Confusión
No existe movimiento durante un tiempo mayor a 16 segundos	Tedio
No existe movimiento durante un tiempo mayor a 8 segundos	Aburrimiento
Movimiento oscilatorio	Interés
Trayectoria con destino final fuera del espacio de trabajo	Frustración
Movimiento errado en el contexto del problema o bloque se suelta después de tomarlo	Curiosidad

La Tabla 11 presenta las reglas identificadas mediante juicio de experto para una implementación del agente de reconocimiento de emociones a partir de la ejecución de la tarea. Si el bloque presenta una trayectoria continua con velocidad constante, la regla de lógica difusa indica una representación emocional de "comodidad". En caso que existan cambios de trayectoria en la manipulación del bloque, la regla indica que una emoción de "duda". En caso que no exista movimiento de los bloques durante un tiempo configurable mayor a 10 segundos, o que hay un bloque fuera del estado de reposo durante este tiempo, se considera una emoción de "confusión". En caso que el estudiante deposite el bloque fuera del espacio de trabajo, se considera una emoción de "frustración".

Las reglas fueron definidas a partir de un juicio de experto y pueden ser modificadas en la implementación, de tal forma que, mediante la aplicación de técnicas de lógica difusa, es posible representar el conocimiento de juicio de experto y medir el resultado en la ejecución de la tarea. Sin embargo, si la definición del problema cambia, es necesario verificar este conjunto de reglas para adaptarlo al nuevo contexto del proceso.

En caso que varios conjuntos difusos se activen al mismo tiempo para una variable, se utiliza el método de centro de masa para la "defuzzificación". La implementación de lógica difusa requiere que se configuren tablas de inferencia dadas las variables definidas. La Tabla 12 presenta las inferencias configuradas para las variables tiempo en movimiento en las columnas y velocidad en las filas.

Tabla 12 Tabla de inferencia velocidad vs tiempo en movimiento. Elaboración propia.

	Seguido	Separado	Muy Separado
Rápido	Euforia	Confianza	Indiferencia
Medio	Revelación	Comodidad	Aburrimiento
Lento	Confusión	Incomodidad	Tedio

5 PRUEBA EXPERIMENTAL

Este capítulo presenta el caso de estudio en laboratorio que fue diseñado para realizar la prueba de concepto del módulo de reconocimiento multimodal a través de un montaje experimental de laboratorio utilizando el robot Baxter. Inicialmente se describe el caso de estudio, luego se detalla el protocolo experimental seguido y el grupo que participó en la ejecución de la prueba. Finalmente, se detalla el mecanismo de validación a partir de los resultados de la prueba utilizando información recolectada mediante un panel de expertos.

5.1 Caso de estudio de laboratorio

El experimento realizado incluye la participación de un instructor que, con ayuda del robot Baxter, presenta un problema con bloques al estudiante. El problema seleccionado se denomina "el juego de la rana"[63] y se describe a continuación:

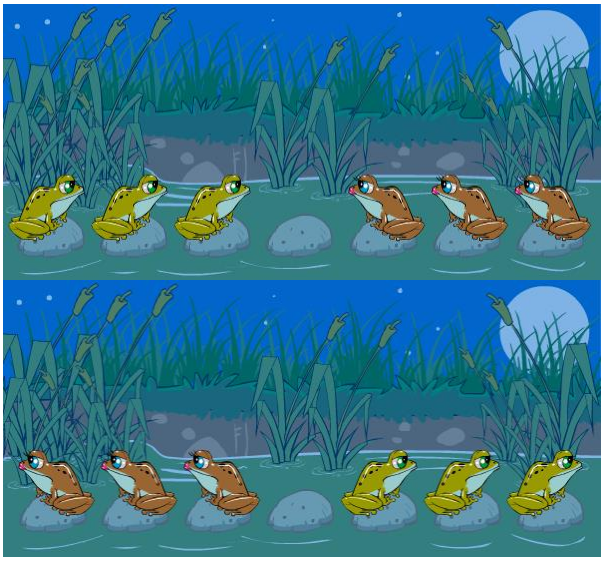


Figura 15 Ilustración de "el juego de la rana". Disponible en [64].

Existen dos grupos de ranas, todas las ranas de un mismo grupo tienen el mismo color. El objetivo es pasar las ranas al otro lado del tablero, tal como se muestra en la Figura 15. Cada rana puede moverse una posición hacia adelante o saltar sobre otra rana. No es posible mover hacia atrás, siempre se deben mover hacia adelante. Aunque parezca un ejercicio sencillo, se llega a posiciones de bloqueo que obligan a reiniciar el juego hasta encontrar la solución correcta.

Este ejercicio puede modelarse con un número igual de ranas a cada lado (n), sobre una cuadrícula de $2n + 1$ posiciones y con figuras del mismo color. Con el propósito de despertar el interés del estudiante, inicialmente se le presenta el ejercicio con cuatro ranas a cada lado. Una vez encuentra dificultades para hallar la solución, se le anima a resolver el problema con

una complejidad menor, en este caso, dos ranas a cada lado, luego con tres ranas y, finalmente, el problema completo con las cuatro ranas.

Adicionalmente, el experimento involucra la participación de un robot humanoide, en este caso, el robot Baxter. Este es un robot producido por la compañía Rethink Robotics, con capacidades de gravedad cero que permiten la interacción y fácil posicionamiento de sus brazos [24]. Cuenta con dos brazos, cada uno con siete grados de libertad en sus movimientos. El robot es programable y cuenta con un sistema operativo ROS [27].

Figura 16 Robot Baxter. Elaboración propia.

El robot fue originalmente diseñado para trabajar lado a lado con personas sin que resulte peligroso en su interacción. Incluye cuatro ruedas que facilitan su posicionamiento y cuatro fijadores de posición que anclan el dispositivo al suelo durante la ejecución de las tareas. Los brazos tienen un alcance de un metro veinte centímetros y cada uno incluye una cámara de video. Cuenta con una pantalla superior que permite mostrar imágenes representando estados emocionales [65]. La capacidad del robot Baxter de interactuar en un entorno que resulte atractivo para los niños y, a la vez, que no revista peligros en el movimiento, lo hace ideal en la interacción requerida por la arquitectura HRS-BDIBESA [44].

Dadas las condiciones descritas para el juego, se programó el robot para que moviera un conjunto de bloques de color rojo y verde, cada uno representando un grupo de ranas para el juego. La Figura 17 muestra el robot ejecutando los movimientos. Los elementos del juego se construyeron mediante cubos, de tal forma que facilitara el agarre por parte del robot y se desplegó sobre un tablero con nueve espacios para la disposición de los bloques durante el

juego. El tablero se fijó a una mesa de tal forma que la interacción con el niño no afectara la distancia requerida por el robot para los movimientos.

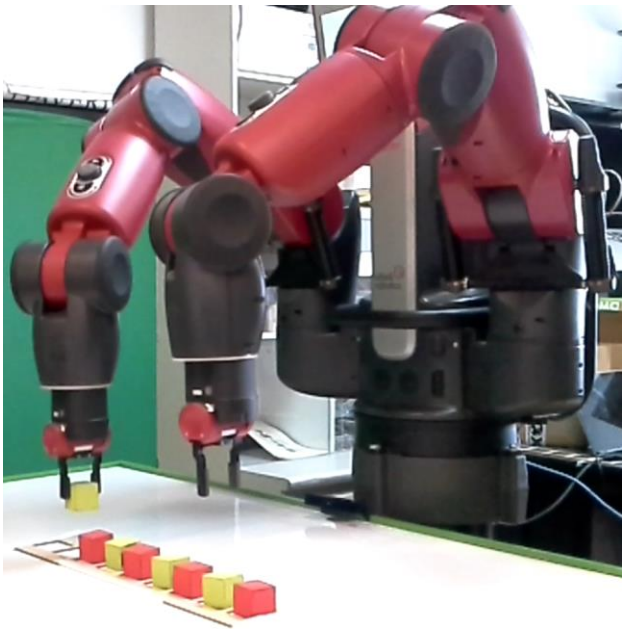


Figura 17 Robot Baxter ejecutando "el juego de la rana". Elaboración propia

5.2 Protocolo experimental

En esta sección se describen los pasos efectuados durante la realización del experimento. Para ello, se invitó un grupo de niños con edades de entre 10 y 13 años y se solicitó aprobación escrita por parte de sus padres para la participación en la actividad como parte de la clase de tecnología. El trabajo se efectuó en el laboratorio de robótica de la Pontificia Universidad Javeriana, que fue acondicionado para realizar los pasos necesarios. Los niños no conocían previamente el problema que iban a resolver, solamente se les había compartido que iban a participar en una actividad con un robot. Mientras los niños pasaban a la ejecución del experimento, los demás niños hicieron parte de una actividad grupal y se les pidió no compartir la experiencia con sus compañeros en tanto que pasaban al laboratorio. Los roles presentes durante cada instancia fueron:

1. Instructor: guía la ejecución de la actividad. Es una persona especializada en educación, en este caso el rol lo desempeñó el estudiante de doctorado John Jairo Páez Rodríguez.
2. Estudiante: es el niño que participa del proceso educativo y quien recibe la guía durante el proceso de aprendizaje, en este caso, los conocimientos para la solución del juego de la rana.
3. Robot Baxter: ayuda durante la ejecución de la tarea e interactúa con el estudiante brindándole retroalimentación sobre el problema y ofrece ayuda en caso que el niño lo requiera.

Los pasos que se siguieron fueron los siguientes:

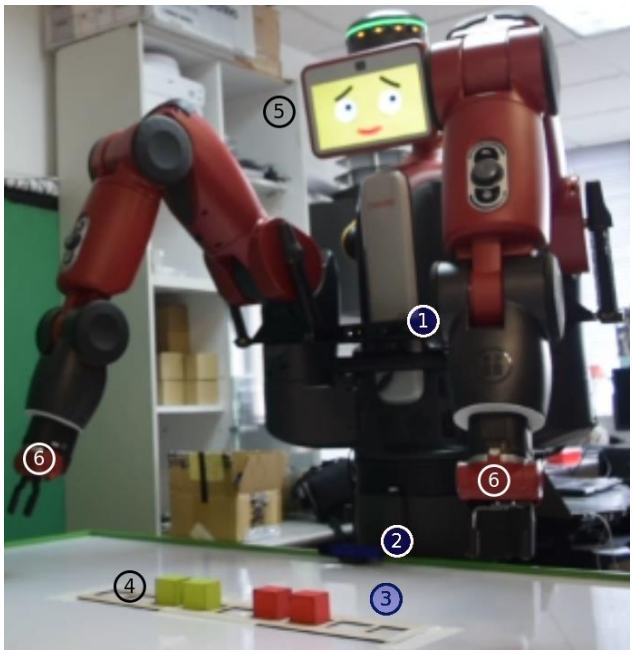
1. El niño pasa al laboratorio y el instructor le presenta al robot Baxter.
2. El instructor presenta las reglas del juego de la rana y los movimientos posibles con los bloques rojos y verdes.
3. El instructor le indica al niño que estará atento en un escritorio a la vista del niño.
4. El instructor activa el robot Baxter quien da la bienvenida al niño y le ofrece su ayuda durante la ejecución de la tarea.
5. El robot le solicita al niño por favor no mover los bloques cuando le vaya a brindar ayuda para evitar accidentes y lo invita a iniciar la solución del problema con dos bloques.
6. En la medida en que el niño se encuentra con problemas para la solución, solicita ayuda al robot diciéndole "Baxter ayúdame".
7. En este prototipo el robot todavía no reconoce los comandos del niño pero, el instructor, que está atento, activa al robot para realizar el movimiento que el niño necesita bien sea realizando un movimiento para avanzar, corregir alguna jugada inválida o solicitarle al niño que inicie nuevamente.
8. Una vez el niño terminaba el problema con dos bloques, el robot lo invitaba a resolverlo con tres bloques y, finalmente, con las cuatro disponibles.
9. Se prepararon un conjunto de frases a utilizar durante la interacción, por ejemplo, la presentación del robot, "terminé, ahora puedes continuar", "veo que estás aprendiendo", expresiones como "ajá!!" para mostrar que el movimiento fue correcto o sonidos como "mmmm" cuando el movimiento fuera inválido. El instructor comanda al robot para que ejecute dichas frases para que el niño sintiera una interacción más elaborada por parte del robot.
10. Adicionalmente, el instructor podía cambiar la expresión de la pantalla del robot para que mostrara felicidad o descontento.
11. Debido a limitaciones de tiempo, la actividad terminaba en un tiempo máximo de quince minutos.

Para la ejecución del experimento, se prepararon cámaras y dispositivos para grabar la interacción y que sirvieran como sensores para la detección de emociones por parte del niño durante el proceso de aprendizaje.

- Cámara Intel RealSense: esta es una cámara especializada que facilita el seguimiento al movimiento de los bloques. Es una cámara pequeña de 10 cm de ancho por 1 cm de altura y que se ubicó al mismo nivel que los bloques del juego como sensor del "agente modal de interacción / tarea". Esta cámara también capta la expresión del estudiante para el agente modal de rostro.
- Dispositivo Kinect: este dispositivo se ubicó entre los brazos del robot Baxter, a una mayor altura que la Intel RealSense, de tal forma que pudiera captar la posición de los brazos y cabeza del estudiante y sirviera como sensor para el agente de postura.
- Xiaomi Mi Band 2: este es un wearable en forma de manilla con capacidad de medir el pulso del niño. Cuenta con una aplicación para el celular que recopila los datos medidos. Los datos así obtenidos, se usaron como entrada al agente modal fisiológi-

co. A cada niño se le solicitó ubicar la manilla en su mano derecha durante la ejecución del experimento.

- Micrófono: mediante este dispositivo se grabaron los sonidos del experimento. Este sensor es la entrada para los agentes modales de tonalidad y de lenguaje.
- Video lateral: la ejecución del experimento se grabó con la cámara integrada de un MacBook Pro, esta cámara se ubicó de tal forma que pudiera captar las expresiones del niño y la interacción con los bloques, se utilizó para presentar los resultados de la detección de emociones al juicio de expertos.
- Cámara de Video posterior: esta cámara también se ubicó en forma lateral pero de espaldas al niño, de tal forma que pudiera captar de mejor manera la interacción con el robot y que los expertos tuvieran un punto de vista adicional a la cámara de video lateral.



1. Dispositivo Kinect
2. Cámara Intel RealSense
3. Mesa de trabajo
4. Bloques (ranas)
5. Expresión del robot
6. Pinzas del robot

Figura 18 Ubicación de dispositivos y bloques. Elaboración propia

La Figura 18 muestra la ubicación del dispositivo Kinect entre los brazos del robot y la cámara Intel RealSense al mismo nivel de los bloques para captar el movimiento de los mismos. A su vez, la Figura 19 muestra los dispositivos involucrados, la imagen corresponde al punto de vista de la cámara RealSense. El wearable se ubica en la mano derecha del niño.

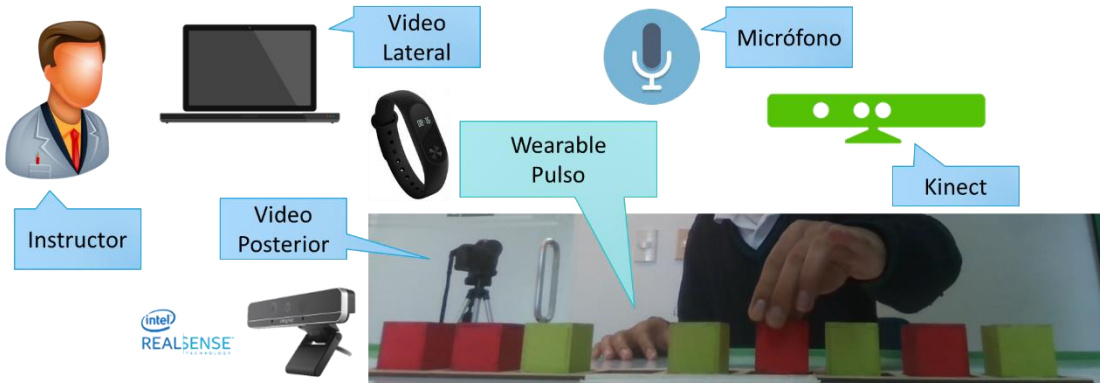


Figura 19 Despliegue experimental. Elaboración propia

5.3 Validación

El experimento se realizó en dos sesiones de medio día cada una. En la primera sesión participaron tres niñas y un niño, en la segunda sesión participó un grupo de ocho niños. Los niños de la primera sesión fueron invitados y su participación se utilizó como piloto para ajustar la posición de las cámaras, mejorar las expresiones del robot y los diferentes elementos del despliegue experimental. Los niños en la segunda sesión eran estudiantes de grados 6, 7 y 8 de educación secundaria del Gimnasio los Robles de Bogotá. Los estudiantes hacen parte del grupo de robótica del colegio y fueron invitados al experimento como parte de su proceso educativo. Las edades de los niños varían entre los 11 y los 13 años. Como se aclaró antes, se contó con la aprobación por parte de los padres de familia y del colegio para participar en la actividad.

La validación del experimento se realizó mediante un panel de expertos. Para ello, fueron invitadas a evaluar los resultados dos personas expertas en educación:

- Experto 1: Psicóloga con especialización en psicología clínica y desarrollo infantil y especialización en psicología clínica y autoeficacia personal. Cuenta con 16 años de experiencia como psicóloga y se ha enfocado en el trabajo con niños.
- Experto 2: Licenciada en preescolar con 13 años de experiencia. Durante los últimos 10 años se ha desempeñado como profesora del Colegio Anglo Americano de la ciudad de Bogotá.

A cada experto se le presentaron los videos de las grabaciones de la segunda sesión, junto con una encuesta diseñada para recopilar su percepción durante la ejecución de la actividad. La encuesta fue desplegada en Google Drive y los datos fueron recibidos en formato de hoja de cálculo. La encuesta solicita que se ingresen los datos a intervalos regulares máximo de dos minutos o en intervalos menores si el evaluador detecta cambios emocionales o si considera que hubo un cambio importante para señalar. La información de la encuesta se presenta a continuación, todas las respuestas son obligatorias en los diferentes intervalos:

1. Minutos / segundos (Tipo duración): Posición: Minutos:Segundos en el vídeo, por ejemplo :21:40 para minuto 21 segundo 40.
2. Alegría (Tipo escala lineal de 0 a 5): 1 es un estado neutral 5 es un estado de emoción. 0 significa No Aplica o no es reconocible.
3. Tristeza (Tipo escala lineal de 0 a 5): 1 es un estado neutral 5 es un estado de emoción. 0 significa No Aplica o no es reconocible.
4. Ira (Tipo escala lineal de 0 a 5): 1 es un estado neutral 5 es un estado de emoción. 0 significa No Aplica o no es reconocible.
5. Sorpresa (Tipo escala lineal de 0 a 5): 1 es un estado neutral 5 es un estado de emoción. 0 significa No Aplica o no es reconocible.
6. Miedo (Tipo escala lineal de 0 a 5): 1 es un estado neutral 5 es un estado de emoción. 0 significa No Aplica o no es reconocible.
7. Desagrado (Tipo escala lineal de 0 a 5): 1 es un estado neutral 5 es un estado de emoción. 0 significa No Aplica o no es reconocible.
8. Indiferencia / Curiosidad (Seleccionar una opción a partir de las descripciones):
    a. Tedio
    b. Aburrimiento
    c. Indiferencia
    d. Interés
    e. Curiosidad
    f. Fascinación
    g. No Aplica / No Reconocible
9. Ansiedad / Confianza (Seleccionar una opción a partir de las descripciones):
    a. Ansiedad
    b. Preocupación
    c. Incomodidad
    d. Comodidad
    e. Optimismo
    f. Confianza
    g. No Aplica / No Reconocible
10. Frustración / Euforia (Seleccionar una opción a partir de las descripciones):
    a. Frustración
    b. Vacilante
    c. Confusión
    d. Intuición
    e. Revelación
    f. Euforia
    g. No Aplica / No Reconocible

La encuesta fue diseñada para recopilar tanto las emociones comunes (alegría, tristeza, ira, sorpresa, miedo, desagrado), como los ejes emocionales del modelo especializado en educación, tal como se presentó en el capítulo del estado del arte, sección 3.3. La Figura 20 muestra la distribución del formulario de la encuesta para cuatro preguntas. La primera pregunta ubica el momento (minutos y segundos) en que ocurre la percepción de la emoción, las preguntas a continuación determinan la sensación respecto a las emociones comunes y, finalmente, se solicita la percepción en cada uno de los tres ejes emocionales del modelo de Kort & Reilly [5].

Los resultados de la encuesta fueron tabulados y comparados con el resultado del módulo de reconocimiento en los diferentes experimentos. Los capítulos a continuación detallan el análisis de datos y las conclusiones del trabajo.

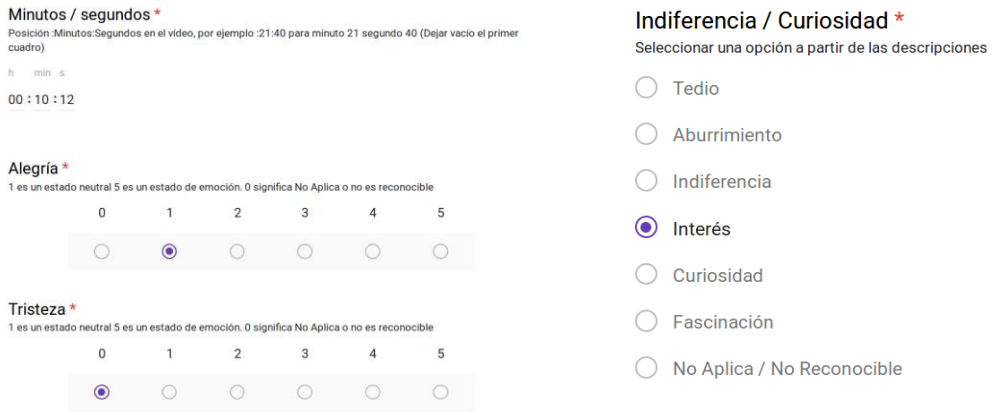


Figura 20 Fragmento del formulario de la encuesta. Elaboración propia.

6 ANÁLISIS DE DATOS

A partir de los resultados de la prueba experimental presentada en el capítulo anterior, a continuación se presentan los resultados analizando los datos recolectados para los diferentes agentes modales y, en general, para el módulo de reconocimiento multimodal. El prototipo implementado utiliza los agentes modales de rostro, postura, lenguaje y de interacción / tarea, al igual que el agente de procesamiento multimodal. Estos resultados guían la elaboración de las conclusiones y recomendaciones de trabajo futuro que serán desarrolladas al final del trabajo.

Como parte del análisis de datos que se realizará durante el capítulo, se debe calcular el porcentaje de precisión de los agentes en relación con las emociones del modelo de Kort & Reilly [5]. Este cálculo se realiza asignando valencias a cada emoción con valores entre -1 y 1, tal como se presenta en el artículo original. La precisión se mide como una medida de la distancia usando la ecuación (11).

$$precisión = \frac{|v_p - v_e|}{2} \quad \textbf{(11)}$$

En dónde v_p es la valencia percibida por el módulo de reconocimiento y v_e es la valencia de la emoción determinada a través del panel de expertos.

6.1 Agente modal de rostro

Tal como se presentó en la sección 4.2 Selección de componentes, el componente escogido para la prueba experimental fue el Affectiva Affdex SDK. Para ello, se implementó la clase affdex::VideoDetector con los siguientes parámetros:

- processFrameRate: 25, este valor corresponde a la tasa de fotogramas proveniente de la cámara Intel RealSense.
- maxNumFaces: 1, este valor se debe a la presencia de un único estudiante.

El parámetro faceConfig se utilizó como variable independiente para confirmar la variabilidad de los resultados de reconocimiento frente a los dos valores posibles:

- FaceDetectorMode.LARGE_FACES
- FaceDetectorMode.SMALL_FACES

La implementación a partir del stream de la cámara también se puede hacer con la clase affdex::CameraDetector que recibe los mismos parámetros. Como resultado, se obtuvo la matriz de confusión de la Tabla 13 para el agente modal en el caso de la configuración mediante FaceDetectorMode.SMALL_FACES, y los resultados de la

Tabla 14 para la configuración mediante FaceDetectorMode.LARGE_FACES. Se encuentra que, bajo las condiciones de la prueba experimental efectuada, el resultado es mejor con la

configuración SMALL_FACES para las emociones de alegría, tristeza, ira y desagrado. Por otra parte, la configuración LARGE_FACES resultó más precisa para el reconocimiento de la sorpresa y tuvo una leve mejora para la emoción de miedo.

Tabla 13 Matriz para el agente modal de rostro con SMALL_FACES. Elaboración propia.

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
Agente Modal Rostro	Alegría	83	0	0	18	0	0
	Tristeza	0	82	0	0	12	0
	Ira	0	0	79	0	0	15
	Sorpresa	17	0	0	75	12	5
	Miedo	0	18	7	7	63	8
	Desagrado	0	0	14	0	13	72

Tabla 14 Matriz para el agente modal de rostro con LARGE_FACES. Elaboración propia.

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
Agente Modal Rostro	Alegría	78	0	0	12	0	0
	Tristeza	0	77	0	0	9	7
	Ira	0	0	72	0	0	11
	Sorpresa	22	7	5	83	11	9
	Miedo	0	16	11	5	65	5

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
	Desagrado	0	0	12	0	15	68

Estos resultados sugieren que las condiciones del experimento son relevantes para la configuración a seleccionar, es decir, dada la posición de la cámara RealSense que es la fuente de datos para este agente modal, el rostro del niño toma una parte pequeña de la escena por cuanto la configuración SMALL_FACES adquiere relevancia como parámetro para un mejor reconocimiento. Por otra parte, si se llegara a utilizar una cámara especializada con la capacidad de hacer seguimiento al rostro del niño, probablemente la configuración de LARGE_FACES tomaría mayor importancia. Finalmente, ambas configuraciones tienen un buen desempeño para este caso, pero se concluye que es necesario tener en cuenta el tamaño relativo del rostro de la escena para futuras implementaciones.

6.2 Agente modal de postura

En relación con la evaluación del agente modal de postura, se aplicó el mecanismo de análisis de Maldonado et al [66]. Para cada conjunto de tres nodos correspondientes a una unión Jp=[xp,yp,zp], Jq=[xq,yq,zq] y Jr=[xr,yr,zr], se calcula el ángulo de la unión, aplicando la ecuación (12).

$$\theta = \arccos\left(\frac{(J_p - J_q)\cdot(J_r - J_q)}{\|J_p - J_q\|\|J_r - J_q\|}\right) \quad \textbf{(12)}$$

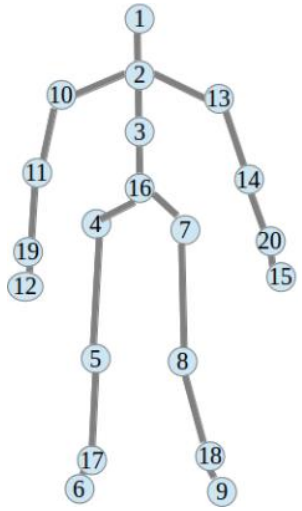


Figura 21 Representación de nodos de Kinect. Tomado de [66].

La Figura 21 muestra la representación de nodos proporcionada por el dispositivo Kinect, cada juntura tiene un número identificador único asociado. Los datos se obtuvieron usando la librería j4k [67]. Debido a que la grabación incluyó únicamente el torso, se utilizaron las 10 junturas a continuación para el cálculo de ángulos.

1. (1, 2, 3)
2. (1, 2, 10)
3. (1, 2, 13)
4. (2, 10, 11)
5. (10, 11, 19)
6. (11, 19, 12)
7. (2, 13, 14)
8. (13, 14, 20)
9. (14, 20, 15)
10. (2, 3, 16)

Para el experimento se utilizó lógica difusa aplicando el algoritmo de cálculo Fuzzy C-Means. Con el propósito de hacer un análisis comparativo, se varió el número de clústeres entre 3 y 5 y, al mismo tiempo, se varió el grado de difusividad para los valores 1.1, 2, 3 y 10. Se utilizó Matlab para calcular el algoritmo Fuzzy C-Means. Como resultado se obtienen los centroides y la función de pertenencia para un número de centroides n.

$$[centroides, U] = fcm(datos, n)$$

Se utilizaron marcos tomados cada 20 segundos y se asociaron a las emociones más cercanas identificadas. Se emplearon los datos de la selección de componentes para el entrenamiento. Como resultado, los mejores porcentajes de exactitud se alcanzaron con un grado de difusividad 2 y un número de 5 clústeres, no obstante, la diferencia no es amplia al compararlo con los valores para 4 clústeres. La Tabla 15 presenta los resultados para cada eje emocional variando el número de clústeres con grado de difusividad 2.

Tabla 15 Matriz para el agente modal de postura por número de clústeres. Elaboración propia.

Eje / n clústeres	3	4	5
Tedio / Fascinación	42	76	81
Ansiedad / Confianza	55	79	82
Frustración/ Euforia	53	74	78

6.3 Agente modal de lenguaje

El componente escogido para la prueba experimental fue Synesketch, la sección 4.2 Selección de componentes, presenta los detalles de los criterios aplicados. Para la implementación, se configuraron los archivos synesketch_lexicon.txt y keywords.xml y se procesaron las palabras mediante la clase EmotionalState con los siguientes parámetros:

- emotions: lista con el conjunto de emociones a evaluar.

- valence: 0, lo que provoca una evaluación normal de las emociones a diferencia de los valores -1 y 1 que las evalúa con caracterizaciones de negativo o positivo respectivamente.

El parámetro generalWeight se utilizó como variable independiente para ajustar la sensibilidad de la librería. Para ello se utilizaron los valores 0.25, 0.5 y 0.75. El valor 0.5 refleja una intensidad emocional promedio. La Tabla 16 presenta los resultados por cada emoción reconocida por Synesketch. Los valores de pesos altos provocaron que únicamente las emociones más destacadas fueran reportadas, los valores de pesos bajos presentaron un mejor resultado para este experimento, esto se puede explicar en la medida en que los estudiantes no utilizan palabras extremas para el uso de emociones con valencia negativa.

Tabla 16 Matriz para el agente modal de lenguaje variando la sensibilidad. Elaboración propia.

		generalWeight		
		0.25	0.5	0.75
Agente Modal Lenguaje	Alegría	72	69	57
	Tristeza	68	65	59
	Ira	65	61	57
	Sorpresa	7	7	5
	Miedo	15	6	0
	Desagrado	10	12	5

Tabla 17 Matriz para el agente modal de lenguaje generalWidth 0.25. Elaboración propia.

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
Agente Modal Lenguaje	Alegría	72	0	0	25	0	0
	Tristeza	11	68	12	0	17	20

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
(0.25)	Ira	5	4	65	0	0	0
	Sorpresa	12	5	0	7	16	33
	Miedo	0	23	4	25	15	37
	Desagrado	0	0	19	43	52	10

Para Synesketch, se encontró que, bajo las condiciones de la prueba experimental, las emociones de alegría, tristeza e ira tienen el mejor desempeño, en contraposición las emociones de sorpresa, miedo y desagrado presentaron resultados con bajo índice de reconocimiento, esto se debe a que los estudiantes no utilizaron palabras precisas para dichos estados emocionales que fueron mejor identificados con el agente modal de rostro. La Tabla 17 muestra la matriz de confusión utilizando el parámetro seleccionado 0.25.

Es importante tener en cuenta que, con la presencia del robot Baxter, se notó una menor cantidad de palabras expresadas por parte de los estudiantes respecto al resultado de la selección de componentes en la que el instructor animaba a los niños a expresar su pensamiento en relación a la solución del problema. Esto podría mejorarse modificando la interacción del robot con el niño, de tal forma que se anime también la expresión de sus ideas durante el ejercicio.

6.4 Agente modal de interacción / tarea

Como se indicó en la sección 4.6, este agente utiliza la cámara especial Intel RealSense que cuenta con funcionalidades de tracking de objetos, a partir de la información obtenida, el agente calcula variables de posición, velocidad, aceleración y tiempo en movimiento para los diferentes bloques. A partir de reglas de lógica difusa configuradas para el agente, el sistema puede encontrar un conjunto de emociones en el aprendizaje correspondientes al modelo de Kort & Reilly [5].

Al aplicar el agente durante la prueba experimental, se obtuvo la matriz de confusión de la Tabla 18. A partir de los resultados, se encuentra que las emociones de tedio y aburrimiento se confunden con cierta facilidad, al igual que las emociones de interés y curiosidad. Es posible que estas emociones no sean claramente diferenciables mediante reglas o que se pudiera mejorar el modelo disminuyendo el número de emociones en cada eje a tres únicamente. No obstante, el reconocimiento actual de emociones por parte del agente permitiría tomar acciones en el contexto del proceso educativo debido a que la cercanía de estas emociones podría conllevar acciones similares por parte del instructor en el contexto del problema.

Tabla 18 Matriz para el agente modal de interacción / tarea. Elaboración propia.

		Panel de Expertos			
		Tedio	Aburrimiento	Interés	Curiosidad
Agente Modal Tarea	Tedio	85	35	0	0
	Aburrimiento	15	65	0	0
	Interés	0	0	79	23
	Curiosidad	0	0	21	77

Por otra parte, este agente modal cuenta con un reconocimiento mejor (85%) en el eje Ansiedad / Confianza, comparado con el resultado del agente de reconocimiento de postura (82%), el agente de rostro (77%) y el agente de lenguaje (55%). La Tabla 19 muestra la comparación entre la precisión medida para los agentes por separado en los ejes emocionales del modelo. El resultado para eje Tedio / Fascinación es cercano al registrado por los agentes de postura y rostro, sin embargo, se distancia bastante en los resultados del eje Frustración/ Euforia (60%), en donde los agentes de postura (78%) y rostro (76%) cuentan con resultados superiores.

Tabla 19 Comparación entre agentes. Elaboración propia.

Eje / Agente	Interacción / tarea	Postura	Rostro	Lenguaje
Tedio / Fascinación	76	81	79	50
Ansiedad / Confianza	85	82	77	55
Frustración/ Euforia	60	78	76	41

6.5 Agente de procesamiento multimodal

A continuación se presentan los resultados de la aplicación del agente de procesamiento multimodal, conforme al comportamiento indicado en la sección 4.5. Este agente utiliza los mensajes enviados por los agentes modales, de tal forma que se consolidan los resultados para lograr un reconocimiento de emociones unificado. Para los datos a continuación, el agente de reconocimiento de postura se ejecutó con el parámetro de 5 clústeres y el agente modal de

rostro con SMALL_FACES; por su parte, el agente modal de lenguaje se configuró con generalWeight 0.25.

La Tabla 20 muestra la matriz de confusión resultado del procesamiento multimodal para las emociones comunes. El resultado en el reconocimiento que se logra es positivo y se puede ver más claro en la

Tabla 21 en la que se detalla la mejora o disminución en la precisión del reconocimiento comparado contra el mayor alcanzado con los agentes modales por separado, desde el punto de vista de los positivos / positivos. En promedio, la mejora resultó del 8%, en especial, corrigiendo el reconocimiento para las emociones de ira y miedo. Sin embargo, ocurre una disminución del 5% en la precisión de la emoción "sorpresa".

Finalmente, la Tabla 22 muestra el resultado del reconocimiento multimodal para los ejes emocionales del modelo, en este caso, se puede apreciar una mejora del 5% en la precisión total y con una tendencia más homogénea en la mejora del proceso de reconocimiento al aplicar la técnica multimodal.

Tabla 20 Matriz para el reconocimiento Multimodal de emociones comunes. Elaboración propia.

		Panel de Expertos					
		Alegría	Tristeza	Ira	Sorpresa	Miedo	Desagrado
Agente Multimodal	Alegría	84	0	0	15	0	0
	Tristeza	0	83	5	0	14	9
	Ira	0	7	84	0	0	3
	Sorpresa	16	4	0	79	0	5
	Miedo	0	6	7	0	74	11
	Desagrado	0	0	4	6	12	72

Tabla 21 Diferencia en el reconocimiento de emociones comunes. Elaboración propia.

	Multimodal	Máximo	Diferencia

		Multimodal	Máximo	Diferencia
Agente Multimodal	Alegría	84	78	8%
	Tristeza	83	77	8%
	Ira	84	72	17%
	Sorpresa	79	83	-5%
	Miedo	74	65	14%
	Desagrado	72	68	6%

Tabla 22 Diferencia en el reconocimiento de ejes emocionales. Elaboración propia.

		Multimodal	Máximo	Diferencia
Agente Multimodal	Tedio / Fascinación	86	81	6%
	Ansiedad / Confianza	89	85	5%
	Frustración/ Euforia	82	78	5%

En la Tabla 23 se presentan los cálculos de los factores de confianza (f_{ij}) descritos anteriormente en la sección 4.5. Estos valores muestran que para las emociones de alegría e ira, los agentes de rostro y lenguaje tienen factores de confianza similares, un poco mayor para el agente de lenguaje en el caso de alegría (89%) y para el agente de rostro en el caso de ira (90%). El caso de sorpresa, el agente de rostro presenta un factor de confianza superior (69%) en comparación con el de lenguaje (61%), aunque todavía cercano. Para las demás emociones comunes, tristeza, miedo y desagrado, los resultados muestran una diferencia significativa en favor del agente de rostro, lo que indica que, en este experimento, cuenta con un nivel de confianza mucho más alto para dichas emociones.

Tabla 23 Cálculo de factores de confianza. Elaboración propia.

	Rostro	Lenguaje

	Rostro	Lenguaje
Alegría	86	89
Tristeza	91	66
Ira	90	85
Sorpresa	69	61
Miedo	69	38
Desagrado	84	51

Finalmente, en la Tabla 19 se presentaron los resultados comparativos para los ejes emocionales de todos los agentes modales. Se encontró que los agentes de postura y de interacción / tarea obtuvieron resultados superiores frente a aquellos de los agentes modales de rostro y lenguaje. En el eje Ansiedad / Confianza el agente de interacción / tarea logró un 85% de precisión, y en los ejes de Tedio / Fascinación (81%) y Frustración/ Euforia (78%) los mejores resultados los obtuvo el agente de postura. Consideramos que esto se debe a que estos agentes fueron especialmente diseñados para el reconocimiento de estos ejes emocionales. El agente de lenguaje obtuvo resultados bajos en los tres ejes, sin embargo, los resultados para agente de rostro fueron buenos e incluso superiores a los del agente de tarea, en especial para el eje de Frustración/ Euforia en donde obtuvo 76% de precisión frente al 60% del agente de tarea. Este comportamiento se debe a que los ejes emocionales se logran diferenciar a partir de la información de emociones comunes en las que el agente modal de rostro presenta un desempeño superior frente al agente de lenguaje, esto es, las emociones de sorpresa, miedo y desagrado.

7 CONCLUSIONES Y RECOMENDACIONES

Este trabajo demuestra cómo un enfoque multimodal para el reconocimiento de emociones, puede enriquecer el proceso educativo y aportar información relevante del estado emocional del estudiante. Esta información permite brindar retroalimentación efectiva al sistema educativo, en particular, tomar acciones que encaminen al niño a sortear las dificultades en el proceso de aprendizaje. La arquitectura desarrollada permite agregar nuevos agentes modales de manera flexible, sin estar atados a un conjunto específico de componentes y ayuda a que la información fluya al agente de reconocimiento multimodal. Las contribuciones que consideramos de mayor relevancia son las siguientes:

- Reconocimiento Multimodal de Emociones: el diseño del módulo de reconocimiento multimodal permitió una mejora del 8% en el reconocimiento de emociones comunes y de un 5% para emociones en el modelo especializado en educación. Se elaboró una arquitectura de reconocimiento basada en un enfoque de Sistemas Multiagente (SMA) en la que los agentes interactúan en forma cooperativa con un bajo acoplamiento, permitiendo agregar nuevos agentes de reconocimiento modal.
- Modelo de Emociones Educativas: se incorporó al diseño del reconocimiento una nueva gama de emociones que no había sido profundizada, en tanto que estudios previos se habían enfocado en el reconocimiento de emociones comunes. En este trabajo se incorpora al módulo de reconocimiento la habilidad de reconocer emociones especialmente dirigidas al proceso educativo.
- Reconocimiento de Emociones de Postura: se diseñó un agente especializado y se elaboró un prototipo para el reconocimiento de emociones a partir de la posición corporal, utilizando los datos proporcionados por el dispositivo Kinect. Los resultados del prototipo y la prueba experimental muestran que la información de postura brinda datos relevantes en el análisis de emociones para el caso educativo.
- Reconocimiento de Emociones de Tarea / Interacción: el reconocimiento de emociones a partir de la interacción con los elementos del entorno es un punto que ha sido poco estudiado. En este trabajo se diseñó un agente especializado a partir de la interacción del niño con los bloques que hacían parte del problema presentado durante la prueba experimental. Los resultados mostraron que estos datos enriquecen el proceso de reconocimiento de emociones y brindan información relevante para el sistema educativo.

El resultado del trabajo en relación con el reconocimiento multimodal de emociones, cuenta con posibilidades de ser aplicado y ampliado en las actividades del proceso de educación, de tal forma que sea incorporado al trabajo regular en instituciones educativas. Al proporcionar información de contexto clave, contribuye al desarrollo del proceso y mejora la experiencia para estudiantes y maestros. De forma complementaria, esta tecnología puede ser extendida a sistemas que brinden soporte emocional como es el caso del cuidado de personas de la tercera edad, personas con enfermedades y, en general, aquellos entornos en los que un acercamiento humano desde el punto de vista de las emociones brinde una mejor relación entre el usuario y los elementos tecnológicos. Por otra parte, el trabajo cuenta con proyección en el segmento de robots asistentes en tanto que brinda información relevante para la toma de decisiones a partir de la interacción hombre máquina y la expresión emocional por parte de las personas.

La capacidad de los sistemas para obtener información a partir de la comunicación no verbal es un punto que debe ser estudiado y mejorado. Este trabajo puede ser extendido y se recomienda sea aplicado en otros contextos de interacción hombre máquina en los que la percepción de emociones del usuario mejore la experiencia y la usabilidad. Por otra parte, y como trabajo futuro, la implementación de agentes enriquecidos podría favorecer la capacidad de reconocimiento y, en especial, la información disponible para la retroalimentación al sistema educativo, de tal forma que se tomen acciones adecuadas al estado del aprendizaje.

Un punto de mejora que se detectó a partir de la prueba experimental es la necesidad de un lazo de retroalimentación inhibitorio por parte del sistema educativo, de tal forma que, cuando el robot inicie sus movimientos o expresiones, se pueda detener la operación de algunos agentes especializados. Un caso específico sería que, cuando el robot realice movimientos, se informe esta situación a los agentes que puedan verse obstaculizados por los brazos del robot o por la interacción del mismo con los bloques o elementos de la tarea.

El área de reconocimiento de emociones en general, admite mayores esfuerzos por parte de los investigadores para encontrar métodos más adecuados a las necesidades de evaluación en línea y bajo consumo de recursos de procesamiento. Nuevas técnicas de inteligencia artificial pueden aplicarse para desarrollar agentes modales que se integren con la arquitectura propuesta.

Finalmente, es importante destacar las posibilidades que brinda el robot Baxter a diferentes tipos de investigación, en particular, en la Pontificia Universidad Javeriana. En contextos educativos, el robot cuenta con habilidades que pueden ser aprovechadas para diseñar nuevas interacciones aplicables en clases de tecnología o en el desarrollo de habilidades analíticas por parte de los estudiantes. A los niños que participaron en la prueba experimental, les resultó muy atractiva la posibilidad de interactuar con el robot Baxter. La experiencia fue positiva en la interacción con el robot humanoide y en ningún momento manifestaron haberse sentido intimidados por su presencia.

En resumen, las principales contribuciones de este trabajo de grado radican en elaborar una arquitectura flexible para el reconocimiento multimodal de emociones, la incorporación de emociones especializadas en el proceso educativo más allá de las emociones comunes previamente estudiadas, y el diseño de agentes especializados en el reconocimiento de emociones a partir de la postura y la interacción con los elementos que hacen parte de la actividad educativa. También se integran una serie de dispositivos, el robot Baxter, Kinect, Intel RealSense, wearables, mediante los cuales se logra incorporar información de postura y de interacción con los bloques. Esta puesta en funcionamiento es novedosa, puesto que en los trabajos de investigación identificados durante la profundización del estado del arte, no se encontró un montaje tan completo en el que también hiciera parte la interacción con el robot humanoide y que formara parte de un proceso educativo.

8 REFERENCIAS

[1] M. Feidakis, T. Daradoumis, y S. Caballé, «Emotion measurement in intelligent tutoring systems: what, when and how to measure», en *Intelligent Networking and Collaborative Systems (INCoS), 2011 Third International Conference on*, 2011, pp. 807–812.

[2] Z.-W. Hong, Y.-M. Huang, M. Hsu, y W.-W. Shen, «Authoring Robot-Assisted Instructional Materials for Improving Learning Performance and Motivation in EFL Classrooms», *Educ. Technol. Soc.*, vol. 19, n.º 1, pp. 337-349, ene. 2016.

[3] Y.-M. Jang, R. Mallipeddi, S. Lee, H.-W. Kwak, y M. Lee, «Human intention recognition based on eyeball movement pattern and pupil size variation», *NEUROCOMPUTING*, vol. 128, pp. 421-432, mar. 2014.

[4] Z. Callejas, D. Griol, y R. Lopez-Cozar, «Predicting user mental states in spoken dialogue systems», *EURASIP J. Adv. SIGNAL Process.*, 2011.

[5] B. Kort y R. Reilly, «Analytical models of emotions, learning and relationships: towards an affect-sensitive cognitive machine», en *Conference on virtual worlds and simulation (VWSim 2002)*, 2002.

[6] M. Alemi, A. Meghdari, y M. Ghazisaedy, «Employing Humanoid Robots for Teaching English Language in Iranian Junior High-Schools», *Int. J. HUMANOID Robot.*, vol. 11, n.º 3, sep. 2014.

[7] G. Keren y M. Fridin, «Kindergarten Social Assistive Robot (KindSAR) for children's geometric thinking and metacognitive development in preschool education: A pilot study», *Comput. Hum. Behav.*, vol. 35, pp. 400-412, 2014.

[8] H. Lehmann, I. Iacono, K. Dautenhahn, P. Marti, y B. Robins, «Robot companions for children with down syndrome A case study», *Interact. Stud.*, vol. 15, n.º 1, pp. 99-112, 2014.

[9] A. Edwards, C. Edwards, P. R. Spence, C. Harris, y A. Gambino, «Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher"», *Comput. Hum. Behav.*, p. , 2016.

[10] M. Fridin, «Kindergarten social assistive robot: First meeting and ethical issues», *Comput. Hum. Behav.*, vol. 30, pp. 262-272, ene. 2014.

[11] R. A. Calvo y S. D'Mello, «Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications», *IEEE Trans. Affect. Comput.*, vol. 1, n.º 1, pp. 18-37, jun. 2010.

[12] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, y K. R. Scherer, «On the acoustics of emotion in audio: what speech, music, and sound have in common», *Front. Psychol.*, vol. 4, may 2013.

[13] A. Esposito, A. M. Esposito, y C. Vogel, «Needs and challenges in human computer interaction for processing social emotional information», *PATTERN Recognit. Lett.*, vol. 66, pp. 41-51, nov. 2015.

[14] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, y P. Kuksa, «Natural Language Processing (Almost) from Scratch», *J. Mach. Learn. Res.*, vol. 12, pp. 2493-2537, ago. 2011.

[15] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, y B. Benhabib, «A Survey of Autonomous Human Affect Detection Methods for Social Robots Engaged in Natural HRI», *J. INTELLIGENT Robot. Syst.*, vol. 82, n.º 1, pp. 101-133, abr. 2016.

[16] D. Schacter, C. Wang, G. Nejat, y B. Benhabib, «A two-dimensional facial-affect estimation system for humanrobot interaction using facial expression parameters», *Adv. Robot.*, vol. 27, n.º 4, pp. 259-273, mar. 2013.

[17] F. Alonso-Martin, M. Malfaz, J. Sequeira, J. F. Gorostiza, y M. A. Salichs, «A Multimodal Emotion Detection System during Human-Robot Interaction», *SENSORS*, vol. 13, n.º 11, pp. 15549-15581, nov. 2013.

[18] P. Barros, D. Jirak, C. Weber, y S. Wermter, «Multimodal emotional state recognition using sequence-dependent deep hierarchical features», *NEURAL Netw.*, vol. 72, pp. 140-151, dic. 2015.

[19] J. A. Prado, C. Simplicio, N. F. Lori, y J. Dias, «Visuo-auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction», *Int. J. Soc. Robot.*, vol. 4, n.º 1, pp. 29-51, nov. 2012.

[20] S. Dobrisek, R. Gajsek, F. Mihelic, N. Pavesic, y V. Struc, «Towards Efficient Multi-Modal Emotion Recognition», *Int. J. Adv. Robot. Syst.*, vol. 10, ene. 2013.

[21] S. K. D'Mello y A. Graesser, «Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features», *USER Model. USER-Adapt. Interact.*, vol. 20, n.º 2, pp. 147-187, jun. 2010.

[22] M. Saneiro, O. C. Santos, S. Salmeron-Majadas, y J. G. Boticario, «Towards Emotion Detection in Educational Scenarios from Facial Expressions and Body Movements through Multimodal Approaches», *Sci. WORLD J.*, 2014.

[23] F. Alonso-Martin y M. A. Salichs, «Integration of a Voice Recognition System in a Social Robot», *Cybern. Syst.*, vol. 42, n.º 4, pp. 215-245, 2011.

[24] S. Calinon, «A tutorial on task-parameterized movement learning and retrieval», *INTELLIGENT Serv. Robot.*, vol. 9, n.º 1, pp. 1-29, ene. 2016.

[25] J. M. Romano, J. P. Brindza, y K. J. Kuchenbecker, «ROS open-source audio recognizer: ROAR environmental sound detection tools for robot programming», *Auton. ROBOTS*, vol. 34, n.º 3, pp. 207-215, abr. 2013.

[26] A. Jain, S. Sharma, T. Joachims, y A. Saxena, «Learning preferences for manipulation tasks from online coactive feedback», *Int. J. Robot. Res.*, vol. 34, n.º 10, pp. 1296-1313, sep. 2015.

[27] P. Liang, L. Ge, Y. Liu, L. Zhao, R. Li, y K. Wang, «An Augmented Discrete-Time Approach for Human-Robot Collaboration», *DISCRETE Dyn. Nat. Soc.*, 2016.

[28] Z. Zhang, Y. Liu, A. Li, y M. Wang, «A novel method for user-defined human posture recognition using Kinect», en *2014 7th International Congress on Image and Signal Processing*, 2014, pp. 736-740.

[29] S. Hahn *et al.*, «Comparing Stochastic Approaches to Spoken Language Understanding in Multiple Languages», *IEEE Trans. AUDIO SPEECH Lang. Process.*, vol. 19, n.º 6, pp. 1569-1583, ago. 2011.

[30] R. Sarikaya, G. E. Hinton, y A. Deoras, «Application of Deep Belief Networks for Natural Language Understanding», *IEEE-ACM Trans. AUDIO SPEECH Lang. Process.*, vol. 22, n.º 4, pp. 778-784, abr. 2014.

[31] K.-T. Song, M.-J. Han, y S.-C. Wang, «Speech signal-based emotion recognition and its application to entertainment robots», *J. Chin. Inst. Eng.*, vol. 37, n.º 1, pp. 14-25, 2014.

[32] E. González y M. Torres, «Organizational approach for agent oriented programming», en *8th Int. Conf. on Enterprise Information Systems-ICEIS*, 2006, pp. 75–80.

[33] C. A. Iglesias, M. Garijo, J. C. González, y J. R. Velasco, «Analysis and design of multiagent systems using MAS-CommonKADS», en *International Workshop on Agent Theories, Architectures, and Languages*, 1997, pp. 313–327.

[34] J. Brinkkemper y A. Solvberg, «Tropos: A framework for requirements-driven software development», *Inf. Syst. Eng. State Art Res. Themes*, p. 11, 2000.

[35] E. González, «Inteligencia Computacional Redes Neuronales». Departamento de Ingeniería de Sistemas. Facultad de Ingeniería. Pontificia Universidad Javeriana, mar-2013.

[36] W. Correia, L. Rodrigues, F. Campos, M. Soares, y M. Barros, «The methodological involvement of the emotional design and cognitive ergonomics as a tool in the development of children products», *WORK- J. Prev. Assess. Rehabil.*, vol. 41, n.º 1, pp. 1066-1071, 2012.

[37] O. M. Lourenco, «Developmental stages, Piagetian stages in particular: A critical review», *NEW IDEAS Psychol.*, vol. 40, n.º B, pp. 123-137, ene. 2016.

[38] S. Lee *et al.*, «On the effectiveness of Robot-Assisted Language Learning», *RECALL*, vol. 23, n.º 1, pp. 25-58, ene. 2011.

[39] I. C. Bacivarov y V. L. M. Ilian, «The paradigm of utilizing robots in the teaching process: a comparative study», *Int. J. Technol. Des. Educ.*, vol. 22, n.º 4, pp. 531-540, nov. 2012.

[40] S. L. Chu, G. Angello, M. Saenz, y F. Quek, «Fun in Making: Understanding the Experience of Fun and Learning through Curriculum-based Making in the Elementary School Classroom», *Entertain. Comput.*, p. , 2016.

[41] V. Nacher, F. Garcia-Sanjuan, y J. Jaen, «Interactive technologies for preschool game-based instruction: Experiences and future challenges», *Entertain. Comput.*, vol. 17, pp. 19-29, 2016.

[42] B. Robins y K. Dautenhahn, «Tactile Interactions with a Humanoid Robot: Novel Play Scenario Implementations with Children with Autism», *Int. J. Soc. Robot.*, vol. 6, n.º 3, SI, pp. 397-415, ago. 2014.

[43] A. Roy *et al.*, «INNS Conference on Big Data 2015 Program San Francisco, CA, USA 8-10 August 2015 Big Data Analytics as a Service for Affective Humanoid Service Robots», *Procedia Comput. Sci.*, vol. 53, pp. 141-148, 2015.

[44] J. J. Páez y E. González, «Human-Robot Scaffolding, a cognitive architecture to foster the metacognitive and emotional support using a Baxter robot».

[45] B. R. Steunebrink, M. Dastani, y J.-J. C. Meyer, «A formal model of emotion triggers: an approach for BDI agents», *Synthese*, vol. 185, pp. 83-129, abr. 2012.

[46] J. L. Salmeron, «Fuzzy cognitive maps for artificial emotions forecasting», *Appl. SOFT Comput.*, vol. 12, n.º 12, pp. 3704-3710, dic. 2012.

[47] A. Batliner *et al.*, «Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech», *Comput. SPEECH Lang.*, vol. 25, n.º 1, SI, pp. 4-28, ene. 2011.

[48] G. Littlewort *et al.*, «The motion in emotion - A CERT based approach to the FERA emotion challenge», en *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, 2011, pp. 897-902.

[49] M. Buonamente, H. Dindo, y M. Johnsson, «Hierarchies of Self-Organizing Maps for action recognition», *Cogn. Syst. Res.*, vol. 39, pp. 33-41, sep. 2016.

[50] M. Karg, A.-A. Samadani, R. Gorbet, K. Kuehnlenz, J. Hoey, y D. Kulic, «Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation», *IEEE Trans. Affect. Comput.*, vol. 4, n.º 4, pp. 341-U157, dic. 2013.

[51] D. Novak, M. Mihelj, y M. Munih, «A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing», *Interact. Comput.*, vol. 24, n.º 3, pp. 154-172, may 2012.

[52] J. C. Castillo *et al.*, «Software Architecture for Smart Emotion Recognition and Regulation of the Ageing Adult», *Cogn. Comput.*, vol. 8, n.º 2, pp. 357-367, abr. 2016.

[53] S. K. D'Mello, N. Dowell, y A. Graesser, «Does It Really Matter Whether Students' Contributions Are Spoken Versus Typed in an Intelligent Tutoring System With Natural Language?», *J. Exp. Psychol.-Appl.*, vol. 17, n.º 1, pp. 1-17, mar. 2011.

[54] S. K. D'Mello, B. Lehman, y N. Person, «Monitoring affect states during effortful problem solving activities», *Int. J. Artif. Intell. Educ.*, vol. 20, n.º 4, pp. 361–389, 2010.

[55] A. Metallinou, M. Woellmer, A. Katsamanis, F. Eyben, B. Schuller, y S. Narayanan, «Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification», *IEEE Trans. Affect. Comput.*, vol. 3, n.º 2, pp. 184-198, jun. 2012.

[56] S. Poria, E. Cambria, N. Howard, G.-B. Huang, y A. Hussain, «Fusing audio, visual and textual clues for sentiment analysis from multimodal content», *Neurocomputing*, vol. 174, Part A, pp. 50-59, ene. 2016.

[57] F. Ringeval *et al.*, «Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data», *Pattern Recognit. Lett.*, vol. 66, pp. 22-30, nov. 2015.

[58] U. Krcadinac, P. Pasquier, J. Jovanovic, y V. Devedzic, «Synesketch: An Open Source Library for Sentence-Based Emotion Recognition», *IEEE Trans. Affect. Comput.*, vol. 4, n.º 3, pp. 312-325, sep. 2013.

[59] K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, y H. Tsujino, «Design and Implementation of Robot Audition System `HARK' - Open Source Software for Listening to Three Simultaneous Speakers», *Adv. Robot.*, vol. 24, n.º 5-6, pp. 739-761, 2010.

[60] F. Dai, Y. Li, y G. You, «Development of an Intelligent Assistant Robot based on Embedded RTOS», *J. Robot. Netw. Artif. LIFE*, vol. 2, n.º 3, pp. 200-204, dic. 2015.

[61] C. Vendome, M. Linares-Vasquez, G. Bavota, M. D. Penta, D. German, y D. Poshyvanyk, «License Usage and Changes: A Large-Scale Study of Java Projects on GitHub», en *2015 IEEE 23rd International Conference on Program Comprehension*, 2015, pp. 218-228.

[62] E. González, J. A. Ávila, y C. J. Bustacara, «BESA: Arquitectura para Construcción de Sistemas MultiAgentes», presentado en Conferencia Latinoaméricana de Informática, La Paz- Bolivia, 2003, p. 70.

[63] C.-Y. Lee, M.-J. Chen, y W.-L. Chang, «Effects of the Multiple Solutions and Question Prompts on Generalization and Justification for Non-Routine Mathematical Problem Solving in a Computer Game Context», *Eurasia J. Math. Sci. Technol. Educ.*, vol. 10, n.º 2, pp. 89-99, abr. 2014.

[64] «Logicgames.com - Frog Jump Game». [En línea]. Disponible en: http://www.logicgames.com/webgames/frogjump.html. [Accedido: 06-nov-2017].

[65] J. Pransky, «The Pransky interview: Dr Rodney Brooks, Robotics Entrepreneur, Founder and CTO of Rethink Robotics», *Ind. ROBOT- Int. J.*, vol. 42, n.º 1, pp. 1-4, 2015.

[66] C. Maldonado, H. V. Rios-Figueroa, y A. Marin-Hernandez, «Improving action recognition by selection of features», en *2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*, 2016, pp. 1-6.

[67] A. Barmpoutis, «Tensor Body: Real-Time Reconstruction of the Human Body and Avatar Synthesis From RGB-D», *IEEE Trans. Cybern.*, vol. 43, n.º 5, pp. 1347-1356, oct. 2013.