UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

SISTEMA DE APRENDIZAJE EMOCIONAL PARA LA INTERACCIÓN HOMBRE-ROBOT BASADO EN CAPACIDADES EMOCIONALES

(EMOTION LEARNING SYSTEM FOR AFFECTIVE HUMAN-ROBOT INTERACTION BASED ON EMOTIONAL AFFORDANCES)

FELIPE ANDRÉS CID BURGOS

DEPT. DE TECNOLOGÍA DE LOS COMPUTADORES Y LAS COMUNICACIONES

2014

UNIVERSIDAD DE EXTREMADURA

TESIS DOCTORAL

SISTEMA DE APRENDIZAJE EMOCIONAL PARA LA INTERACCIÓN HOMBRE-ROBOT BASADO EN CAPACIDADES EMOCIONALES

FELIPE ANDRÉS CID BURGOS

DEPT. DE TECNOLOGÍA DE LOS COMPUTADORES Y LAS COMUNICACIONES

Conformidad del Director:

Dr. Pedro Miguel Núñez Trujillo

2014

Sistema de aprendizaje emocional para la interacción Hombre-Robot basado en capacidades emocionales

(Emotion Learning System for Affective Human Robot Interaction based on Emotional Affordances)

Felipe Andrés Cid Burgos

Cáceres, 2014
Universidad de Extremadura
Doctoral Dissertation

Dedicado a mi familia, amigos y al
Laboratorio RoboLab de la Universidad de Extremadura

Resumen:

Uno de los principales retos de la robótica social en los últimos años es el desarrollo de robots sociales autónomos capaces de realizar tareas complejas e interactuar tanto con las personas como con los elementos del entorno, todo ello de forma intuitiva y similar a la humana. Siguiendo este enfoque, para dar mayor naturalidad a la interacción, las plataformas robóticas deberían ser capaces de comprender el comportamiento y las intenciones del ser humano, realimentándose y aprendiendo durante la comunicación. Además, la capacidad de expresar emociones por parte de un robot autónomo en una interacción añadiría un valor adicional, a veces crucial, para una comunicación más natural y cercana. Sin embargo, los procesos de aprendizaje del comportamiento humano suelen ser invasivos y poco naturales, adquiriendo muy poca información desde fuentes tan completas y robustas como son el lenguaje natural o la propia imitación. Además sucede que este aprendizaje está normalmente restringido a entornos controlados, los cuales toman en consideración principalmente las capacidades del robot para imitar las habilidades físicas aprendidas del humano. Por su parte, existen pocos trabajos en la literatura que centren su investigación en cómo aprende un robot autónomo, de forma natural, su comportamiento afectivo. La mayoría de estos trabajos definen modelos de comportamiento fijos en el robot, sin la posibilidad de aprendizaje y, por tanto, de adaptarse a nuevas situaciones durante la interacción.

En esta Tesis Doctoral se presenta un sistema de aprendizaje emocional para una interacción Hombre-Robot que implementa este concepto por medio de dos enfoques complementarios. En primer lugar, esta Tesis contribuye con un sistema para el reconocimiento e imitación de la información emocional del usuario basado en el análisis del lenguaje natural del humano. Tanto la expresión facial, como la voz humana y el lenguaje corporal durante la interacción, son analizados por el sistema para extraer un conjunto de características faciales, acústicas y corporales, respectivamente, que son a posteriori empleadas para obtener el estado emocional del interlocutor durante la comunicación. A lo largo de todo el trabajo se emplea un enfoque bayesiano que permite estimar un conjunto discreto de estados emocionales, en concreto los estados de felicidad, tristeza, miedo, enfado y el propio estado neutral del humano. A su vez, esta Tesis contribuye con un sistema de imitación de emociones, adaptable a diferentes agentes, mediante el modelado de cada una de las emociones anteriores. Para su desarrollo y evaluación, el sistema de imitación presentado utiliza la cabeza robótica $Muecas$ que, dado su diseño antropomórfico, permite expresar información emocional por medio de las expresiones faciales, la generación de mensajes verbales a través de audio sintético y el lenguaje corporal del propio movimiento del cuello y la boca.

En segundo lugar, esta información emocional adquirida desde los usuarios ha sido utilizada en esta Tesis para el desarrollo de un sistema de aprendizaje del comportamiento afectivo del robot durante la interacción. Esta contribución se basa en la extensión del concepto clásico de *Affordances*, lo que se ha denominado en este trabajo como *affordances* emocionales. Las *affordances* emocionales representan la relación existente entre diferentes elementos afectivos, entre ellos los objetos del entorno y los propios estados emocionales del usuario, con respecto a las posibles reacciones del robot a lo largo de una comunicación para conseguir un efecto determinado en el usuario. Por lo tanto, en este enfoque presentado, se pretende que el agente robótico, de forma autónoma y no predefinida, pueda modificar el estado emocional del usuario

mediante la predicción del efecto que tendrán las reacciones emocionales del robot, o las propias acciones del agente sobre los objetos del entorno. Para ello se implementó un sistema de aprendizaje de estas *affordances* emocionales que sigue nuevamente una perspectiva bayesiana y que hace uso de los sistemas de reconocimiento e imitación de emociones, junto con la interacción con objetos del entorno.

Finalmente, en esta Tesis Doctoral se implementa un modelo de comportamiento afectivo del robot basado en *affordances* emocionales, capaz de utilizar tanto las reacciones del agente como sus acciones sobre el entorno para, gradualmente, modificar el estado emocional del usuario.

Abstract:

Along recent years, one of the main challenges of social robotics is the development of autonomous social robots able to develop complex tasks and interact both with people and elements from the environment, all that in an intuitive way and similar to humans. Following this approach, in order to make interaction more natural, robotics platforms should be able to understand the behavior and interactions of human beings, learning and retrieving feedback from communication. Also the ability of expressing emotions from an autonomous robot in an interaction would add an extra value, sometimes crucial, for a more natural and close communication. However, human behavior learning processes are usually invasive and not natural, obtaining very few information from as robust and complete sources such as the natural language o the imitation itself. On top of that, this learning is usually restricted to controlled environments, which take into account mainly the abilities of the robot to imitate the learned human physical skills. Besides, related literature has very few works with a research focus in how an autonomous robot learn in a natural way, its affective behavior. Most of these works define behavior models permanent in the robot, without the chance of learning and, in consequence, of adapting to new situations during its interaction.

In this PhD Thesis an emotional learning system intended for a Human-Robot interaction that implements this concept by the means of two complementary approaches is presented. Firstly, this Thesis contributes with a recognition and imitation system of the user emotional information based in the analysis of the human natural language. The facial expression, as well as the human voice and the body language during the interaction, are analysed by the system in order to extract a set of facial, acoustics and corporal characteristics respectively, that are a posteriori employed in order to obtain an emotional state of the spokesperson during the communication. Along the whole work a bayesian approach that allows to estimate a discrete emotional set is used, specifically states of happiness, sadness, fear, angriness, and the neutral human state itself. In turn, this Thesis contributes with an emotional imitation system, adaptable to different agents, modeling each one of the previous emotions. In order to test and experiment it, $Muecas$, a robotics head is used, that, due to its anthropomorphic design, allows to express emotional information through facial expressions, verbal messages generation, through its sintetics audio system, and corporal language of mouth and neck.

Secondly, this acquired emotional information from users has been used in this Thesis for the development of an affective behavior learning system of the robot during interaction. This contribution is based in the extension of the classic concept of $Affordances$, which is usually named in this work as $emotional\ affordances$. The $emotional\ affordances$ represent an existing relation between different affective elements, between them and the environment and the emotional states of the spokesperson itself (change for user), with respect to the possibilities of the robot along a communication in order to obtain a certain effect by the user. Consequently, with this perspective, the autonomous not predefined modification of the emotional state of the user is pretended through the prediction and effect that emotional reactions of the robot would make, or even the agent actions through the environment objects themselves. In order to do that an $emotional\ affordances$ learning system was implemented that follows, again, a bayesian perspective and makes use of emotions recognition and imitation systems, along with the interaction of objects from the environment.

Finally, in this PhD Thesis an affective behavior of the robot based in $emotional\ affordances$ is implemented, able to use the relations of the agent as well as its actions on the environment in order to, gradually, modify the user emotional state.

Índice general

Capítulo 1

Introducción

1.1. Motivaciones

En un futuro cercano, los robots serán una parte esencial de la vida diaria de los humanos, interactuarán con diferentes elementos y usuarios en entornos comunes, como nuestras propias casas, oficinas, fábricas o en espacios abiertos, nos ayudarán en nuestras tareas cotidianas y, en definitiva, mejorarán nuestra calidad de vida. Por poner un claro ejemplo, diversos estudios afirman que el envejecimiento de la población de los países desarrollados traerá consigo un aumento masivo del uso de estos robots como cuidadores, ayudando a las tareas domésticas y facilitando una mayor autonomía de las personas mayores y dependientes. Siguiendo este concepto, cada robot debería poseer capacidades sociales que le permitan interactuar de forma similar a la humana por medio de reglas sociales, como el uso del lenguaje natural en una interacción, o también disponer de capacidades para percibir y aprender cómo relacionarse durante una interacción, no sólo con los usuarios, sino también con los elementos del entorno.

Los robots sociales, justo aquellos robots capaces de interaccionar con otros usuarios o robots siguiendo normas sociales, aprendidas o programadas, son un tema de estudio en auge dentro de la comunidad robótica internacional. Dada la complejidad de cualquiera de las tareas a las que tiene que enfrentarse, desde el propio movimiento por el entorno de trabajo, hasta la interacción más sencilla (por ejemplo, simplemente coger una taza de una mesa para ofrecérsela al usuario), numerosos grupos de investigación centran su esfuerzo en el desarrollo de algoritmos y métodos que superan alguno de estos grandes retos. Entre ellos, la interacción entre un humano y un robot, conocida normalmente en el ámbito como IHR (*Interacción Humano-Robot*), es en sí misma una tarea amplia con numerosas posibilidades de estudio.

Una interacción real entre un usuario humano y un robot conlleva resolver diferentes problemas. Como ejemplo, lo que para un humano es innato, cómo localizar al otro interlocutor por la voz o por el sentido de la vista, para un robot es todo un gran reto a resolver. Si a esto se le une que durante la interacción suceden numerosas situaciones que son incontrolables para un agente robótico, el abanico de posibilidades de trabajo se amplía aún más. Tratar de dotar al robot con la capacidad de aprender, de forma autónoma, a interactuar con un usuario y solventar situaciones desconocidas durante la misma mediante un aprendizaje previo de casos parecidos, se convierte entonces en una posible solución para mucho de estos retos.

Por este motivo, y en el caso particular que se presenta en esta Tesis Doctoral, el proceso de aprendizaje de las reglas sociales asociadas con una comunicación entre un humano y un robot por medio del lenguaje natural, dependen de la capacidad para comprender no sólo las

intenciones y el comportamiento del usuario, sino también conocer su estado emocional. Esto último, que el robot conozca cómo se encuentra en usuario durante una comunicación (triste, cansado, o muy alegre, por ejemplo), permite encaminar la interacción entre ambos agentes, tal y como realmente ocurre en conversaciones entre personas. Así, los robots sociales del futuro dispondrán de estas capacidades entre sus funciones, aquellas que permitan realmente adaptarse de forma autónoma a cambios inesperados dentro de una interacción en tiempo real. Esta es la razón de que muchos estudios, sobre todo en la última década, hayan considerado la información emocional como pieza clave en la IHR.

Por otro lado, la capacidad de percibir y aprender cómo relacionarse con los elementos del entorno y cómo hacer uso de los mismos en una interacción con un humano es también otro tema de interés creciente dentro de la robótica social. La idea de programar un robot con todas las posibles soluciones es, en sí misma, una respuesta inviable dentro del ámbito de estudio en el que se trabaja actualmente. Por ello, surgen diferentes teorías en la literatura que abordan el problema, siendo quizá la más interesante aquella que permite dotar de estas capacidades al robot mediante procesos de aprendizaje biológicamente inspirados por imitación. Es decir, el robot se enfrenta al problema de aprender cómo hacer uso de los objetos observando e imitando cómo lo hace un humano (u otro robot). Está claro que, en general, la imitación es uno de los mecanismos más completos para el aprendizaje, no ya en robots, sino también en todos los seres vivos [k. Dautenhahn and Nehaniv, 2002].

El aprendizaje por imitación es un método supervisado, pero no invasivo, que permite el desarrollo de modelos perceptivos biológicamente inspirados, como las $Affordances$ descritas por $Gibson$ [Gibson, 1979]. Según este autor, la experiencia previa del ser humano con objetos del entorno hace posible entender cómo relacionarse con otros elementos que desconoce, todo ello basándose en las propias limitaciones de la persona y en la forma concreta de estos objetos. La mayor parte de los trabajos en este sentido se han realizado para tareas de manipulación o $grasping$, pero muy poco, por no decir nada, se ha realizado para integrar sistemas de aprendizaje dentro del comportamiento afectivo de un robot. Es decir, sería interesante dotar al robot con la capacidad de entender qué elementos están asociados con unas emociones u otras, o cómo, dependiendo de qué objeto se le ofrezca, un usuario puede modificar su estado anímico. Esto último impulsa el análisis y la implementación de nuevos conceptos que integren y permitan al robot aprender sobre las relaciones entre los diferentes elementos físicos y los estados emocionales del usuario, o como en otros enfoques, que considere la importancia del contexto del entorno en las mismas.

Esta Tesis Doctoral profundiza en muchos de estos conceptos, tratando de solventar algunos de los problemas antes citados. A lo largo de los diferentes capítulos de este trabajo, se trata el problema de cómo dotar al robot de capacidades afectivas, de forma que pueda interactuar con un humano de la formas más parecida a como lo hacemos en nuestra vida cotidiana. En este documento se ahonda en el comportamiento afectivo de un robot durante la comunicación, cómo disponer de un modelo de comportamiento, aprendido durante la interacción, para hacer más natural lo que de por sí, una comunicación con un robot, parece contra natura.

1.2. Objetivos

Los objetivos de esta Tesis Doctoral se centran en el desarrollo e implementación de sistemas inteligentes, capaces de aprender habilidades sociales por medio de la adquisición y proce-

samiento de la información de su entorno de un modo similar al humano. Específicamente, el trabajo presentado está enfocado en el desarrollo de sistemas que sean capaces de comprender y responder a gran parte de las intenciones y emociones humanas durante una comunicación natural entre un humano y un agente robótico antropomórfico. Tanto el lenguaje natural usado en la interacción como las acciones llevadas a cabo sobre diferentes elementos del entorno son usados en este trabajo para dotar al robot con capacidades emotivas.

La principal hipótesis que sustenta esta Tesis Doctoral es que un robot puede llegar a ser capaz de aprender a interaccionar de una forma natural, modificando su respuesta emocional según el contexto de la interacción y llegado el caso, facilitar un cambio emocional en el interlocutor. Todo ello conlleva el desarrollo de una serie de algoritmos que permitan no sólo el reconocimiento del estado emocional del humano, sino ser capaz de generar un conjunto básico de emociones para tal fin. A su vez, es necesario facilitar al robot con capacidades de aprendizaje de comportamiento emocionales, de forma que sus acciones y reacciones puedan ser adaptadas al contexto de la comunicación.

El uso del lenguaje natural para reconocer e imitar la información emocional del usuario por medio de agentes sociales se presenta como uno de los retos principales en este trabajo. A lo largo de esta Tesis se intenta cubrir una gran gama de formas de comunicación, tanto verbales como no verbales, mediante sistemas de reconocimiento que analicen este tipo de información afectiva y sean capaces de responder a través de las capacidades y limitaciones físicas del robot. Además, se establece un marco para la generación de emociones por parte del robot, de forma que éste pueda llegar a imitar las emociones detectadas previamente. Este sistema de generación de emociones se plantea como un sistema no cerrado, adaptable al uso de cualquier robot que disponga de la capacidad de expresar emociones mediante movimientos faciales, gestos corporales o el habla.

Sin lugar a dudas, el gran reto perseguido en esta Tesis Doctoral consiste en dotar al robot con la capacidad de interaccionar de forma natural con un usuario concreto. Ello conlleva el uso, por parte del robot, de información afectiva. De esta forma, se ha incluido como objetivo principal la definición de un sistema de aprendizaje del comportamiento afectivo del robot que tome como base la observación del usuario y su propia relación con el entorno.

Por último, para completar el estudio y probar el sistema de aprendizaje propuesto, se plantea la realización de una serie de experimentos reales para cada uno de los sistemas a desarrollar en este trabajo. De los resultados de estos experimentos se espera obtener interesantes conclusiones que pueden abrir nuevas hipótesis y líneas de trabajo futuro.

1.3. Contribuciones

Las contribuciones de esta Tesis Doctoral están relacionadas con el campo científico de la Interacción Humano-Robot. A lo largo del trabajo presentado se han realizado diferentes aportaciones que se concretan en sistemas y métodos que hacen uso, durante una interacción real, de información afectiva, agentes robóticos antropomórficos y elementos del entorno. A continuación se describen las principales contribuciones de esta Tesis:

I Reconocimiento e imitación de emociones

En una comunicación entre humanos, la información transmitida entre los interlocutores no sólo es información verbal, sino que un alto contenido del mensaje lo ofrece la comunicación no verbal. En el caso de interacciones entre robots y humanos, para dotar a las

mismas de naturalidad, se ha desarrollado un reconocedor de emociones multimodal. Este sistema toma diferentes fuentes de información para extraer la emoción humana. La expresión facial, los movimientos corporales, así como el habla del interlocutor durante una comunicación natural, son analizados de forma independiente para extraer características visuales y auditivas que permitan estimar el estado emocional entre un conjunto discreto. Como aportaciones principales, se han implementado dos sistemas de reconocimiento de emociones independientes para la expresión facial. El primero de ellos hace uso de la imagen de color obtenida por una cámara estándar. El segundo de estos sistemas emplea una cámara RGB-D para usar la información de profundidad en el proceso de estimación.

Una de las formas más simple de comunicación entre un humano y un robot es aquella en la que el robot realiza una imitación de los movimientos y expresiones del humano. Basándose en el lenguaje natural, esta Tesis ha implementado un sistema de generación de emociones que es usado como sistema de imitación durante una comunicación afectiva. La principal aportación en este punto es el modelado de cada una de las emociones básicas como un conjunto de movimientos faciales y del cuerpo del robot, así como modificaciones en el habla sintetizada. Este sistema de generación de emociones no está ligado a ningún agente en particular y puede ser usado por cualquier robot antropomórfico, adaptándose a las propias limitaciones impuestas por el diseño del mismo.

II Sistema de aprendizaje de comportamientos afectivos basado en *affordances emocionales*

La principal aportación de esta Tesis Doctoral es el sistema de aprendizaje de comportamientos afectivos para un robot autónomo. Este sistema hace uso de un concepto introducido en este trabajo, las *affordances* emocionales, para representar la relación existente entre los objetos del entorno, la reacción que debe tener el robot ante los mismos y el efecto sobre el usuario de dicha reacción. Visto de otro modo, las *affordances* emocionales generan una relación real entre los estados emocionales del usuario y los cambios que sufre dependiendo de los múltiples elementos del entorno y de la propia respuesta emocional del agente robótico. Con este sistema de aprendizaje se consigue un primer paso para orientar de forma autónoma el estado emocional del interlocutor durante una interacción.

Dentro de este mismo tópico, una segunda contribución relevante está asociada al uso de modelos de comportamiento emocional en el robot, los cuales utilizan el sistema de aprendizaje de esta Tesis para orientar el estado del interlocutor durante la comunicación.

III Experimentos

Por último, el conjunto de experimentos tratados en esta tesis permitieron verificar y cuantificar las diferentes técnicas y métodos presentados en cada capítulo en un entorno controlado (escenario afectivo) y con usuarios no entrenados.

1.4. Estructura del documento

Esta Tesis Doctoral está organizada en una serie de partes diferenciadas, cada una de ellas divididas a su vez en un conjunto de capítulos. Estos capítulos desarrollan los contenidos asociados, tanto los conceptos básicos necesarios para su comprensión, como la revisión de la

literatura relacionada con el trabajo realizado y, en su caso, los experimentos que demuestran la funcionalidad de cada sistema. En las dos partes que describen los principales enfoques planteados en esta Tesis, se ha optado por incluir la revisión del estado del arte al comienzo de las mismas, en lugar de realizar una revisión de la literatura en cada capítulo.

I Sistemas de reconocimiento e imitación de emociones

Los sistemas de reconocimiento e imitación descritos en la Parte I tienen como propósito la adquisición de la información emocional del usuario a través de tres diferentes enfoques basados en el lenguaje natural y un enfoque multimodal. Estos enfoques analizan la información asociada a las expresiones faciales, lenguaje corporal y la voz humana, respectivamente, para estimar el estado emocional, el comportamiento y los movimientos de las articulaciones del usuario que puedan ser imitadas a posteriori por las habilidades físicas de un agente robótico social.

En el caso de los Capítulos 3, 4, 5 y 6, se describen los sistemas implementados para el reconocimiento del estado emocional del usuario mediante diferentes modos de comunicación, verbal y no verbal. Por un lado, los Capítulos 3 y 5 se basan en el análisis del lenguaje corporal (comunicación no verbal), a través del análisis de la expresión facial y de los movimientos corporales. Las características extraídas en el análisis, así como los experimentos llevados a cabo son descritos en ambos capítulos. Por otro lado, el Capítulo 4 realiza un proceso de extracción y clasificación similar a los presentados en los capítulos anteriores, utilizando en este caso la comunicación verbal y características acústicas extraídas de los elementos claves de la prosodia de la voz. El Capítulo 6 esta basado en un sistema de reconocimiento multimodal, que utiliza tanto la información visual relacionada a las expresiones faciales, como la información verbal asociada a la voz para estimar el estado emocional del usuario. Finalmente, el Capítulo 7 presenta el método de imitación y generación de emociones aportado en este trabajo.

Los experimentos descritos en esta parte son realizados tanto con usuarios no entrenados en comunicaciones reales como con bases de datos obtenidas de la literatura.

II Sistema de aprendizaje de comportamientos afectivos basado en *affordances* emocionales

La Parte II de esta Tesis Doctoral describe el sistema de aprendizaje de comportamientos afectivos implementado. El término *affordances* emocionales es introducido en el Capítulo 8, así como su diferencia con respecto a las tradicionales *affordances* perceptuales usadas en la literatura para el aprendizaje, por parte del robot, de tareas de manipulación de objetos del entorno. En el Capítulo 9 se presenta en detalle la metodología empleada en esta teoría por medio de todos los elementos ($i.e.$, objetos, marcas, agentes humanos y robóticos) necesarios en el proceso de aprendizaje y evaluación dentro de un escenario afectivo real.

En el Capítulo 10 se describe el desarrollo e implementación de los diferentes experimentos relacionados con las *affordances* emocionales, así como los resultados obtenidos dentro de un escenario afectivo. La estructura propuesta en este capítulo se centra en un método de aprendizaje que permita al robot afectar u orientar el estado emocional del usuario por medio de estímulos afectivos.

III Conclusiones y trabajo futuro

La Parte III resume el trabajo realizado durante esta Tesis Doctoral. En el Capítulo 11 se describen las conclusiones más relevantes de este documento, además de proveer una descripción detallada de las principales contribuciones, mientras que el Capítulo 12 describe las líneas de trabajo futuro en este campo.

IV Publicaciones

El Capítulo 13 describe la lista de trabajos publicados, según su orden cronológico, durante el desarrollo de esta Tesis Doctoral.

V Apéndice, Índice y Bibliografía

Esta Tesis concluye con un apéndice, un índice alfabético y la bibliografía consultada durante el desarrollo del trabajo. Por su parte, el apéndice se divide en A, B y C, que describen de forma detallada aquellas librerías externas usadas en la implementación de cada uno de los sistemas desarrollados, mientras que el índice alfabético ayuda al lector a encontrar los términos incluidos en este documento, y que son resaltados en el texto con el símbolo $\surd$ en los margenes.

$\surd$

Parte I

Sistemas de reconocimiento e imitación de emociones

Capítulo 2

Estado del arte

2.1. Emociones humanas

Dentro del campo de la psicología, durante décadas los investigadores han intentado llegar a una comprensión común acerca de ¿$Qué\ es\ una\ emoción$?, o qué se puede considerar una $emoción$. Realmente no existe una comprensión global de este término, a pesar de que cada humano tiene un conocimiento relativo del concepto, y que incluso tiene la capacidad innata de reconocer y generar físicamente estas emociones.

En la literatura, la definición general de este concepto suele referirse a *un estado afectivo, a menudo acompañado de características fisiológicas y mentales específicas, que afectan directamente los pensamientos y el comportamiento en los humanos*, si bien esta definición puede llegar a considerarse incompleta, ya que realmente no cubre el alcance total de este concepto, que abarca desde cambios biológicos, fisiológicos y del comportamiento, hasta nuestra propia percepción de las cosas, los organismos y del entorno.

Por este motivo, la investigación acerca de este tema ha evolucionado constantemente en búsqueda de una definición formal del concepto de $emoción$ desde sus orígenes, sus efectos, sus expresiones, y su uso en la vida diaria. Esta evolución es perceptible al analizar los estudios de las emociones desde los grandes filósofos como Aristóteles, Descartes [Descartes, 1647], Spinozza [Spinoza, 1677], James [James, 1884] o Damasio [Damasio, 2003]. Cada una de estas teorías entregan una visión general de las emociones desde diferentes puntos de vistas asociados a múltiples campos de investigación, como la psicología, la medicina o la sociología, entre otros.

Dentro de lo que se conoce como la *Teoría de las Emociones*, se pueden observar corrientes predominantes desde diferentes puntos de vistas, destacando las fisiológicas, las neurológicas y las cognitivas. Las primeras, las teorías fisiológicas, sugieren que las emociones son generadas por cambios en el cuerpo de los humanos. En las teorías neurológicas, se postula la hipótesis de que la actividad neuronal del cerebro genera respuestas emocionales. En cambio, las teorías cognitivas describen que los pensamiento y procesos mentales formulan las estados emocionales en las personas. Es importante considerar la importancia de estos principios que, a pesar de poseer enfoques diferentes, describen una relación directa entre los cambios fisiológicos y mentales con respecto a las emociones.

En este capítulo de revisión del estado de la cuestión, a pesar de que se pueden encontrar múltiples enfoques del concepto de emociones, sólo se describirán teorías que sean aplicadas o que tengan relación en el contexto de la robótica, como las mencionadas en trabajos como [Fong et al., 2003]. Estos autores hacen referencia a tres enfoques reconocidos que describen la

teoría de las emociones en humanos:

- En el primero, se cita a uno de los trabajos más extendidos y difundidos acerca de la teoría de las emociones, como es el estudio de P. Ekman recogido en su libro [Ekman, 1999]. En sus trabajos, Ekman intenta clasificar las emociones en seis categorías discretas y primarias, y las define como universales y biológicamente básicas.
- El segundo enfoque tiene como objetivo categorizar las emociones en escalas continuas y dimensiones básicas, asociadas a la $Excitación$ y la $Valencia$. Este tipo de trabajo es descrito en [Schlossberg, 1954] y [Russell, 1980].
- En el tercer enfoque se presenta una teoría que fusiona lo más importante de los estudios antes descritos, haciendo uso de categorías discretas y escalas continuas. Este último se describe en trabajos como [Plutchik and Kellerman, 1980].

Estos enfoques fueron seleccionados de la literatura debido a que representan algunas de las visiones mas significativas y complejas respecto a la teoría de las emociones. A continuación se profundiza en los contenidos de esta teoría, por estar directamente relacionado con el trabajo presentado en esta Tesis Doctoral.

2.1.1. Teoría de emociones de Ekman

La teoría de las emociones propuesta por Ekman tiene como objetivo normalizar las emociones del ser humano, esto es, categorizar las emociones en un concepto que las defina como discretas, cuantificables y físicamente distintas. Para lograr esto, Ekman se basó en el principio descrito originalmente por Darwin [Darwin, 1872], acerca de cómo las expresiones faciales son universalmente percibidas y relacionadas con una emoción específica, incluso con usuarios de diferentes culturas que no poseen un conocimiento previo acerca de qué expresión representa cada emoción. En [Ekman et al., 1983], se corroboró este principio por medio de un experimento donde una serie de actores representaban diferentes expresiones faciales, mientras se analizaban los cambios psicológicos y físicos para cada una de estas expresiones.

Este experimento arrojó como resultado que las emociones son reconocidas universalmente, incluso sin un aprendizaje que asocie las expresiones faciales con emociones específicas. Esto último permitió a Ekman describir una conjunto de seis emociones básicas, con características biológicas y psicológicas únicas, que se corresponden con el enfado, el miedo, la felicidad, la tristeza, el disgusto y la sorpresa.

2.1.2. Teoría de emociones de Russell

En el caso del enfoque dado por Russell, se afirma que es posible relacionar los espacios emocionales en un sistema de referencia de dos dimensiones. En este sistema, el primer eje X estará asociado a la $Valencia$ de las emociones en un orden del positivo (0^o) al negativo (180^o), mientras, en el eje Y estará asociado al nivel de $Excitación$ (llamado a veces de $Activación$) en un orden del positivo para altos niveles de excitación (90^o) y negativo para bajos niveles de excitación (270^o). Este modelo que categoriza las emociones de acuerdo a la valencia y el nivel de excitación es denominado *modelo circumplejo del afecto de Russell*, y se ilustra en la Figura 2.1.

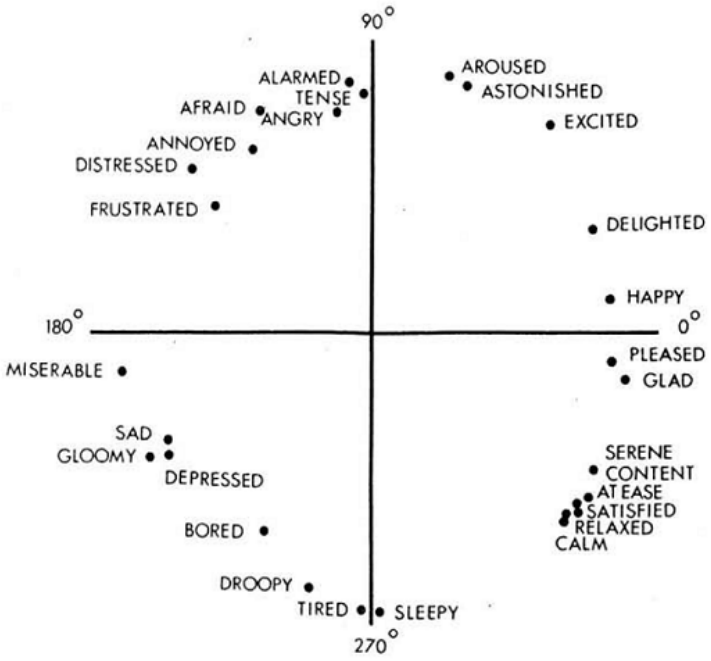


Figura 2.1: Modelo Circumplejo del afecto de Russell; La valencia es representada por 0^o (positiva) y 180^o (negativa), mientras, el nivel de excitación es representado por 90^o (elevado) y 270^o (bajo) (Figura adquirida de la publicación [Russell, 1980])

A continuación, y basándose en esta teoría, se describen brevemente las dimensiones utilizadas para representar y categorizar las emociones a partir de los niveles de activación y valencia:

1. *Valencia:* es un concepto que permite evaluar de forma positiva o negativa una emoción por parte de un usuario [Lang et al., 1990], basándose en la percepción y reacción de estímulos externos. Este concepto sigue la hipótesis de que estados emocionales con la misma valencia están asociados a respuestas emocionales similares por parte del humano.

   Las emociones con valencia positiva, dentro de los cuadrantes $C1$ y $C4$, están asociados a estados emocionales como la felicidad, la sorpresa o la calma. Por su parte, las emociones con valencia negativa, dentro de los cuadrantes $C2$ y $C3$, están asociados a estados emocionales como el miedo, el enfado o la tristeza.

2. *Nivel de excitación:* es un factor cuantitativo utilizado para evaluar el nivel de excitación de una emoción, relacionado a factores perceptibles como los movimientos, las expresiones y otras respuestas físicas.

   En el caso de las emociones con un alto nivel de excitación, dentro de los cuadrantes $C1$ y $C2$, están asociados estados emocionales como la felicidad, la sorpresa o el miedo. Mientras, en las emociones con bajo nivel de excitación, dentro de los cuadrantes $C3$ y $C4$, se encuentran los estados emocionales como la calma, el disgusto o la tristeza.

A pesar de los descrito anteriormente, lejos de las condiciones de los estudios realizados en muchas otras publicaciones, el nivel de excitación o valencia dentro de cada estado emocional suele variar entre diferentes personas, en algunos casos debido el entorno. No obstante, esta variación sólo se refiere al nivel de la valencia y de excitación, sin cambiar de cuadrante en el

Cuadrante	representación Valencia/Intensidad
C1	Valencia Positiva (0^o) y Nivel de excitación Positivo (90^o)
C2	Valencia Negativa (180^o) y Nivel de excitación Positivo (90^o)
C3	Valencia Negativa (180^o) y Nivel de excitación Negativo (270^o)
C4	Valencia Positiva (0^o) y Nivel de excitación Negativo (270^o)
Neutral	zona de valencia y nivel de excitación neutral

Cuadro 2.1: Clasificación de los estados emocionales según su cuadrante en el modelo de Russell [Russell, 1980].

modelo de Russell (Ver Cuadro 2.1). Esto suele estar justificado en los estudios de Darwin que describen el hecho de que los estados emocionales conllevan una cierta disposición a actuar de una manera determinada.

Finalmente, en los trabajos de Russell [Russell, 1980] se describen una cantidad de 28 palabras ($affect\ word$ [Whissell et al., 1986]) que representan los estados emocionales de los humanos en el modelo circumplejo (Ver Figura 2.1). Sin embargo, estos estados emocionales no son todos emociones discretas y varían constantemente durante el tiempo, lo cual demuestra la importancia de otros enfoques como el presentado por Ekman y Freisen, que normalizan las emociones en únicamente seis categorías primarias diferentes.

2.1.3. Teoría de emociones de Plutchik

Dentro de la literatura, este tercer enfoque fue descrito por R. Plutchik [Plutchik and Kellerman, 1980], y se presenta como uno de los estudios más completos y analizados en la psicología actual. A diferencia de otros autores, Plutchik no considera a las emociones como simples estados emocionales asociado al concepto de $Emociones\ Básicas$, sino que analiza este término como un tema aun más complejo. Esto se debe a que las emociones presentan características únicas que les permiten evolucionar a lo largo del tiempo, con el fin de elevar las capacidades reproductivas y de supervivencia de cada organismo, describiendo que las emociones biológicamente primitivas suelen ser el detonante en conductas asociadas a situaciones de supervivencia. Por ejemplo, la respuesta del miedo en situaciones de peligro nos inspira a respuestas como escapar, pelear o simplemente no moverse.

Para explicar lo anterior, Pluchik describe diez postulados en los que se apoya su teoría de la evolución de las emociones básicas:

1. *Animales y humanos:* el concepto de emoción es aplicable a todos los niveles evolutivos y se aplica tanto a animales como humanos, debido a que ambos experimentan las mismas emociones de forma similar.

2. *Historia evolutiva:* las emociones poseen una historia evolutiva y han evolucionado en diversas formas de expresión en las distintas especies.

3. *Principios de supervivencia:* las emociones cumplen una función adaptativa en la tarea de ayudar a los organismos a lidiar con los problemas fundamentales de la supervivencia que plantea el medio ambiente.

4. *Patrones prototipo:* a pesar de las diferentes formas de expresión de las emociones en diferentes especies, existen ciertos elementos comunes, o patrones, que se pueden identificar.

5. *Emociones básicas:* hay un pequeño número de emociones básicas o primarias.

6. *Combinaciones:* todas las demás emociones son estados mixtos o derivados; es decir, que se producen como combinaciones, mezclas o compuestos de las emociones primarias. Por ejemplo, el amor es una combinación de alegría (emoción primaria) y confianza (emoción primaria).

7. *Construcciones hipotéticas:* se reconoce que las emociones primarias son construcciones hipotéticas o estados idealizados cuyas propiedades y características sólo se pueden inferir de varios tipos de pruebas.

8. *Opuestos:* las emociones primarias se clasifican en pares de opuestos polares. Por ejemplo, la alegría es lo opuesto a la tristeza, anticipación es lo opuesto de sorpresa y el disgusto es lo opuesto de confianza.

9. *Similitud:* todas las emociones tienen diferentes grados de similitud entre sí.

10. *Intensidad:* cada emoción puede existir en diversos grados de intensidad o niveles de excitación.

Siguiendo estos diez postulados, la teoría de Plutchik trabaja bajo el concepto de ocho emociones básicas, que se concretan en la alegría, la confianza, el miedo, la sorpresa, la tristeza, la anticipación, la ira y el disgusto. Todas estas emociones primarias se pueden combinar para generar emociones secundarias o terciarias denominadas $Dyad$. No obstante, la existencia de relaciones entre múltiples emociones asociadas a postulados como emociones opuestas o combinadas, llevo a Plutchik a crear la rueda de las emociones que se ilustra en la Figura 2.2. En esta figura se observa cómo la intensidad de las emociones disminuye a medida que se aleja del centro, y aumenta a medida que se acerca al mismo. Visualmente esto se representa a través de la intensidad del color, siendo más oscuro el color de la emoción con una mayor intensidad.

2.1.4. Comparativa

En la literatura relacionada con este campo de investigación, es común encontrar múltiples estudios asociados a la teoría de las emociones. Como se ha comentado anteriormente y como se ha podido comprobar con los tres enfoques analizados, incluso entre los propios investigadores más relevantes se discrepa con respecto al número exacto de emociones básicas en los humanos. Esta discrepancia entre el número de emociones básicas es recogida por algunos autores en sus resúmenes, como es el caso del trabajo presentado en [Ortony and Turner, 1990], donde se analiza las diferentes emociones consideradas por diferentes autores, como se ilustra en el Cuadro 2.2. En la literatura también es posible encontrar revisiones similares centradas en ciertos campos específicos, como es el caso de [Cowie and Cornelius, 2003], que se basaba en el análisis del habla y su relación con las emociones.

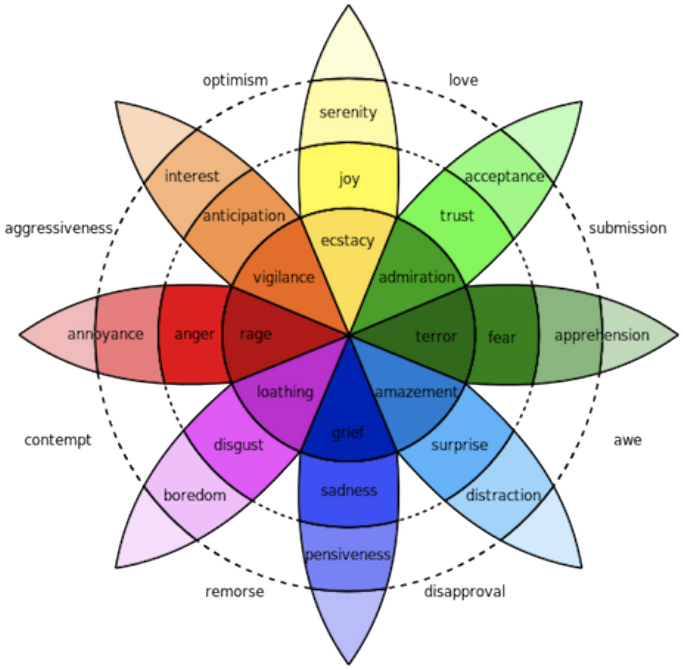


Figura 2.2: Rueda de las emociones de Plutchik (Figura obtenida desde la publicación [Plutchik, 2002]).

2.1.5. Las emociones en Interacciones humano-computador y humano-robot

√ Dentro de la interacción Humano-Computador (***IHC***), el término *computación afectiva* (*affective computing*) descrito por R. Picard [Picard, 2000], representa el concepto de que las maquinas necesitan ser capaces de reconocer, interpretar e incluso imitar emociones humanas. El objetivo detrás de esta hipótesis es que los sistemas informáticos precisan percibir información visual o auditiva, de forma similar a la humana, para identificar las emociones del usuario.

El uso de la información emocional por parte de sistemas computacionales está destinado a proveer una interacción natural con el usuario, mediante el uso de modelos de comportamiento social que intenten emular la interacción entre organismos vivos. Para ello se precisa de sistemas que adquieran información emocional de los usuarios, y con respecto a esto, es necesario el uso de sensores pasivos que capturen toda la información visual y acústica relacionada con los movimientos y el comportamiento de los humanos. Esta información será posteriormente analizada y procesada para obtener datos acerca de los elementos de la comunicación verbal (por ejemplo, la voz) o no verbal (como las expresiones faciales o el lenguaje corporal) que están afectados directamente por las emociones.

√ Por otro lado, en una interacción humano-robot (***IHR***), esta información emocional adquiere una característica especial, debido a que los robots pueden no sólo reconocer este tipo de información, sino también generar la misma, mejorando significativamente la percepción de la naturalidad por parte del usuario. Estas emociones generadas por los agentes, denominadas *emociones artificiales*, tienen como objetivo realimentar la interacción y afectar el estado emocional del usuario, siguiendo la propia teoría de las emociones.

Tanto en las IHC como en las IHR existen diferentes formalizaciones de los estados emocionales de los humanos, los cuales son utilizados en los sistemas de reconocimiento de emociones basados en expresiones faciales, voz humana o lenguaje corporal, entre otros. A continuación, se describirán brevemente los mas importantes.

- La primera formalización está basada en los estudios de Ekman [Ekman et al., 2002], el cual presenta un sistema de categorización con seis emociones básicas (enfado, disgusto, miedo, felicidad, tristeza y sorpresa).
- La segunda formalización se basa en los trabajos de Parrott [Parrott, 2001], quien describe un método de categorización de 136 estados emocionales, divididos en grupos de emociones primarias, secundarias y terciarias. Este trabajo también cuenta con seis emociones primarias: amor, alegría, sorpresa, enfado, tristeza y miedo.
- La tercera y última formalización tiene su origen en los estudios de Plutchik [Plutchik and Kellerman, 1980], quien divide las emociones de los humanos en primarias, secundarias y terciarias, por medio de su propio sistema de categorización denominado *Rueda de las emociones de Plutchik*. Este sistema cuenta principalmente con 8 emociones básicas, la alegría, la confianza, el miedo, la sorpresa, la tristeza, la anticipación, la ira y el disgusto.

Entre estas formalizaciones, el enfoque más utilizado dentro del campo del reconocimiento de emociones está ligado con los trabajos de Ekman, debido principalmente a la enorme implicación de sus estudios en la normalización de las emociones dentro de los sistemas de reconocimiento basados en expresiones faciales.

Por último, el objetivo final en la incorporación de este tipo de información emocional dentro de la IHR (o IHC) es mejorar aspectos como la empatía, la atención, o la naturalidad de la interacción. Para ello, normalmente estos sistemas, luego de reconocer una emoción, actúan según diálogos basados en normas sociales y con diferentes modalidades de comunicación (por ejemplo, la voz o el lenguaje corporal). La influencia en la comunicación del intercambio de información emocional entre un agente robótico y las personas es analizado en algunos trabajos como [Paiva et al., 2004] y [Breazeal, 2002] .

2.2. El lenguaje natural en IHR afectivas

En la actualidad, las nuevas técnicas desarrolladas para interacciones entre robots y humanos se centran en la utilización de algoritmos y métodos que permiten al usuario comunicarse y, por tanto, relacionarse con los robots de forma similar a como lo hacen los humanos. Del conjunto de técnicas estudiadas, el lenguaje natural desempeña uno de los papeles fundamentales, siendo el método natural de comunicación que un humano adquiere de forma espontánea desde su entorno, y que es utilizado para transferir ideas, conceptos, emociones e intenciones durante una interacción. Los lenguajes o idiomas, como el español, son considerados lenguajes naturales puesto que se originaron de forma espontánea entre un grupo de individuos como una necesidad para comunicarse. Durante años ha sido el método más estudiado para hacer más naturales las IHR, pero aún así, todavía permanecen varios aspectos relevantes sin resolver.

Para obtener una interacción más real y más creíble, los sistemas de IHR deben ser capaces de responder apropiadamente a los usuario por medio de reacciones afectivas

[Zeng et al., 2008], pero también conocer el estado emocional del usuario y actuar en consecuencia. Dado que el lenguaje natural está íntimamente relacionado con la comunicación humana, las emociones se transmiten por medio de información verbal y no verbal, o lo que es lo mismo, utilizando diferentes canales de comunicación (por ejemplo, la voz, el lenguaje corporal o los propios gestos y expresiones faciales). Es necesario, por tanto, dotar a los robots de capacidades para reconocer y generar emociones durante la interacción [Picard, 2000].

Dentro de una interacción entre un humano y un robot, los sistemas de reconocimiento de emociones hacen uso del lenguaje natural para extraer información del estado emocional del interlocutor. Muchos de estos sistemas trabajan con un único modo o canal de información, como son los sistemas basados en voz, en expresiones faciales o en el análisis del lenguaje corporal. Sin embargo, los últimos trabajos presentan sistemas multimodales, esto es, sistemas que hacen uso de más de un canal de información para estimar la emoción humana. Para ello, precisan de sistemas de sincronización de cada uno de los modos, así como la capacidad de combinar las salidas en una única emoción. Normalmente, en estos sistemas se suele utilizar un enfoque basado en una modalidad predominante, como las expresiones faciales, y reforzar la salida del sistema con la información emocional obtenida desde otras fuentes, como la voz [Sebe et al., 2005], [Jaimes and Sebe, 2005].

Por su parte, para generar emociones durante una IHR el robot también puede hacer uso de estos mismos canales de comunicación. En el caso de la voz, por ejemplo, las reacciones afectivas se logran modificando características de la prosodia (por ejemplo, el énfasis o la intensidad). De igual forma, el robot puede generar expresiones faciales o movimientos corporales a la hora de reaccionar afectivamente [Breazeal, 2002]. Está claro que el propio diseño del robot facilita estas expresiones, y por ello es importante que disponga de elementos móviles para asemejar estas emociones a las humanas.

Finalmente, es importante mencionar que el uso del lenguaje natural asociado a la información emocional en diálogos con robots no es suficiente para generar una sensación de naturalidad en el dialogo. No son sólo los canales de comunicación, ni la capacidad para percibir o generar la información emocional en una interacción lo que compone este tipo de dialogo, sino que existen unas características sociales, culturales y del propio contexto de la situación que son las que realmente generan un dialogo basado en el lenguaje natural entre seres humanos [Fong et al., 2003].

A continuación se analizan los principales canales de comunicación usados en IHR afectivas, y los principales trabajos que encontramos en la literatura.

2.2.1. Voz humana

La voz humana es uno de los medios de comunicación más eficiente para la transmisión de ideas, emociones e intenciones dentro del lenguaje natural. Este canal representa un método no invasivo para el intercambio de información, que ha sido utilizado en el mundo de la robótica en numerosas aplicaciones. Esta amplia gama de usos de la voz humana está relacionada con el hecho de que tiene la capacidad de intercambiar dos tipos de información en una interacción: la objetiva y la subjetiva. La primera información suele estar relacionada al contenido del mensaje verbal que se transmite por medio del canal. Mientras, la segunda información está relacionada a las emociones del locutor que se transmiten mediante el habla, la cual suele ser identificada por medio de cambios en las características acústicas, principalmente pertenecientes a la prosodia, tales como el $pitch$, la velocidad de la voz, el énfasis o la energía.

Por un lado, en relación a la información objetiva durante la IHR, los robots suelen estar equipados con sistemas $Text\text{-}to\text{-}Speech$ (TTS), que permiten, a partir de información de texto, generar audio sintético. Esto abre un interesante abanico de posibilidades, sin embargo, estos sistemas están limitados y hoy por hoy, la naturalidad de la comunicación suele ser bastante escasa. En casi todos los sistemas TTS, comerciales o gratuitos, el humano percibe este tipo de mensaje como un discurso monótono y artificial (no el contenido del mensaje en sí, sólo la forma de transmitirlo), lo que suele provocar bastante rechazo [Bartneck, 2002]. La voz generada no posee ningún cambio significativo relacionado con el énfasis, el volumen o el $pitch$, siendo estos cambios comunes en la voz humana. Cómo de limitado están los sistemas TTS a la hora de generar mensajes con carga emotiva respecto a otros canales como las expresiones faciales o el lenguaje corporal se ha estudiado en trabajos como [Bartneck, 2002].

Para solucionar este problema, existen estudios que intentan emular la información emocional dentro de mensajes verbales sintéticos a través de sistemas TTS. Uno de los trabajos más representativos es el desarrollado por Cahn [J.Cahn, 1990], que comienza analizando los cambios en la prosodia de la voz en cada estado emocional hasta crear un modelo que permita manipular determinados sistemas TTS y así generar contenido emocional a través de mensajes verbales. En los últimos años se han implementado técnicas que consideran las componentes emocionales en la síntesis de voz, la mayoría para el idioma inglés [Murray and Arnott, 1993], [Hoult, 2004]. En el caso del idioma español, existen estudios que centran el análisis prosódico del español y su relación con las emociones [Montero et al., 1999] , [Iriondo et al., 2000]. Algunos de estos sistemas que expresan diferentes tipos de estados emocionales por medio de la voz, fueron desarrollados para plataformas robóticas con métodos de vocalización como el del robot $Kismet$ [Breazeal, 2002].

Por otro lado, como se ha comentado anteriormente, la voz no sólo permite la transmisión de información objetiva, sino también subjetiva, relacionada con la emoción de los interlocutores. Dentro de las IHR, la mayor parte de los trabajos que estudian el reconocimiento de las emociones por el análisis del habla categorizan la información del estado emocional del usuario en seis emociones básicas, las cuales están basados en los estudios desarrollados por Ekman. Realmente, el uso de estas emociones básicas en estos trabajos no ofrece una justificación detallada de la razón de la misma, aun más si se incluyen los aspectos mas relevantes de la voz, pero se presenta como una normalización valida que ha sido desarrollada y discutida a través de múltiples estudios como [Sebe et al., 2005]. Un ejemplo del uso de las emociones básicas dentro de los sistemas de reconocimiento es posible encontrarlo en trabajos como [Murray and Arnott, 1993], donde se estudia la correlación existente entre algunos estados emocionales y las características acústicas pertenecientes a la prosodia de la voz humana. Esta correlación es descrita brevemente en el Cuadro 2.3, que presenta cómo los elementos cuantitativos de la voz pueden ser asociados con estados emocionales a través ensayos prácticos.

Por su parte, la adquisición de la información verbal en este tipo de sistemas usualmente está basado en la cuantificación de la energía en la señal de voz, de forma que se pueda discriminar desde la señal de audio, los ruidos, la música, el sonido ambiente o la voz del usuario que cumple con el rol de interlocutor durante la interacción. En el caso de las características elegidas por los investigadores para estimar la emoción a partir de la voz, existen estudios que definen dos grupos de características auditivas [Iliou and Anagnostopoulos, 2010] y [Schuller et al., 2004]: en el primer grupo se encuentran las características de alto nivel, es decir, aquellas relacionadas con el propio mensaje, o con características espectrales de la señal de voz. Estos sistemas, si bien dan buenos resultados, tienen un fuerte problema con la dependencia que presentan a los

fonemas, al hardware de adquisición de señal acústica, y el propio contenido de cada expresión [Atassi et al., 2011]. En el segundo grupo están las características de bajo nivel, esto es, aquellas relacionadas con elementos de la prosodia [Moberg, 2007]. Gran parte de los trabajos en el reconocimiento de emociones basados en voz suele trabajar con variables asociadas con características de bajo nivel [Nogueiras et al., 2001], por su bajo coste computacional y los propios resultados alcanzados.

2.2.2. Expresiones faciales

Las expresiones faciales se pueden definir como un conjunto de distorsiones de la actividad muscular, en zonas específicas de la cara, para expresar una idea o un concepto. Sin lugar a dudas, representan una de las fuentes de información emocional mas completas y robustas dentro del lenguaje natural de los humanos. Dentro del concepto de expresiones faciales, uno de los investigadores mas reconocidos de este campo como es P. Ekman, quien describe que las expresiones faciales son universales y representados por medio de estados emocionales específicos basados en seis emociones básicas, tal y como fue presentado en la sección anterior. Esta característica de las respuestas faciales fue descrita en su trabajo [Ekman and Friesen, 1971], por medio de la siguiente declaración "*las seis emociones estudiadas eran las que se habían encontrado por más de un investigador como discriminables dentro de una misma cultura letrada*".

Las expresiones faciales juegan un rol clave en el intercambio de información emocional entre humanos y robots, y parte de esta importancia la motiva la normalización y comprensión de las emociones dadas por Ekman. En los últimos años se ha desarrollado considerablemente la técnica y hoy por hoy se consiguen tasas de aciertos elevadas en la detección de emociones. Además, el constante desarrollo de robots sociales diseñados con características antropomórficas o caricaturizadas facilitan la interacción por medio de la generación de expresiones faciales, a través de los elementos de la cara (como la boca, las cejas, los párpados o los ojos, por ejemplo) [Breazeal, 2002], [Cid et al., 2014].

√ El reconocimiento de las expresiones faciales está, en general, basado en el sistema ***FACS*** (*Facial Action Coding System*), que se ha impuesto como el estándar en el análisis de las características faciales para la estimación del estado emocional del usuario. Este sistema trabaja identificando y categorizando el comportamiento de la actividad muscular facial de los humanos en cada expresión facial, a través de las denominadas Unidades de Acción AUs (*Action Units*), siendo cada AUs una distorsión facial específica causada por la actividad muscular de un pequeño grupo de músculos faciales. El reconocimiento de cada expresión facial está basado en la cuantificación de las deformaciones de los músculos faciales en cada expresión, donde las deformaciones musculares afectan visiblemente a los elementos de la cara, como la boca, las cejas, entre otros. Así, normalmente se definen las expresiones faciales como un conjunto especifico de AUs. Dada la importancia que tienen en esta Tesis Doctoral, una descripción más detalla del sistema FACS se encuentra en la Sección A.1 del Apéndice A.

La literatura relacionada al reconocimiento de emociones es muy extensa, pues son décadas trabajando en este tema. Existen resúmenes interesantes que desglosan paso a paso la mayor parte de los algoritmos, citando autores y métodos diferentes, y aportando estudios comparativos. Un excelente trabajo, tomado como base en esta Tesis Doctoral es el presentado en [Bettadapura, 2012]. La mayor parte de los sistemas presentan un esquema similar: parten de la información visual, y de ahí extraen un conjunto de características del rostro de la persona. Una vez se obtienen estos elementos principales de la cara, se suele incluir un bloque clasi-

ficador que, dada las características a la entrada, ofrezca una estimación de la emoción a la salida. Los sistemas de reconocimiento se diferencian en los algoritmos o estrategias concretos de cada una de estas fases (en el tipo de características usadas, en el clasificador elegido, etc) [Bettadapura, 2012]. En este mismo informe se destaca cómo existen diferentes partes en el rostro que permiten clasificar más fácilmente las emociones, como son las cejas o la comisura de los labios. Como punto de partida, los sistemas de detección y clasificación de emociones se dividen en tres clases fundamentales: aproximaciones basadas en el flujo óptico, detección y seguimiento de características y aproximaciones basadas en el alineamiento del modelo. En realidad, hoy en día la mayor parte de los algoritmos son una combinación de todas estas técnicas e incluyen un clasificador a la hora de estimar la emoción.

La aproximación basada en el flujo óptico usa campos de movimiento densos calculados en áreas específicas de la cara tales como la boca y los ojos. Intenta relacionar los vectores de movimiento con las emociones faciales usando plantillas de movimiento extraídas sobre un conjunto de campos de movimiento de entrenamiento y aplicando posteriormente Modelos ocultos de Markov (por ejemplo, los trabajos pioneros [Tsapatsoulis et al., 1999] y [Otsuka and Ohya, 1997] entre otros, o los más recientes como [Anderson and Owan, 2003]). En la segunda aproximación la estimación de la emoción se obtiene a partir de un conjunto pequeño de características relevantes en la escena. El análisis se realiza en dos etapas: en primer lugar se procesa el $frame$ del vídeo para la detección de las características necesarias, (los ojos, la nariz, la boca...), posteriormente se analiza el movimiento de dichos elementos (como el trabajo [Bartlett et al., 2005], utilizando la transformada $Wavelets$ de $Gabor$ o el trabajo [Aleksic and Katsaggelos, 2006], utilizando características del rostro extraídas en el contorno de los labios y cejas, seguido de PCA - *Principal Component Analysis*). En definitiva, se basa en la obtención de elementos característicos que fácilmente puedan ser asociados a las AUs o combinaciones de las mismas. En base a ellas, los autores obtienen la estimación de la emoción detectada. La tercera aproximación alinea un modelo 3-D de la cara para estimar tanto el movimiento del objeto como la orientación, como los trabajos [Kotsi et al., 2008], realizando ambos medidas sobre el modelo $Candide-3$. Un análisis más extenso puede verse en el trabajo antes mencionado [Bettadapura, 2012].

2.2.3. Lenguaje corporal

Los últimos estudios referentes al reconocimiento de emociones por medio de la información visual, han llevado a los investigadores a tomar en consideración un modo de comunicación muy poco explorado, como es el lenguaje corporal durante la interacción. La relevancia de este lenguaje con respecto a otras fuentes de información, como el audio, se debe a que el lenguaje corporal posee una correlación directa con respecto a las expresiones faciales, como se ha estudiado en los trabajos [Meeren et al., 2005] y [Peelen and Downing, 2007]. Esta relación describe cómo el usuario expresa las mismas emociones por medio de las expresiones y el lenguaje corporal durante la rutina diaria, lo que causa que el reconocimiento de emociones basado en la información visual, ya sea por medio de expresiones faciales o el lenguaje corporal, genere el mismo resultado.

El reconocimiento de emociones basado en el análisis del lenguaje corporal sigue un procedimiento similar a los basados en expresiones faciales, si bien el proceso es mucho más complejo puesto que requiere procesar una mayor cantidad de información relacionada con los grados de libertad del cuerpo humano, y todo ello en tiempo real. En primer lugar, se realiza la extrac-

ción y posterior análisis de las características corporales adquiridas por el robot, tomadas principalmente de las manos, la cabeza, los hombros, y las piernas. A partir de las mismas, se utiliza algún método de clasificación para obtener la emoción. Los primeros sistemas hacían uso de cámaras RGB para adquirir la información visual, y a partir de las mismas obtenían características relacionadas con el movimiento humano y las emociones, como la cantidad de movimiento, el índice de contracción del cuerpo, la velocidad, la aceleración o la fluidez del movimiento, entre otros. (Ver [Kessous et al., 2010]). Sin embargo, en los últimos años, el desarrollo de sensores que incorporan información de vídeo RGB y profundidad han permitido un seguimiento y un análisis conciso de la información relacionada a la posición 3D, lo que permite analizar de forma dinámica la silueta 3D del usuario, abriendo un nuevo abanico de posibilidades.

Uno de los grandes problemas para valorar el uso del lenguaje corporal como fuente de información en el reconocimiento de emociones, está relacionado con el hecho de que la mayor parte de los estudios analizados fueron realizados mediante actores, lo que no arroja resultados significantes para su uso en aplicaciones reales. Aún así, revisiones como la presentada en [Fong et al., 2003], describen que más del 90 % de los gestos corporales se producen durante el proceso del habla, proporcionando gran parte de la información redundante necesaria en la interacción [Krauss et al., 1991], [McNeill, 1992].

Por otro lado, en relación con la generación de reacciones afectivas, en [Nakata et al., 1998] se describe un ejemplo interesante del uso del lenguaje corporal para expresar estados emocionales por medio de un robot. El robot, en este caso, expresa impresiones emocionales mediante varios bailes que son medidos y comparados con las categorías recogidas en el Análisis de Movimiento Laban [Laban, 1980]. Esta teoría reciente relacionaba el movimiento en la danza con las emociones, que se ha extendido igualmente a otros campos de aplicación, incluida la IHR.

Finalmente, en el Cuadro 2.4 se resumen los niveles de activación y valencia con respecto a movimientos o gesto específicos. Estos datos fueron analizados desde los estudios mas relevantes en este campo [Kessous et al., 2010]). Mientras, extraído del trabajo [Fong et al., 2003] se presenta el Cuadro 2.5, que describe la relación entre algunos movimientos del cuerpo con respecto a un limitado número de estados emocionales, siendo estos movimientos, en muchos casos, involuntarios y donde se genera una gran cantidad de intensidad emocional que los hace visibles a simple vista.

	Emociones básicas	Criterios de inclusión
Plutchik [Plutchik and Kellerman, 1980]	Aceptación, enfado, anticipación, disgusto, alegría, miedo, tristeza, sorpresa	Relación con procesos biológicos adaptativos.
Arnold [Arnold, 1960]	Enfado, aversión, coraje, abatimiento, deseo, desesperación, miedo, odio, esperanza, amor, tristeza	Relación con las tendencias de acción.
Ekman [Ekman et al., 2002]	Enfado, disgusto, miedo, alegría, tristeza	Expresiones faciales universales
Frijda [Frijda, 1986]	Deseo, felicidad, interés, sorpresa, asombro, tristeza	Formas de acción disponibles
Gray [Gray, 1982]	Rabia y terror, ansiedad, alegría	Estructurado (invariante)
Izard [Izard, 1971]	Enfado, desprecio, disgusto, angustia, miedo, culpa, interés, alegría, vergüenza, sorpresa	Estructurado (invariante)
James [James, 1884]	Miedo, dolor, amor, rabia	Implicación corporal
McDougall [McDougall, 1926]	Enfado, disgusto, euforia, miedo, sumisión, emoción tierna, asombro	Relación con los instintos
Mowrer [Mowrer, 1960]	Dolor, placer	Desaprender los estados emocionales
Oatley [Oatley and Johnsin-Laird, 1987]	Enfado, disgusto, ansiedad, felicidad, tristeza	No requieren contenido proposicional.
Panksepp [Panksepp, 1982]	Esperanza, miedo, rabia, pánico	Estructurado (invariante)
Tomkins [Tomkins, 1984]	Enfado, interés, desprecio, disgusto, angustia, miedo, alegría, vergüenza, sorpresa	Densidad de actividad neuronal.
Watson [Watson, 1930]	Miedo, amor, rabia	Estructurado (invariante)
Weiner [Weiner and Graham, 1984]	Felicidad, tristeza	Atribución independiente

Cuadro 2.2: En [Ortony and Turner, 1990] se presenta una lista de emociones básicas seleccionadas de estudios relevantes (Cuadro obtenido de la publicación [Ortony and Turner, 1990]).

Emociones	Vel. de la voz	Promedio del $Pitch$	Rango del $Pitch$	Intensidad	Calidad de la voz
Enfado	Ligeramente rápido	muchísimo más rápido	muy amplio	Elevada	Procede de la respiración
Felicidad	rápido o lento	muy elevado	muy amplio	Elevada	A todo volumen
Tristeza	Ligeramente lento	Ligeramente lento	Ligeramente mas estrecho	reducido	resonante
Miedo	Muy rápido	muchísimo más rápido	Elevada	Normal	irregular
Disgusto	muchísimo más lento	muchísimo más lento	Ligeramente mas amplio	reducido	retumbante

Cuadro 2.3: Relación entre elementos cuantitativos de la voz humana con respecto a estados emocionales básicos, siendo cuantificados en relación aun estado emocional de la voz neutro (Cuadro obtenido de la publicación [Sebe et al., 2005]).

Emociones	Valencia	Activación	Gesto
Enfado	Negativa	Alta	Descenso/Ascenso violento de las manos y brazos.
Felicidad	Positiva	Alta	Movimiento rápido del cuerpo circulares.
Tristeza	Negativa	Baja	Desplazamiento lento de las manos y el cuerpo.
Miedo	Negativa	Baja	Movimiento de las manos hacia la cara, contracción del cuerpo.
Disgusto	Negativa	Baja	Movimiento de las manos hacia fuera $dejadme\ sólo$.

Cuadro 2.4: Movimientos corporales relacionados a estados emocionales básicos.

Emociones	Movimientos del cuerpo
Enfado	Mirada feroz; puños apretados; paso ligero; movimientos cortos.
Miedo	Doblar la cabeza; hombros encogidos; obligado a mantener los ojos cerrando o mirando fijamente.
Felicidad	Acelerado; movimientos aleatorios; sonriendo.
Tristeza	Comisura de los labios hacia abajo; llorando.
Sorpresa	Ojos muy abiertos; respiración retenida; la boca abierta.

Cuadro 2.5: Movimientos corporales relacionados a estados emocionales, siendo información original del estudio [Frijda, 1986] (Cuadro obtenido de la publicación [Fong et al., 2003]).

Capítulo 3

Sistema de reconocimiento de emociones basado en el análisis de las expresiones faciales

3.1. Introducción

Los robots sociales están destinados a interactuar con los usuarios de una forma similar a como lo harían dos humanos en una conversación, lo que se conoce con el nombre de lenguaje natural. Estos sistemas suelen ser amigables, intuitivos y no invasivos para los interlocutores, de forma que se busca crear una atmósfera de naturalidad durante la comunicación. En este contexto, ser capaz de reconocer las emociones de los participantes durante la conversación se convierte en un objetivo prioritario dentro de esta robótica social, lo que provoca que el reconocimiento de emociones se haya convertido en uno de los campos científicos más explorados en los últimos años dentro de las Interacciones Humano-Robot (IHR). En estos sistemas, las expresiones faciales suponen una poderosa fuente de información emocional asociada al comportamiento humano, siendo la base de un gran número de trabajos relacionados con la IHR afectiva [Zeng et al., 2008].

La mayor parte de estos trabajos se realizan a partir de la información de color capturada por cámaras acopladas al robot. La lectura de estos sensores es posteriormente procesada para extraer características faciales que puedan ser relacionadas con emociones, y normalmente, como paso final del sistema, se hace uso de algún tipo de clasificador para estimar el estado emocional del usuario. Como fue explicado en el Capítulo 2.1, existen un gran número de trabajos que se diferencian en partes concretas de este proceso, pero todos ellos se caracterizan por seguir un patrón similar [Bettadapura, 2012].

En los últimos años, la aparición de sensores de bajo coste en el mercado que combinan color y profundidad ha permitido su uso en los mecanismos de reconocimiento de emociones a partir de expresiones faciales. Disponer de información tridimensional directamente con el sensor (y no tener que recurrir a procesado de imágenes estéreo, por ejemplo) permite un nuevo abanico de características faciales que igualmente pueden ser utilizadas en el proceso de reconocimiento. Los sistemas mantienen la misma estructura, pero se aprovechan de las posibilidades que presentan este tipo de sensores.

En este capítulo se describen dos métodos diferentes para el reconocimiento de emociones basado en expresiones faciales. El primero de ellos hace uso de las imágenes capturadas por una

cámara RGB-D conectada al robot y la posibilidad de emplear la información 3D del mismo para extraer un modelo de malla de la cara, mientras que el segundo se caracteriza por el uso de un sensor RGB. Ambos sistemas basan su funcionamiento en las Unidades de Acción (AUs), descritas en el Apéndice A.1. Como ya se comentó el el capítulo anterior, estas AUs permiten categorizar el tipo y la intensidad de cada deformación de los músculos faciales utilizando el *Facial Action Code System* (FACS) [Ekman et al., 2002]. Los dos sistemas presentan un esquema similar, parten de la información del sensor, extraen características de la imagen (RGB-D o RGB, respectivamente), y usan un clasificador bayesiano para estimar el estado emocional del usuario de entre los cinco posibles. Ambos sistemas son presentados con detalle, describiendo cada una de las fases y presentando resultados tras probarlo con diferentes usuarios, tanto en edad, género o rasgos faciales. A su vez, los dos métodos han sido comparados con algoritmos similares del estado del arte, destacando sus mejoras en la precisión y robustez de los resultados.

3.2. Sistema de reconocimiento de expresiones faciales basado en Candide-3

3.2.1. Descripción del sistema

En este capítulo se presenta un sistema de reconocimiento de expresiones faciales basado en un modelo de malla 3D. El sistema propuesto utiliza las librerías de desarrollo $Kinect\ for\ Windows\ SDK$ [Microsoft, 2014] y el $Toolkit$ - $Microsoft\ Kinect\ Face\ Tracking$, todo ello a partir del componente $WinKinectComp$ del $framework$ RoboComp. Este componente adquiere y pre-procesa la información referente a las características faciales del usuario, y constituye la entrada para el sistema descrito. El proceso de adquisición está basado en el uso del sensor RGB-D ($Kinect$) que combina la información de color y profundidad para obtener un modelo de malla $Candide$-3 [Ahlberg, 2001] que se ajusta sobre la cara del usuario, recogiendo la mayor parte de las expresiones del mismo. El uso de la malla *Candide-3* permite un seguimiento eficiente de la posición, de la orientación de la cabeza y de las deformaciones de los músculos y de cada uno de los elementos de la cara. Esta información es transmitida a través del $middleware$ de comunicación ICE (*Internet Communication Engine*) [Henning and Spruiell, 2005] a otros componentes encargados de procesar las características faciales para posteriormente estimar el estado emocional del usuario.

El sistema de reconocimiento de emociones que se presenta en esta sección contiene, desde el punto de vista del diseño, tres etapas diferentes que se ilustran en la Figura. 3.1:

- *Adquisición de datos*: este primer proceso es el encargado de la adquisición, preprocesamiento y transferencia de la información del componente $WinKinectComp$. La información transferida está relacionada no sólo con los nodos que componen el modelo de malla $Candide$-3, sino también con la posición, la orientación o la propia imagen RGB del usuario.

- *Extracción de características faciales*: la información transmitida por el componente $WinKinectComp$ mediante ICE permite reconstruir el modelo de malla $Candide$-3 por otro equipo. A partir de los puntos del modelo el sistema obtiene un conjunto de N características del rostro F^I=$\{f_i^I | i = 1...N\}$, basándose en las AUs del sistema FACS.

- *Red bayesiana dinámica*: como entrada de esta última etapa se encuentran las características extraídas en el paso anterior. Finalmente, este último proceso utiliza el conjunto de características en la estimación del estado emocional del usuario, y para ello hace uso de un clasificador bayesiano dinámico, un modelo probabilístico que permite trabajar con secuencias de variables en el tiempo. Los posibles estados emocionales de salida del sistema, comunes en todo el documento, son los estados de felicidad, tristeza, miedo, enfado y el propio estado neutral.

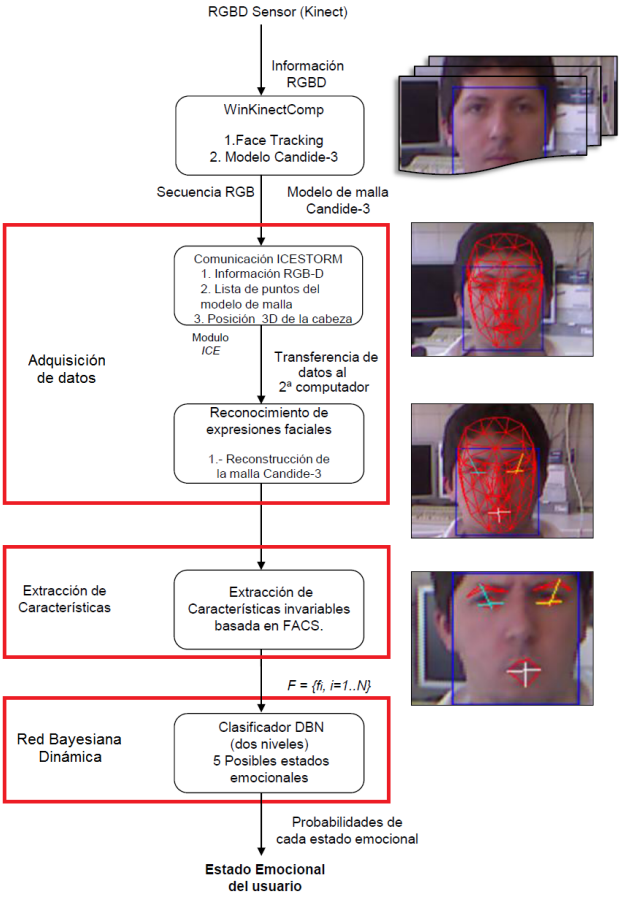


Figura 3.1: Vista general del sistema.

Cada una de estas fases es descrita con detalle en las siguientes subsecciones.

3.2.2. Adquisición de datos y pre-procesamiento

La adquisición de los datos necesarios para el reconocimiento de expresiones faciales está basado en el uso de la información RGB-D proporcionada por el sensor $Kinect$. La imagen de profundidad capturada por este sensor facilita el ajuste del modelo de malla $Candide-3$ sobre la cara del interlocutor. Este conjunto de puntos 3D que conforman la malla, junto con la posición y orientación de la cara, es transmitido desde el componente $WinKinectComp$ (funcionando como un servidor en el Sistema Operativo $Windows$) a otros componentes en

un mismo ordenador o de forma distribuida a otros equipos de la red (actuando como clientes), todo ello a través del $middleware$ ICE. Los procesos de adquisición y transferencia de la información son descritos a continuación.

3.2.2.1. Componente WinKinectComp

El primer paso en la adquisición de datos para el sistema está relacionado con la información necesaria para la extracción de características faciales. El proceso de captura de datos es realizado por este componente, que adquiere la información del sensor $Kinect$ por medio de la librería de desarrollo $Kinect\ for\ Windows\ SDK$, para su posterior procesamiento y transmisión. La principal función de este componente radica en utilizar la imagen RGB (640x480) y de profundidad, proporcionadas por el sensor, en los algoritmos de detección y seguimiento de caras que implementan el modelo de malla $Candide$-3. Este software forma parte del $framework$ RoboComp, descrito en el Apéndice C, fue desarrollado en colaboración con el grupo de Ingeniería de Sistemas Integrados (ISIS) de la Universidad de Málaga..

Los principales procesos del componente son descritos en detalle, según su orden en el programa:

- *Seguimiento de la cara del usuario (face tracking)*: el seguimiento de caras comienza con la detección del usuario en la imagen, mediante una identificación positiva del usuario (o usuarios) y sus características faciales en tiempo real. En este componente, el proceso de seguimiento permite obtener información de diversa índole, como el número de usuarios, la Región de interés (ROI) de la imagen donde se encuentra la cara del usuario detectado, la distancia del sensor al interlocutor o la propia posición y orientación del usuario, entre otras. Sin embargo, el principal aporte de este proceso al sistema descrito está relacionado con la información 3D que permite ajustar un modelo de malla sobre la cara del usuario.

- *Modelo de malla Candide-3*: esta malla es un modelo deformable basado en polígonos, implementado por el $toolkit\ Microsoft\ Kinect\ Face\ Tracking$. Este modelo está compuesto por una lista de 133 puntos o nodos bidimensionales, interconectados para formar una serie de triángulos a partir de los cuales se realiza un $mapping$ o ajuste entre esta lista de puntos y la imagen del usuario. En un inicio, la malla presenta una expresión facial neutra, basada en FACS. Esta expresión se modifica a través de las AUs globales y locales que controlan los cambios en los polígonos del modelo para imitar las deformaciones de los músculos de la cara en cada expresión facial.

  En la Figura 3.2 se muestra el modelo de malla realizando una expresión facial determinada (sorpresa). La información completa sobre la lista de nodos con sus respectivas posiciones en la malla se encuentra en el Apéndice A.2.

Los datos obtenidos de los procesos antes mencionados son transferidos por medio del componente servidor (o publicador) a varios clientes (subscriptores) mediante ICE. La información publicada por el componente es la siguiente:

- Sensor RGBD ($Kinect$): imagen RGB, imagen de profundidad, tamaño de la imagen y número de bits por píxel.

- Lista de puntos 2D del modelo deformable $Candide$-3.

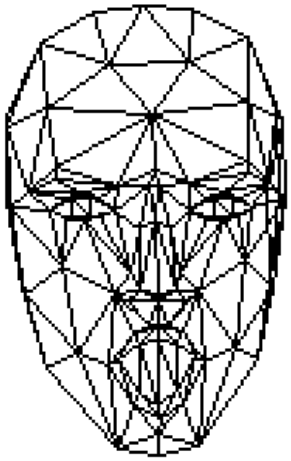


Figura 3.2: Modelo de malla $Candide$-3 para una expresión facial concreta, en este caso, de sorpresa.

- Posición 3D de la cabeza del usuario: Matrices de rotación y traslación, y el tamaño de la cabeza del usuario.

3.2.2.2. Comunicación publicador/suscriptor

El proceso de comunicación basado en el $middleware$ ICE permite, de forma eficiente, la implementación de robots que puedan paralelizar sus sistemas internos en varios equipos. En el caso que nos ocupa, por ejemplo, el uso compartido de los recursos permite realizar múltiples sistemas de reconocimiento que estimen el estado emocional del usuario por diferentes medios. Incluso un mismo medio, como la información visual capturada por el sensor RGB-D, puede ser utilizada por sistemas de reconocimiento que trabajen con la expresión facial (descrito en este mismo capítulo), o bien a través del lenguaje corporal del usuario (descrito en capítulos posteriores). Para ello, en este trabajo, el servicio $IceStorm$ [ZeroC, 2014] es el encargado de transmitir los datos a través de la red por medio de un sistema de distribución de eventos basados en suscripción/publicación. Este servicio permite al suscriptor (cliente) utilizar sólo la información que considere necesaria para su funcionamiento y procesamiento de una lista de elementos disponibles facilitados por el publicador.

En la implementación de este sistema de reconocimiento de expresiones faciales, la transferencia de información es realizada como se muestra en la Figura 3.3. Esta configuración permite transmitir datos a varios programas simultáneos que utilicen un mismo robot para adquirir la información del usuario. El servicio $IceStorm$ permite dividir los procesos en varios programas, reduciendo el procesamiento de la información y disminuyendo el coste computacional de cada equipo.

3.2.3. Extracción de características faciales

El proceso de extracción de características parte de la información proporcionada por el servicio $IceStorm$. Del conjunto de elementos publicados, este módulo extrae las características faciales necesarias para la estimación de la emoción a partir de la lista de puntos o nodos de la malla, y los datos referentes a la posición y orientación del usuario. El modelo

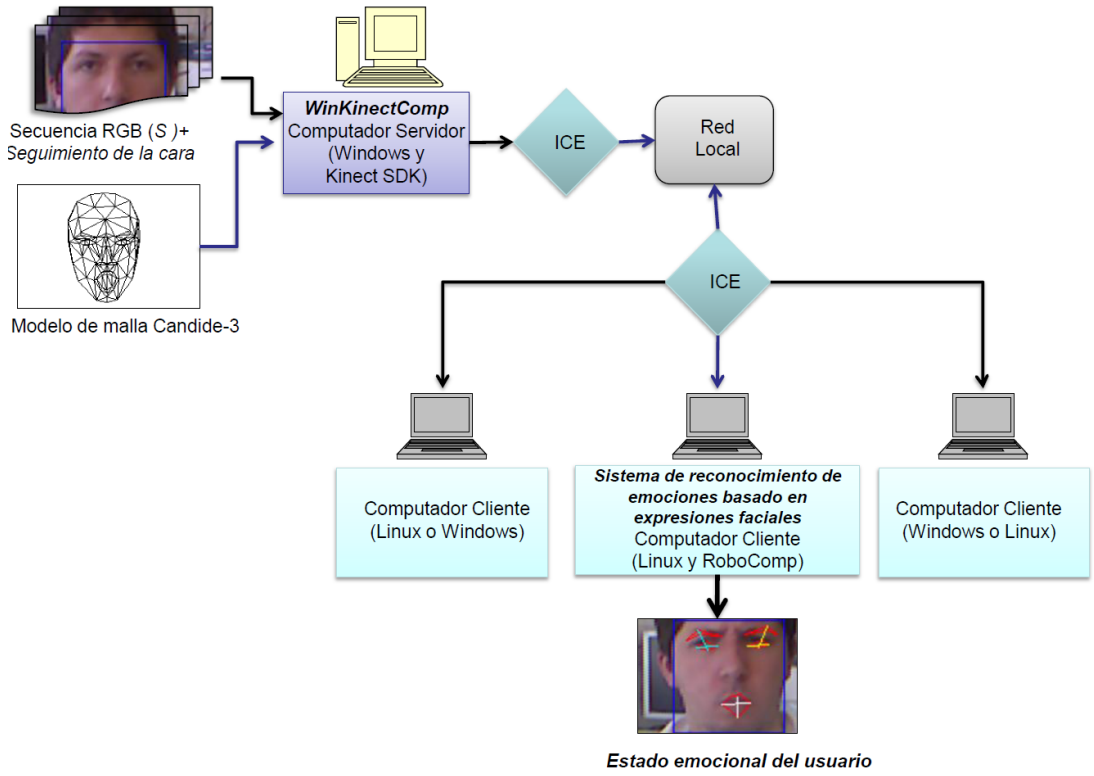


Figura 3.3: Descripción del sistema de comunicación basado en ICE.

de malla $Candide$-3 facilita la detección de las características faciales asociadas a deformaciones de la cara al expresar una emoción. El conjunto de estas N características faciales, F^I=$\{f_i^I | i = 1...N\}$, están directamente relacionadas con las Unidades de acción AUs descritas en el *Facial Action Code System* (FACS) [Ekman et al., 2002]. En la Figura 3.4 se muestran las AUs utilizadas en este trabajo, que han sido seleccionadas del total de AUs por sus propiedades antagónicas y exclusivas (por ejemplo, las $AU1$ y $AU4$ de la figura, asociadas a la distorsión de las cejas, son opuestas entre ellas). Además, para una correcta estimación de cada estado emocional es necesario que las unidades de acción seleccionadas superen un umbral mínimo de intensidad B (es decir, leve evidencia), dentro del rango de intensidad descrito en el Apéndice A.1.

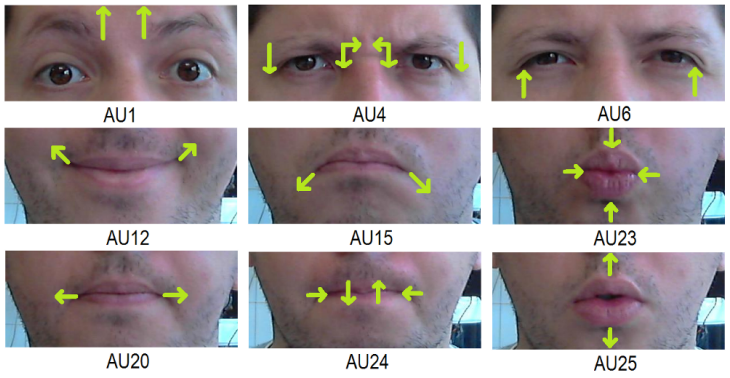
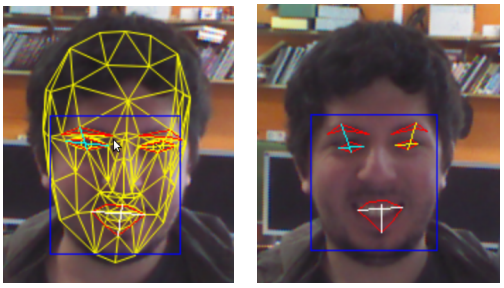


Figura 3.4: Unidades de Acción AUs utilizadas en este sistema.

El proceso para calcular las características faciales está basado en el uso de la distancia Euclídea entre diferentes nodos del modelo de malla, como se muestra en la Figura A.3. Como es de sobra conocida, la distancia Euclídea entre dos puntos o nodos, P = (x_1, y_1, z_1) y Q=(x_2, y_2, z_2), del modelo en un espacio de tres dimensiones, está caracterizada por la ecuación:

$$d(P,Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}, \tag{3.1}$$



(a) (b)

Figura 3.5: Extracción de características faciales basado en el modelo $Candide$-3; a) Modelo de malla sobre la cara del usuario; b) Características extraídas del modelo. (Figura obtenida de la publicación [Cid and Núñez, 2014])

En este sistema, las principales distancias necesarias para estimar el estado emocional del usuario, son las siguientes:

- $\mathbf{d_{eb}}$: esta variable representa la distancia entre el contorno superior de las cejas y el borde inferior de los ojos, en los nodos 17-25 (MiddleTopOfLeftEyebrow – UnderMidBottonLeftEyelid) de la malla descrita en la sección A.2. Esta variable está asociada directamente con la variable EB del clasificador, como se muestra en la Figura 3.6 (d_{eb} - Amarillo).

- $\mathbf{d_{lc}}$: este valor mide la distancia entre las comisuras de la boca, en los nodos 32-65 (OutsideLeftCornerMouth – OutsideRightCornerMouth), estando asociada directamente con la variable LC del clasificador, como se muestra en la Figura 3.6 (d_{lc} - blanco).

- $\mathbf{d_{ma}}$: esta variable cuantifica la distancia entre el contorno superior y el borde inferior de los labios (conocida como la apertura de la boca), en los nodos 8-9 (MiddleTopDipUpperLip – MiddleBottonDipLowerLip). Su valor está directamente relacionado con la variable MA del clasificador descrito en la sección siguiente, como se muestra en la Figura 3.6 (d_{ma} - blanco).

- $\mathbf{d_{mf}}$: este valor representa la distancia entre las comisuras y el borde inferior de los labios (describiendo la forma de la boca), en los nodos 9-32 (MiddleBottonDipLowerLip – OutsideLeftCornerMouth). Además, se asocia directamente con la variable MF del clasificador, como se muestra en la Figura 3.6 (d_{mf} - blanco).

- $\mathbf{d_{ch}}$: esta variable representa la distancia en las mejillas, en los nodos 21-24 (OuterCornerOfLeftEye – InnerCornerLeftEye). La variable CH del clasificador bayesiano se relaciona directamente con este valor, tal y como se muestra en la Figura 3.6 (d_{ch} - blanco).

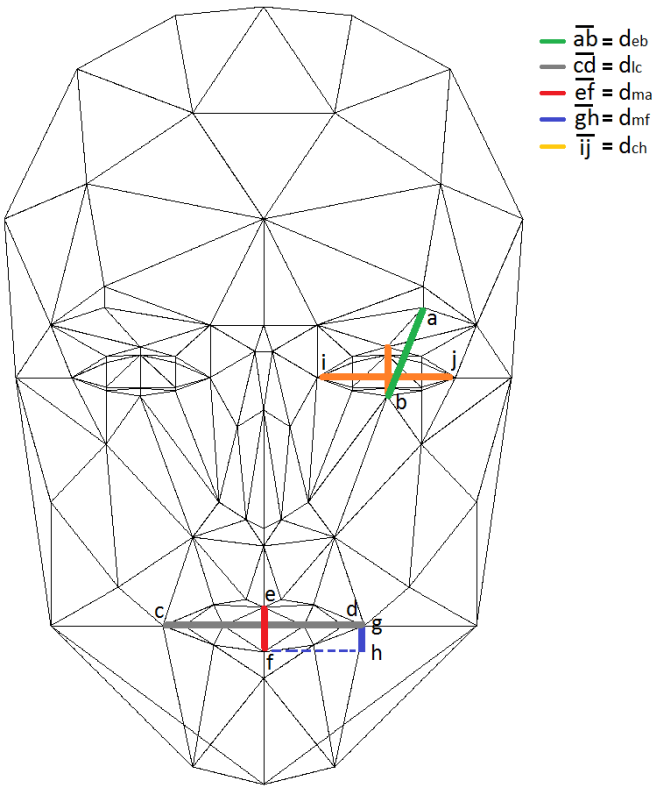


Figura 3.6: Distancias principales para la extracción de características faciales.

Estas distancias obtenidas desde el modelo de malla evitan las limitaciones asociadas a la escala o distancias estáticas al sensor, tradicionales errores introducidos por los sistemas basados en imagen bidimensionales, al normalizar la información por medio de una malla 3D. La principal limitación del sistema proviene de la etapa previa de $tracking$, al depender exclusivamente del modelo de malla facilitado al detector de características. Se ha evaluado esta limitación del algoritmo de seguimiento según la distancia del usuario, llegando a que, en condiciones normales, este rango está definido dentro de una distancia de 0.4 a 2 mt. como se observa en la Figura 3.7. Como será descrito en el apartado correspondiente de resultados experimentales, el sistema basado en cámaras RGB-D y el modelo de malla $Candide-3$ permite mejorar los resultado obtenidos en comparación a sistemas previos que utilizaban la distancia en píxeles [Cid et al., 2013b, Romero et al., 2013].

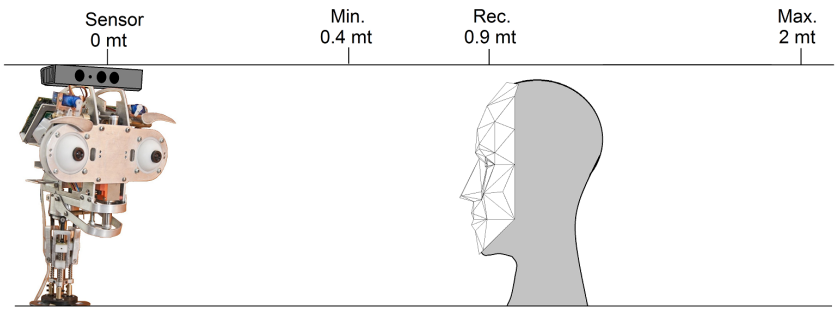


Figura 3.7: Distancias del usuario al sensor.

3.2.4. Red bayesiana dinámica

La última etapa del sistema realiza la función de clasificación, dando a la salida una estimación de la emoción humana de entre los cinco posibles estados usados en esta Tesis Doctoral (neutral, felicidad, tristeza, miedo y enfado). El clasificador se implementa siguiendo una *red bayesiana dinámica* (***DBN*** , del inglés *Dynamic Bayesian Network*), cuya implementación √ es un modelo mejorado del presentado en [Prado, 2012]. La red bayesiana hace uso de las características faciales extraídas en la etapa previa y gracias a las propiedades antagónicas y exclusivas antes mencionadas, se reduce el número de variables utilizadas, de las nueve iniciales (ligadas a las AUs), a tan sólo cinco:

- *EB:* $\{AU1,\ AU4,\ ninguno\}$, esta variable está asociada a los movimientos de las cejas (*Eye-Brows*), y su valor está relacionada con la existencia de la AU1 y AU4.
- *Ch:* $\{AU6,\ ninguno\}$, su valor está asociado a los movimientos de las mejillas (*Cheeks*). Específicamente, indica si se levantan las mejillas, dando lugar a la existencia de la AU6.
- *LC:* $\{AU12,\ AU15,\ ninguno\}$, esta variable está ligada a los movimientos de las esquinas de los labios (*Lip Corners*). En este caso, la probabilidad de identificar los AU12 y AU15 a través de esta variable depende del movimiento de las comisuras de los labios. Por ejemplo: el AU12 presenta mejores probabilidades cuando se detiene el movimiento de los labios. Si las comisuras de los labios se mueven, el AU15 presenta una mayor probabilidad de identificación.
- *MF:* $\{AU20,\ AU23,\ ninguno\}$, esta variable está asociada a la forma de la boca (*Mouth's Form*). Los AU20 y AU23 respectivamente, están relacionados con la acción de estirar o contraer la boca en forma horizontal.
- *MA:* $\{AU24,\ AU25,\ ninguno\}$, este valor mide la apertura de la boca (*Mouth's Aperture*), donde AU24 y AU25, respectivamente, están relacionadas a la acción de presionar los labios o relajarlos y abrir la boca.

La estructura de la red bayesiana consta de dos niveles y una característica de dependencia en el tiempo, como se muestra en la Figura 3.8. El primer nivel de la red contiene la variable principal FE, que representa los posibles estados emocionales del usuario resultantes del proceso de clasificación, ($FE_{[Neutral]}$, $FE_{[Felicidad]}$, $FE_{[Tristeza]}$, $FE_{[Miedo]}$, $FE_{[Enfado]}$). En el segundo nivel, la variable FE se presenta como padre de las cinco variables del segundo nivel que conforman la entrada de la red bayesiana, EB, Ch, LC, MF y MA.

La red descrita necesita ser complementada con información de aprendizaje obtenida de las cinco variables del segundo nivel para cada estado emocional (ver con detalle el Apéndice A.2). Estos datos se consiguen por medio de un entrenamiento inicial, con una pequeña muestra de usuarios que presentan diferentes características faciales, género y edad. Además, para evitar errores, ambigüedades o lagunas (entrenamiento insuficiente) se realiza un pre-procesamiento de la información que consiste en un ajuste gaussiano de los datos adquiridos en el entrenamiento.

Los datos *D* obtenidos por medio del algoritmo de extracción de características faciales propuesto en este capítulo, poseen la siguiente configuración:

$$D = ((x_1, y_1)...(x_5, y_5)), x_i \in \mathbb{R}^d, y_i \in \mathbb{R} \tag{3.2}$$

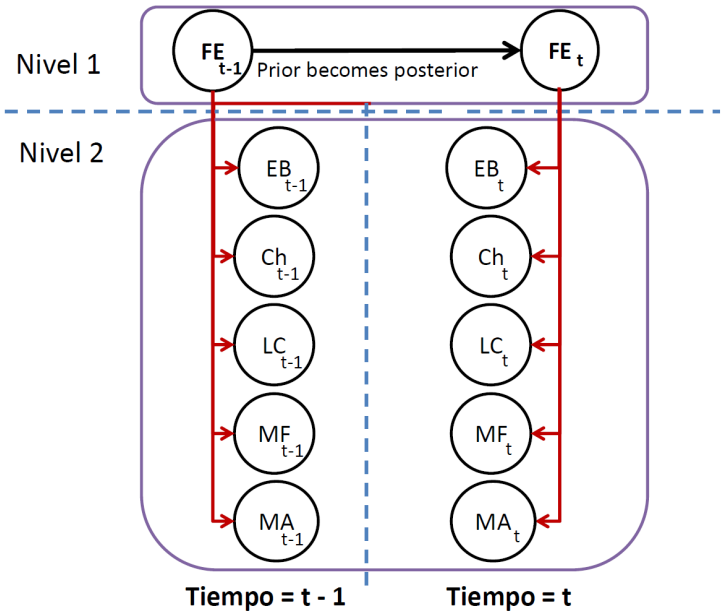


Figura 3.8: Red bayesiana dinámica, donde se muestra un intervalo de 2 tiempos (t-1, t).

Donde y_1 a y_5 son los cinco posibles estados emocional ($FE_{[Neutral]}$, $FE_{[Felicidad]}$, $FE_{[Tristeza]}$, $FE_{[Miedo]}$, $FE_{[Enfado]}$) y x corresponde a uno de las variables aleatorias descritas anteriormente, específicamente (EB, Ch, LE, LC, CB, MF y MA). Dado que los datos de aprendizaje pueden tener lagunas entre sus muestras, se construye un modelo simplificado, asumiendo que $(X_1, ..., X_5)$ son independientes dada la expresión facial FE. Por tanto,

$$X_i \sim N(prior^T x_i, \sigma^2) \tag{3.3}$$

Al principio, el valor de $prior \sim U(1/5)$, sin embargo a lo largo de las iteraciones, el resultado en el instante t se convierte en el valor obtenido en $t-1$. En el calculo de la distribución conjunta asociada a este clasificador bayesiano de expresiones faciales, se utilizaron las variables aleatorias (x) que pertenecen al segundo nivel de la red bayesiana, como se ilustra en la ecuación 3.4.

$$\begin{aligned} & P(FE, EB, CH, LC, MF, MA) \\ &= P(EB, CH, LC, MF, MA \mid FE) \cdot P(FE) \\ &= P(EB \mid FE) \cdot P(CH \mid FE) \cdot P(LC \mid FE) \\ &\cdot P(MF \mid FE) \cdot P(MA \mid FE) \cdot P(FE) \end{aligned} \tag{3.4}$$

Las propiedades antagónicas y exclusivas de las unidades de acción causan que las variables del segundo nivel del clasificador sean independientes entre sí. Por lo tanto, dada la ecuación 3.4 asociada a la distribución conjunta, el $posterior$ se obtiene por medio de la aplicación de la regla de Bayes en la ecuación 3.5.

$$\begin{aligned} & P(FE \mid EB, LE, LC, MF, MA) \\ = & P(EB \mid FE) \cdot P(CH \mid FE) \cdot P(LC \mid FE) \\ & \cdot P(MF \mid FE) \cdot P(MA \mid FE) \\ & \cdot P(FE)/P(EB, CH, LC, MF, MA) \end{aligned} \tag{3.5}$$

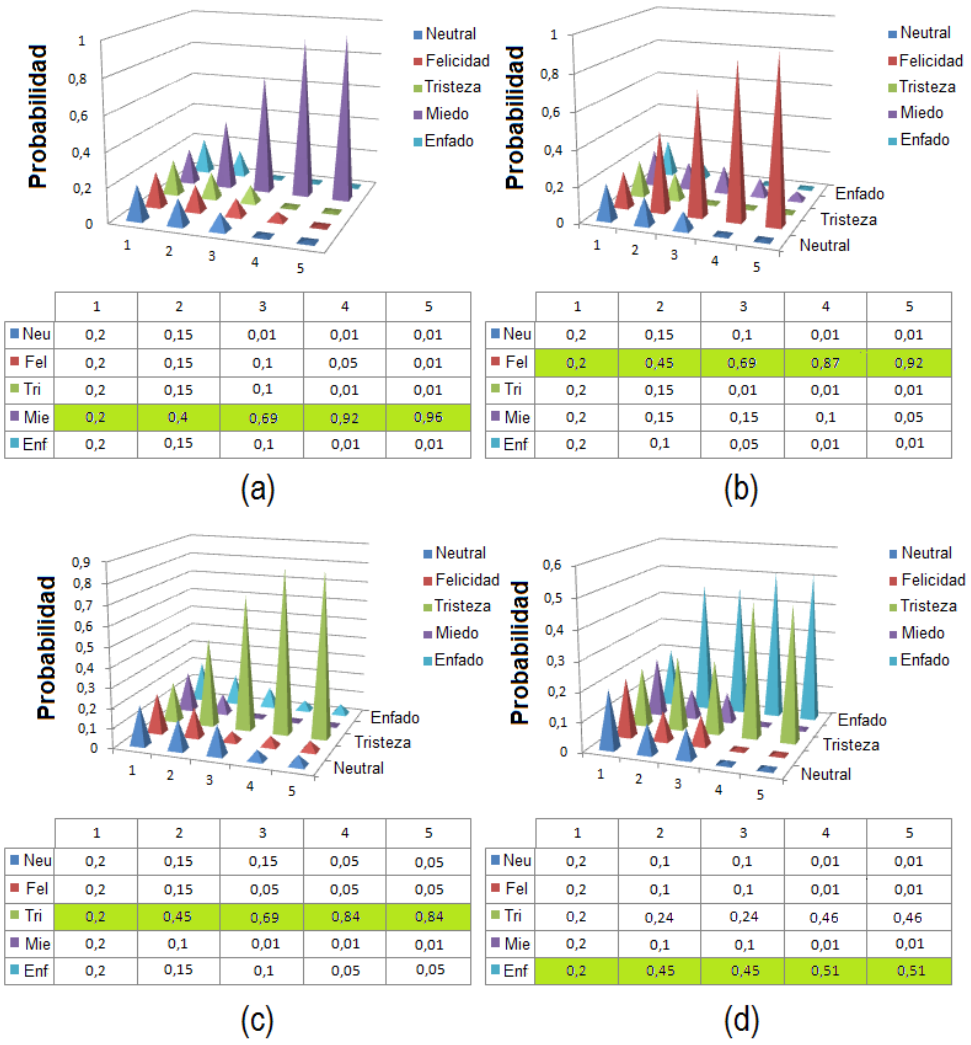
Finalmente, la constante de normalización puede ser calculada por medio de la regla de marginación bayesiana:

$$\begin{aligned} & P(EB, CH, LC, MF, MA) \\ = & \sum_{FE} P(EB \mid FE) \cdot P(CH \mid FE) \cdot P(LC \mid FE) \\ & \cdot P(MF \mid FE) \cdot P(MA \mid FE) \cdot P(FE) \end{aligned} \tag{3.6}$$

En este sistema, el modelo de la red bayesiana posee una propiedad dinámica que causa una convergencia a lo largo del tiempo. El efecto del tiempo está basado en el histograma resultante desde el instante previo, que se utiliza como información a priori para el instante de tiempo actual. Esta convergencia se considera completa, cuando el umbral es superior al 85 % después de 5 instantes consecutivos. Sin embargo, si el umbral no es superado después de este tiempo, el clasificador selecciona la probabilidad más alta (usualmente llamada como $Maximum\ a\ posteriori\ decision$ en teoría bayesiana) como el resultado del clasificación. En la Figura 3.9 se muestra un ejemplo de la convergencia del modelo, donde (a) muestra el estado emocional miedo, (b) el estado felicidad, (c) el estado tristeza y finalmente en (d) se observa un ejemplo de ambigüedad. Este último demuestra que después de 5 tramas el resultado es el estado emocional con mayor probabilidad (en este caso, enfado).

3.2.5. Limitaciones

Las principales limitaciones asociadas al funcionamiento del método propuesto están relacionadas, como suele ser común en este tipo de sistemas, con las propias condiciones de un escenario no controlado. Cambios en la luminosidad de la escena, interferencia con la luz ambiente o el mismo comportamiento de los usuarios, entre otras, suelen provocar fallos en el proceso de clasificación final. En primer lugar, destaca el hecho de que las condiciones de luz con fuentes irregulares o indirectas afectan negativamente el ajuste del modelo de malla $Candide$-3 facilitado por el algoritmo $FaceTracking$. Este problema comienza con un error en la detección de todos los elementos de la cara del usuario (es decir, no encuentra ningún elemento asociado al rostro), creando una implementación aproximada y errónea del modelo que no permite estimar correctamente el estado emocional del usuario. Por otro lado, existe un segundo problema importante relacionado también al algoritmo $FaceTracking$ implementado en el $toolkit$, asociado, en este caso, a problemas en la detección de la boca a distancias superiores a 1.2 metros con fuentes indirectas de luz natural. En estas condiciones, una correcta detección de la boca se hace francamente difícil, generando errores en el reconocimiento de los estados emocionales estimado por el algoritmo presentado en esta Tesis por su fuerte dependencia con las características de este elemento del rostro humano.



	1	2	3	4	5
Neu	0,2	0,15	0,01	0,01	0,01
Fel	0,2	0,15	0,1	0,05	0,01
Tri	0,2	0,15	0,1	0,01	0,01
Mie	0,2	0,4	0,69	0,92	0,96
Enf	0,2	0,15	0,1	0,01	0,01

(a)

	1	2	3	4	5
Neu	0,2	0,15	0,1	0,01	0,01
Fel	0,2	0,45	0,69	0,87	0,92
Tri	0,2	0,15	0,01	0,01	0,01
Mie	0,2	0,15	0,15	0,1	0,05
Enf	0,2	0,1	0,05	0,01	0,01

(b)

	1	2	3	4	5
Neu	0,2	0,15	0,15	0,05	0,05
Fel	0,2	0,15	0,05	0,05	0,05
Tri	0,2	0,45	0,69	0,84	0,84
Mie	0,2	0,1	0,01	0,01	0,01
Enf	0,2	0,15	0,1	0,05	0,05

(c)

	1	2	3	4	5
Neu	0,2	0,1	0,1	0,01	0,01
Fel	0,2	0,1	0,1	0,01	0,01
Tri	0,2	0,24	0,24	0,46	0,46
Mie	0,2	0,1	0,1	0,01	0,01
Enf	0,2	0,45	0,45	0,51	0,51

(d)

Figura 3.9: Resultados desde el clasificador para diferentes estados emocionales (Figura obtenida de la publicación [Cid et al., 2014]).

Estas dos limitaciones están ligadas al componente $WinkinectComp$, siendo un problema recurrente en escenarios reales, pero no relevante dentro de espacios interiores o habitaciones pequeñas para interacciones directas entre el agente robótico y el usuario. El hecho de disponer de estas limitaciones en el sistema presentado, hace necesario tener en consideración los siguientes puntos:

- La distancia del usuario al sensor es afectada negativamente por la cantidad y la dirección de la luz natural que incide durante el funcionamiento del sistema. Por este motivo, la distancia que obtiene los mejores resultados está comprendida entre los 0.5 y 1.5 metros, con bajas condiciones de iluminación.

- El uso de este sistema en condiciones de luz natural irregular necesita el refuerzo de luz artificial que permita una identificación positiva del usuario.

3.2.6. Resultados experimentales

A lo largo de esta sección se realizan una serie de experimentos que persiguen evaluar el rendimiento del sistema propuesto en el capítulo. El componente software presentado, $AffordancesHumanComp$, ha sido desarrollado en C++, dentro del $framework$ RoboComp [Manso et al., 2010]. Las pruebas se realizaron en un ordenador con CPU de 2.8 GHz Intel(R) Core(TM) i7 y 4 Gb de RAM funcionando en $Linux$. En el experimento se contó con un grupo de 40 usuarios con diferente género, edad y rasgos faciales (existencia o no de barba, pelo largo o corto, por ejemplo), que realizan cinco secuencias aleatorias de expresiones faciales por cada usuario, como se describe en el trabajo [Cid et al., 2014]. La Figura 3.10 muestra una parte de los usuarios con la malla $Candide$-3 ajustada al rostro, donde se puede visualizar diferentes expresiones faciales relacionadas con diferentes estados emocionales. En el Cuadro 3.1 se ilustra la matriz de confusión del sistema de reconocimiento propuesto, donde se expresa el porcentaje de cada expresión facial detectada correctamente, P_{fe}, en la diagonal. También se muestran en la tabla los datos referidos a los posibles errores en la clasificación. De los resultados de la tabla anterior se desprende un funcionamiento bastante preciso del sistema, presentando un porcentaje de acierto, en todas las expresiones faciales, mayor al 90 %. Al finalizar este capítulo existe un estudio comparativo con otras técnicas similares,

Test P_{fe}	Tristeza	Felicidad	Miedo	Enojo	Neutral	Errores
Tristeza	**90** %	0 %	0 %	0 %	4 %	6 %
Felicidad	0 %	**98** %	0 %	0 %	0 %	2 %
Miedo	1 %	2 %	**95** %	0 %	0 %	2 %
Enojo	2 %	0 %	0 %	**95** %	0 %	3 %
Neutral	3 %	0 %	0 %	0 %	**92** %	5 %

Cuadro 3.1: Matriz de confusión de los resultados del sistema de reconocimiento de expresiones faciales basado en 5 estados emocionales (Cuadro obtenido de la publicación [Cid et al., 2014]).

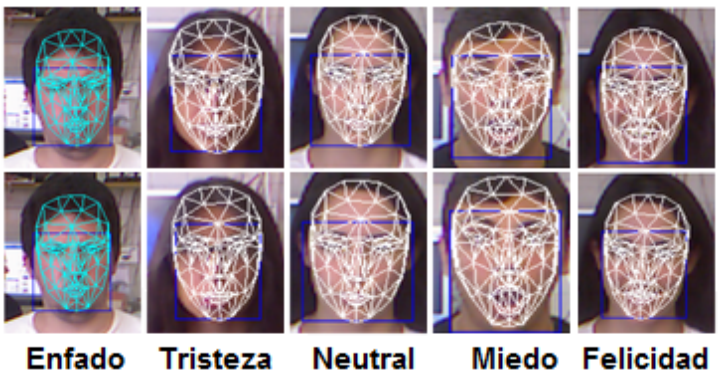


Figura 3.10: Usuarios participantes en las pruebas (Figura obtenida de los datos de la publicación [Cid et al., 2014]

En el Cuadro 3.2 se muestran los errores del sistema descrito en este capítulo. Como puede observarse en la tabla, la mayor parte de estos errores se deben a fallos en la clasificación por

sobre-entrenamiento de la red bayesiana, la ambigüedad de los resultados facilitados por la red o la inexistencia de estados emocionales que superen el umbral requerido.

Errores	Clasificación errónea	Ambigüedad	Bajo el umbral
$P_{Test1(FE)}$	2 %	1 %	1 %

Cuadro 3.2: Detalle de los errores (Cuadro obtenido de la publicación [Cid et al., 2014]).

3.3. Sistema de reconocimiento de expresiones faciales basado en el filtro de Gabor

El sistema de reconocimiento de expresiones faciales basados en filtros de $Gabor$ es descrito con completo detalle en el trabajo [Cid et al., 2013b]. A diferencia del método presentado en la sección 3.2, en este caso se comienza con la adquisición, en tiempo real, de la información visual por medio de una secuencia de vídeo obtenida por las cámaras RGB del robot. Esta secuencia de imágenes es utilizada para detectar la región de interés (ROI_I) de la cara del usuario. A continuación, esta ROI es procesada, de forma que se divide la misma en dos subregiones ROI_{Top} y ROI_{Botton}, esto es, la parte superior e inferior de la cara, respectivamente. A ambas imágenes se les aplica una serie de filtros para contrarrestar los efectos del ruido y de las condiciones de luz ambiente (por ejemplo, las sombras o los reflejos, entre otros). Finalmente, el uso del filtrado por $Gabor$ permite la eliminación de información no relacionada con contornos, dejando una imagen que realza los bordes de la cara del interlocutor. La segunda parte de este sistema está compuesta por un método para la extracción de las características faciales basado en el análisis de estos contornos, y su posterior uso en un clasificador. Esta última etapa se realiza por medio de una red bayesiana dinámica para la estimación del estado emocional del usuario, similar en su estructura a la descrita en la sección anterior. En la Figura 3.11 se ilustran los procesos que componen este sistema, que son descritos en detalle en las siguientes sub-secciones.

3.3.1. Adquisición de datos

Al igual que la mayor parte de los sistemas de reconocimiento de expresiones faciales, la primera etapa es la encargada de procesar la secuencia de vídeo obtenida por el robot, S, y realizar la adquisición de la información visual necesaria para la detección y seguimiento del usuario en cada instante de tiempo. El algoritmo utilizado es descrito en el trabajo de Viola y Jones [Viola and Jones, 2004], que utiliza características $Haar$ y una clasificador en cascada para el reconocimiento del rostro del usuario en tiempo real. La detección de la cara ofrece a su salida una imagen RGB que se procesa para ofrecer un tamaño normalizado, lo que reduce considerablemente la cantidad de información de la imagen original. A continuación la ROI se divide en dos regiones diferenciadas, una específica para la parte superior del rostro, ROI_{top}, centrada en los ojos y cejas, y otra en la parte inferior del mismo, ROI_{botton}, centrado en la boca y en sus movimientos (ver como se muestra en la Figura 3.12a).

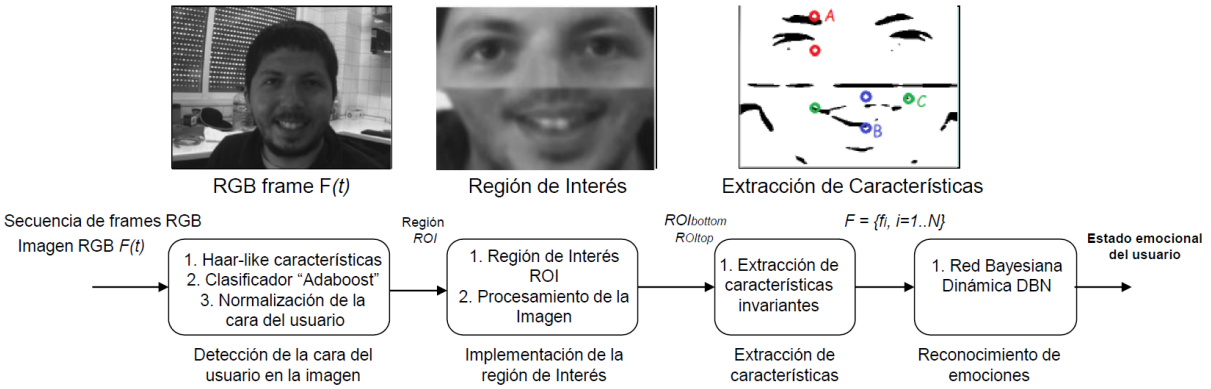


Figura 3.11: Vista general del sistema de reconocimiento de expresiones faciales basado en el filtro de $Gabor$. La figura fluye desde la izquierda a la derecha. Los detalles del funcionamiento del sistema están recogidos en el texto.

3.3.2. Procesado de la región de interés

Esta etapa cumple el rol de procesar la región de interés inicial ROI_I a través de varios algoritmos encaminados a eliminar el ruido, reducir la dependencia con las condiciones de la luz y eliminar la información innecesaria en etapas posteriores. El ruido que tiene su base en elementos característicos de la cara pero que en sí mismo no aportan información emocional (por ejemplo, la barba o las heridas) se eliminan con una serie de filtrados consecutivos a la imagen. En particular, se utiliza un filtro de Mediana, y otro de desenfoque Gaussiano. Finalmente, y con el fin de reducir los efectos de la luz en el proceso de extracción de características, se realiza el siguiente procedimiento basado en el enfoque descrito en [Tan and Triggs, 2010]:

1. Corrección $gamma$: es una transformación no lineal en escala de grises que reemplaza el nivel de gris por las valores de gamma γ definidos por el usuario ($\gamma \in [0, 1]$). La implementación de esta corrección mejora el ajuste del brillo en la imagen, dado que al procesar imágenes RGB pueden ser oscurecidas o blanqueadas, afectando también a las proporciones de los colores rojos, verdes y azules. El propósito de esta corrección es recuperar información de la imagen independiente de la iluminación.

2. Diferencia de Gaussianas (***DoG***): este paso corrige los efectos del sombreado en la imagen. La DoG es un algoritmo utilizado en la detección de bordes, basado en el uso de dos desenfoques gaussianos con diferentes radios de desenfoque, del cual se sustrae una imagen final. Debido a la importancia de los detalles, el radio del desenfoque interior presenta valores más pequeños en comparación al radio de la segunda función gaussiana. Esto es común dada las fuertes variaciones de la luz y el uso de una fuente de iluminación directa. Sin embargo, es importante mencionar que el uso de la DoG causa la reducción del contraste local de las imágenes resultantes en las regiones sombreadas, lo que provoca perdidas de información visual (por ello se utiliza previamente la corrección $gamma$). √

3. Enmascaramiento: es el algoritmo utilizado en este sistema para eliminar información irrelevante de la imagen que afecta a los siguientes procesos en la extracción de carac-

terísticas. Principalmente se eliminan los elementos en los extremos de las imágenes, como el cabello o parte del fondo de la imagen que no corresponde al usuario.

4. Ecualización del Histograma: es la última etapa del procedimiento encargado de mejorar el contraste en una imagen, expandiendo el rango de intensidad de la misma. Para ello, se transforma la distribución original del histograma a una distribución más uniforme y se extiende los valores de intensidad en todo el rango, maximizando el contraste sin perder información [Culjak et al., 2012].

3.3.3. Filtro de *Gabor*

Ambas regiones de interés, ROI_{top} y ROI_{botton}, son utilizadas como información visual de entrada al filtro de $Gabor$. El filtro de $Gabor$ es un filtro lineal utilizado, entre otras ventajas, por sus buenos resultados en la detección de bordes con diferentes orientaciones [Kamarainen, 2012] y con un bajo coste computacional. Los contornos extraídos a partir del filtrado son utilizados para la extracción de las características faciales de los usuarios en las etapas posteriores. Un filtro de $Gabor$ posee una respuesta de impulso en el dominio espacial que consiste en una onda plana sinusoidal de cierta orientación y frecuencia, modulada por una envolvente Gaussiana bi-dimensional. Así, sea $I(u, v)$ la imagen de entrada, la salida del filtro de $Gabor$, $G(u, v)$, viene dada por:

$$G(u,v) = exp(-\frac{1}{2}(\frac{u_\theta^2 + v_\theta^2}{\sigma^2}))cos(2\pi\frac{u_\theta}{\lambda} + \psi) \tag{3.7}$$

Donde θ, λ y ψ están asociados a la onda sinusoidal plana (orientación, longitud de onda y fase), y siendo u_θ y v_θ descritas como:

$$\begin{aligned} u_\theta &= ucos\theta + vsin\theta \\ v_\theta &= -usin\theta + vcos\theta \end{aligned} \tag{3.8}$$

En la Figura 3.12b se muestra la imagen a la salida del filtro de $Gabor$. Como se observa en la figura, los bordes extraídos están asociados a diferentes elementos del rostro humano. A su vez, la eliminación del ruido en la imagen, así como la normalización de la luz ambiental, permite obtener una imagen de bordes clara, sobre la que aplicar los algoritmos de extracción de características.

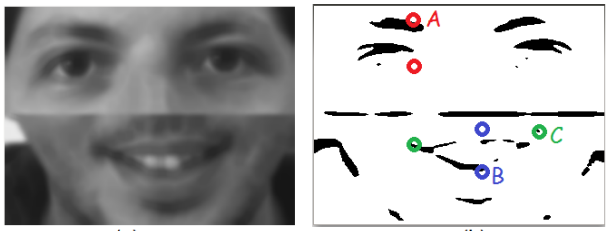


Figura 3.12: Extracción de características faciales basada en contorno; a) Región de interés en la imagen de la cara (ROI_{top} y ROI_{botton}); y b) Extracción de características en la imagen.

3.3.4. Extracción de características faciales

Las características faciales utilizadas en el método presentado están directamente relacionadas con las Unidades de Acción (AUs) descritas en el *Facial Action Code System*, y del conjunto de éstas, únicamente son seleccionadas aquellas que presentan propiedades antagónicas y exclusivas, tal y como se describe en la sección 3.2.3. En la Figura 3.13 se muestran las 11 AUs utilizadas en este sistema, con sus propiedades opuestas (por ejemplo, la $AU24$ y $AU25$). Tomando como base estas AUs, se pueden obtener una serie de características del rostro humano a partir de los contornos extraídos, las cuales están asociadas a las distancias Euclídeas entre el contorno superior de las cejas y el borde inferior de los ojos dA, las comisuras de los labios dB y el contorno superior e inferior los labios de la boca dC.

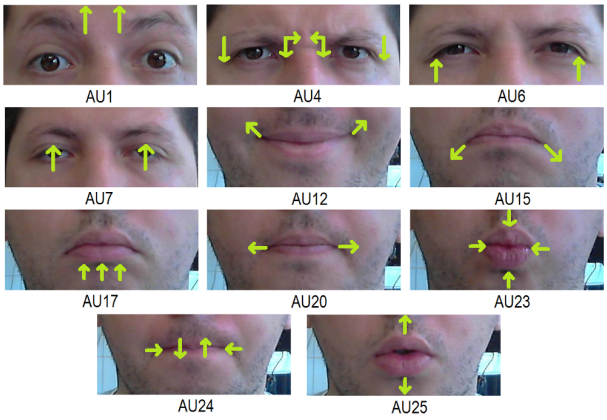


Figura 3.13: Unidades de Acción AUs utilizadas en el sistema presentado en esta sección.

El cálculo de las características faciales por medio de las distancias Euclídea entre dos píxeles diferentes, $P = (x_1, y_1)$ y $Q=(x_2, y_2)$, en un espacio de dos dimensiones, se realiza por medio de la conocida Ecuación 3.9:

$$d(P,Q) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, \tag{3.9}$$

El cálculo de estas distancias sigue un proceso similar al método basado en la malla $Candide - 3$. El proceso parte inicialmente de la imagen resultado del filtro de $Gabor$, se detectan los puntos característicos asociados a los elementos del rostro (parte superior de las cejas, borde inferior del ojo, extremos de los labios, y parte superior e inferior de la boca) por simple análisis de la imagen de borde. Tras ello, se calculan las distancias siguiendo la Ecuación 3.9, y finalmente se normalizan estas distancias según los valores que alcanzan en la expresión facial asociada al estado neutral para ese usuario. Esto último permite reducir la dependencia a la escala del usuario en la imagen y la distancia del usuario al sensor. En la Figura 3.12 se observa la Región de interés y las características extraídas, dA (rojo), dB (azul) y dC (verde).

3.3.5. Red bayesiana dinámica

La red bayesiana utilizada en este sistema posee una estructura similar a la descrita en la sección 3.2.4. La principal diferencia se refiere al uso de 11 Unidades de Acción, con propiedades antagónicas y exclusivas, que son agrupadas en siete grupos de variables para reducir el coste computacional. Este clasificador posee una estructura de dos niveles y una propiedad dinámica

que causa un convergencia en el tiempo, como se ilustra en la Figura 3.14. En el primer nivel se encuentra el nodo padre FE que corresponde al posible estado emocional ($FE_{[Neutral]}$, $FE_{[Felicidad]}$, $FE_{[Tristeza]}$, $FE_{[Miedo]}$ y $FE_{[Enfado]}$) del usuario. El segundo nivel contiene los siete grupos de variables asociados a las unidades de acción AU y son obtenidas a partir de las distancias Euclídeas de diferentes puntos del rostro.

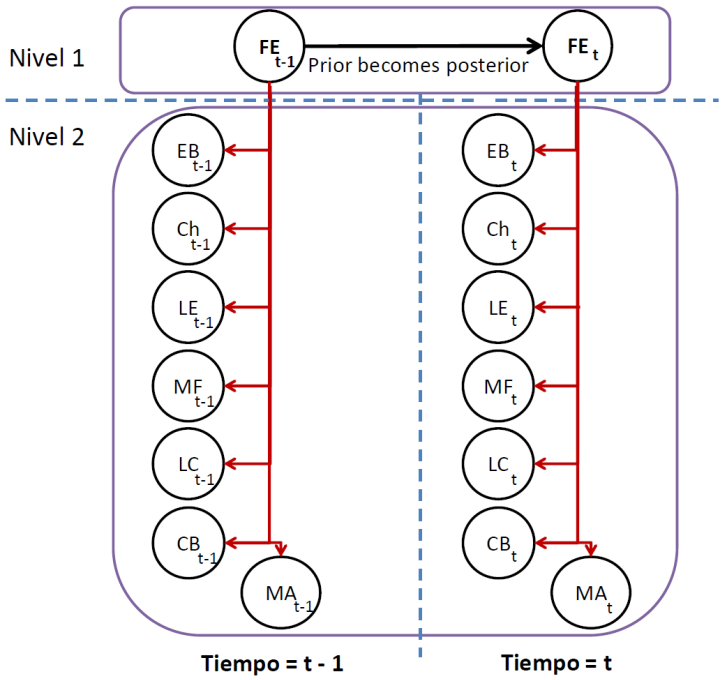


Figura 3.14: Red bayesiana dinámica de dos niveles, donde se muestra un intervalo de 2 tiempos (t-1, t).

Por un lado, las primeras cinco variables (EB, Ch, LC, MF, MA) corresponden a las mismas variables utilizadas en la red bayesiana de la sección 3.2.4. En este caso las variables se extraen del análisis 2D tras el filtro de $Gabor$.

- *EB:* $\{AU1, AU4, ninguno\}$; Esta variable está asociada a los movimientos de las cejas (*Eye-Brows*), y su valor está relacionada a la existencia de la AU1 y AU4.
- *Ch:* $\{AU6, ninguno\}$; Su valor está ligado a los movimientos de las mejillas (*Cheeks*); específicamente, indica si se levantan las mejillas, dando lugar a la existencia de la AU6.
- *LC:* $\{AU12, AU15, ninguno\}$; Esta variable se asocia a los movimientos de las esquinas (comisuras) de los labios (*Lip Corners*). En este caso, la probabilidad de identificar los AU12 y AU15 a través de esta variable depende del movimiento de las comisuras de los labios. Por ejemplo: el AU12 presenta mejores probabilidades cuando se detiene el movimiento de los labios. Si las comisuras de los labios se mueven, el AU15 presenta una mayor probabilidad de identificación.
- *MF:* $\{AU20, AU23, ninguno\}$; Esta variable está asociada a la forma de la boca (*Mouth's Form*). Los AU20 y AU23, respectivamente, están relacionados a la acción de estirar o contraer la boca en forma horizontal.

- *MA:* $\{AU24,\ AU25,\ ninguno\}$; Su valor se relaciona con la apertura de la boca (***Mouth's Aperture***), donde AU24 y AU25, respectivamente, están asociados a la acción de presionar los labios o relajarlos y abrir la boca.

Por su parte, las variables LE y CB son características únicas de este clasificador, y cuyo valor se obtiene como sigue:

- LE: $\{AU7,\ ninguno\}$; Esta variable está asociada al movimiento de los párpados inferiores (***Lower Eyelids***), donde la existencia de la AU7 está relacionada a la acción de elevar los párpados inferiores.

- CB: $\{AU,\ ninguno\}$; Esta variable representa el movimiento de la barbilla (***Chin Boss***). La existencia de la AU17 está relacionada al movimiento de la barbilla hacia arriba.

Tanto el modelo de la red bayesiana como las propiedades que la componen, son idénticas a las presentadas por el clasificador de la Sección 3.2.4. La principal diferencia se refiere al número de variables, que afecta el resultado del sistema y aumenta los datos necesarios en el entrenamiento inicial del mismo. El cálculo de la distribución conjunta asociada a este clasificador bayesiano está basado en las siete variables del segundo nivel de la red, como se ilustra en la ecuación 3.10.

$$\begin{aligned} & P(FE, EB, CH, LC, MF, MA, LE, CB) \\ & = P(EB, CH, LC, MF, MA, LE, CB \mid FE) \cdot P(FE) \\ & = P(EB \mid FE) \cdot P(CH \mid FE) \cdot P(LC \mid FE) \\ & \cdot P(MF \mid FE) \cdot P(MA \mid FE) \cdot P(LE \mid FE) \cdot P(CB \mid FE) \cdot P(FE) \end{aligned} \tag{3.10}$$

3.3.6. Limitaciones

Al igual que se hizo en la sección 3.2.5, en esta sección se presentan las principales limitaciones del sistema propuesto. En este caso, al ser toda información adquirida desde una cámara RGB, los problemas asociados a los cambios en la iluminación son los más determinantes. A pesar del banco de filtros implementado, es difícil contrarrestar el comportamiento de la luz en la escena. En estos sistemas donde sólo se trabaja con información 2D es común encontrar errores en la detección y en el seguimiento del rostro, imprescindible para el correcto funcionamiento del sistema. Durante los experimentos se encontraron limitaciones en la detección asociadas a las orientaciones de la cabeza del usuario. En particular, en el caso del $Roll$ y en menor medida el Yaw, los resultados obtenidos difieren de los esperados. Por su parte, la necesidad de detectar con suficiente resolución el rostro del interlocutor para poder obtener una imagen de borde sobre la que extraer características faciales, hace que la distancia máxima entre el robot y el humano en la interacción no sea superior a 1.2 metros, con luz directa.

3.3.7. Resultados experimentales

Para la evaluación del sistema de reconocimiento se han seguido una serie de experimentos con idea de cuantificar el rendimiento del método propuesto con usuarios no entrenados y en un entorno no controlado. A su vez, se ha probado el mismo con la base de datos $SAVEE$

[Haq and Jackson, 2010]. Las condiciones que se deben tomar en consideración para evaluar estos experimentos se basan en dos enfoques. Por un lado, en términos de software, se utilizó el componente $MuecasEmotionComp$ desarrollado en C++ dentro del $framework$ RoboComp [Manso et al., 2010]. Este componente es el encargado de realizar la estimación del estado emocional del usuario a través de la información visual adquirida del componente $cameraComp$, descrito en el Apéndice C. Por otro lado, las condiciones de hardware están asociadas al robot Muecas, el cual es el encargado de la adquisición de la información visual desde el usuario por medio de dos cámaras RGB - $Point\ Grey\ Dragonfly2\ IEEE$-1394 (DR2-13S2C/M-C) integradas al globo ocular del robot [Cid et al., 2014]. El equipamiento encargado de procesar la información de la prueba de rendimiento está compuesto por un ordenador con una CPU de 2.8 GHz Intel(R) Core(TM) i7 y 4 Gb de RAM funcionando en $Linux$.

El primer experimento tiene como objetivo analizar la respuesta del sistema en condiciones reales, durante una interacción hombre-robot. Se ha trabajado en este punto con un grupo de 22 usuarios que presentan diferentes características, como la edad, el género y la presencia de diferentes características en el rostro (barba, heridas o el tamaño de los pómulos, entre otros). Cada uno de los usuarios realizó diez secuencias aleatorias de expresiones faciales que incluyesen las emociones trabajadas en esta Tesis Doctoral. Se contaba con la presencia de un observador experto, quien al final de cada test evaluaba el correcto desarrollo de la prueba, o por contra, fallos durante la misma. Los resultados de esta primera prueba se muestran en el Cuadro 3.3, en donde se observa que, en conjunto, las emociones con alta intensidad junto con el estado neutral presentan los mejores resultados. La matriz de confusión presentada demuestra que el sistema tiene una alta tasa de acierto, incluso para las emociones de baja intensidad.

Test $P_{Facial1}$	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**90** %	0 %	0 %	0 %	3 %	7 %
Felicidad	0 %	**97** %	0 %	0 %	0 %	3 %
Miedo	1 %	2 %	**93** %	0 %	0 %	4 %
Enfado	2 %	0 %	0 %	**94** %	0 %	4 %
Neutral	3 %	0 %	0 %	0 %	**95** %	2 %

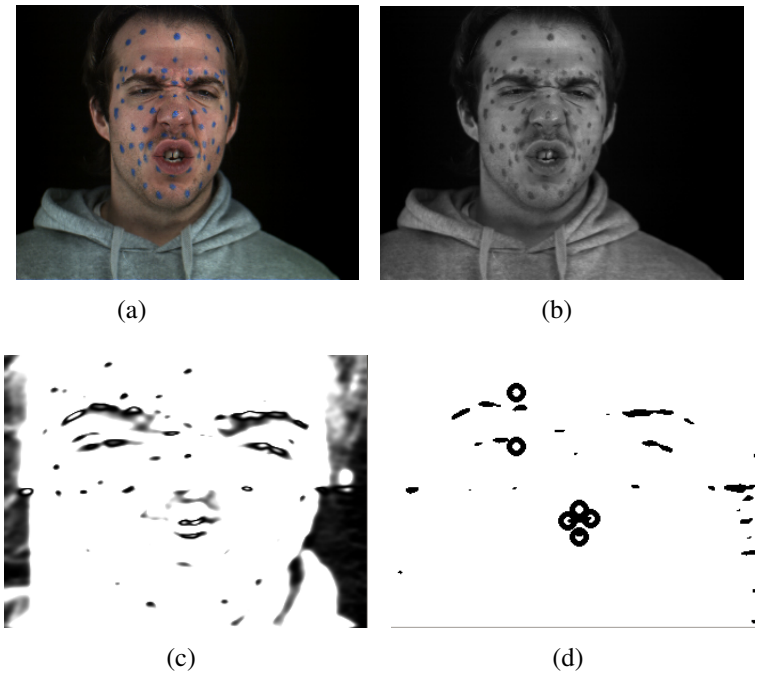
Cuadro 3.3: Resultados del sistema de reconocimiento de emociones basado en el filtrado de $Gabor$ para usuarios no entrenados y entornos parcialmente controlados. (Cuadro obtenido de la publicación [Cid et al., 2013b]).

El segundo experimento consiste en evaluar el sistema de reconocimiento a través de la base de datos audio-visual $SAVEE$ [Haq and Jackson, 2010], que se describe en detalle en el Apéndice A.6. Las pruebas comienzan con el uso de la información visual de la base de datos como entrada del sistema, un conjunto de imágenes (*.jpeg) y archivos de vídeo (*.avi) asociados a los usuarios. Antes de comenzar las evaluaciones se modificaron manualmente las componentes de frecuencia del filtro de $Gabor$ para eliminar las marcas o puntos en la cara, de forma que que no tuvieran influencia en los resultados. En la figura 3.15a se ilustra la imagen del usuario con las marcas en el rostro, mientras en la Figura 3.15d se observa que las marcas han sido eliminadas. Además, para evitar errores relacionados con el entorno donde se toman los datos para cada usuario, se adquirió la información de aprendizaje de la red bayesiana desde la propia base de datos. El procedimiento de este segundo experimento es similar al realizado en

el primer experimento. En primer lugar se eligieron secuencias de vídeo aleatorias con diferentes expresiones faciales y se estimaron los estados emocionales en cada instante de tiempo. Los resultados de esta evaluación se ilustran en el Cuadro 3.4, donde las emociones con valencia negativa presentaron los peores resultados, aunque todos ellos con valores cercanos al noventa por ciento. Como puede observarse tras este segundo experimento, los resultados del sistema en este caso, con un equipo de captura de datos profesional y usuarios entrenados en entorno controlados, confirman la robustez y precisión del método propuesto. En la figura 3.15 se ilustran los procesos del reconocimiento de emociones por medio de la base de datos $SAVEE$.

Test $P_{Facial2}$	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**96** %	0 %	0 %	0 %	1 %	3 %
Felicidad	0 %	**98** %	0 %	0 %	0 %	2 %
Miedo	1 %	3 %	**89** %	0 %	0 %	7 %
Enfado	2 %	0 %	0 %	**93** %	0 %	5 %
Neutral	1 %	0 %	0 %	0 %	**97** %	2 %

Cuadro 3.4: Matriz de confusión del sistema. Evaluación del método propuesto para el reconocimiento de emociones usando la base de datos $SAVEE$.



(a) (b) (c) (d)

Figura 3.15: Figura obtenida de la base de datos audio-visual $SAVEE$; a) Imagen original del usuario KL que genera el estado emocional enfado; b) Imagen en escala de grises del usuario KL; c) Imagen que contiene las sub-regiones ROI_{top} y ROI_{botton},; d) Imagen procesada por $Gabor$ del usuario KL con el estado emocional enfado;

Se analizaron los errores en la estimación a lo largo de los experimentos presentados. Estos valores están descritos en el Cuadro 3.5 para ambos experimentos, $facial1$ y $facial2$, respec-

tivamente. La mayor parte de los errores vienen determinados por fallos en el clasificador, posiblemente por el hecho de que muchos usuarios realizan expresiones poco naturales, actuadas o exageradas en los experimentos, que no poseen un correlación con los datos de aprendizajes utilizados en el entrenamiento de la red. En lo que respecta a los errores asociados a la ambigüedad y los resultados bajo el umbral, se deben principalmente a problemas en la detección del usuario que provocan errores en la convergencia del clasificador y por consecuencia en la estimación del estado emocional del usuario.

Errores	Clasificación errónea	Ambigüedad	Bajo el umbral
$P_{Test(facial1)}$	2 %	1 %	1 %
$P_{Test(facial2)}$	2 %	1 %	2 %

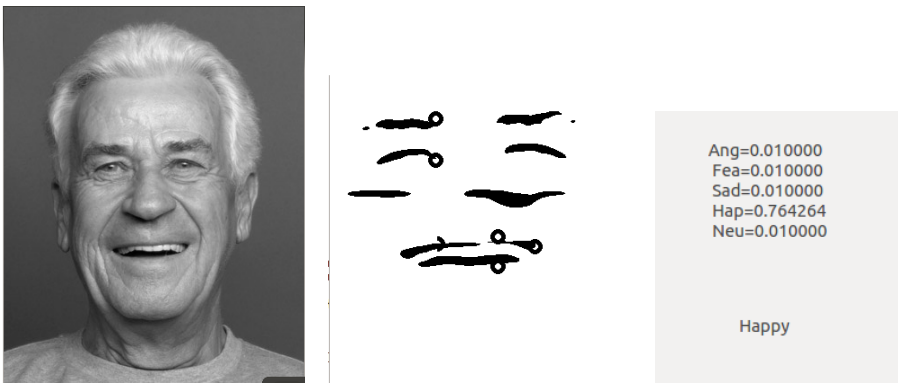
Cuadro 3.5: Detalle de los errores (La información parcial de el cuadro fue obtenida de la publicación [Cid et al., 2013b]).

Finalmente, el sistema de reconocimiento presentado en esta sección fue también utilizado en la base de datos $FACES$, con el fin de probar su funcionamiento en entornos controlados, sin cambios en la luminosidad durante la interacción. Esta base de datos presenta imágenes de alta calidad, sin problemas de iluminación o interferencias, y con sujetos entrenados. En la Figura 3.16 se muestran algunas imágenes de los usuarios de la base de datos $FACES$, con la calidad y condiciones de imagen que caracterizan a esta fuente de información.

(a) (b) (c) (d)

Figura 3.16: Imágenes de los usuarios obtenidas de la base de datos visual $FACES$; a) Estado emocional miedo; b) Estado emocional miedo; c) Estado emocional enfado; d) Estado emocional felicidad;

En la Figura 3.17a se observa a un usuario de la base de datos realizando una expresión facial, asociada en este caso al estado emocional Felicidad. La Figura 3.17b muestra la imagen de salida del filtro de $Gabor$. Debido al bajo número de imágenes y los múltiples estados emocionales, se entrenó la red bayesiana a través del 25 % de las imágenes totales. En el Cuadro 3.6 se observan los resultados del reconocimiento de emociones basado en filtrado de $Gabor$, unos valores bastante inferiores a los presentados en las tablas anteriores. El mayor problema de esta base de datos está relacionado a un error en la estimación (ambigüedad o resultados bajo el umbral) por las exageradas, poco uniformes y actuadas expresiones faciales.



(a) (b) (c)

Figura 3.17: Imágenes obtenidas de la base de datos visual $FACES$; a) Imagen en escala de grises de un usuario que genera el estado emocional felicidad; b) Imagen procesada por $Gabor$ de un usuario con el estado emocional felicidad; y c) Imagen de la interfaz gráfica con el resultado del sistema de reconocimiento.

Test P_{FACES}	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**67** %	0 %	8 %	8 %	8 %	9 %
Felicidad	0 %	**82** %	18 %	0 %	0 %	0 %
Miedo	0 %	0 %	**75** %	17 %	8 %	0 %
Enfado	0 %	0 %	17 %	**83** %	0 %	0 %
Neutral	25 %	8 %	17 %	8 %	**42** %	0 %

Cuadro 3.6: Resultados de la evaluación del sistema de reconocimiento basado en filtrado de $Gabor$ usando la base de datos $FACES$.

3.4. Estudio comparativo

El desarrollo de sistemas de reconocimiento basados en expresiones faciales se presenta como una solución común en la IHR, lo que ha dado lugar a múltiples trabajos que poseen características y métodos similares para el reconocimiento del estado emocional del usuario. En esta sección se realiza un estudio comparativo de los algoritmos descritos en esta Tesis Doctoral con los resultados obtenidos en trabajos relevantes en la materia [Riaz et al., 2009] y [Mayer et al., 2009]. En la evaluación se ha utilizado la base de datos de expresiones faciales $Cohn\text{-}Kanade$. El trabajo descrito en [Riaz et al., 2009] hace uso también del modelo de malla $Candide-3$ junto con una red bayesiana, siendo capaz de reconocer seis estados emocionales. Por su parte, el método presentado en [Mayer et al., 2009], si bien también utiliza el mismo modelo de malla, hace uso de árboles de decisión en el proceso de clasificación para estimar seis estados emocionales. En el Cuadro 3.7 se muestran los resultados de estos sistemas, lo que permite compararlos con los sistemas presentados en este capítulo [Cid et al., 2014] y [Cid et al., 2013b].

Los resultados del Cuadro 3.7 muestran una ligera mejora por parte de los métodos propuestos, destacando el sistema basado en la malla $Candide$-3, en comparación con estos otros trabajos. Esta diferencia en la precisión está relacionada principalmente con el estudio detallado

Sistema de reconocimiento	Precisión	Emociones	Método	Clasificador
Sist. basado en $Candide$-3 [Cid et al., 2014]	94 %	5	Candide-3	Red Bayesiana
Sist. basado en $Gabor$ [Cid et al., 2013b]	93 %	5	Gabor	Red Bayesiana
Riaz et al. [Riaz et al., 2009]	90 %	6	Candide-3	Red Bayesiana
Mayer et al. [Mayer et al., 2009]	87 %	6	Candide-3	Model Tree

Cuadro 3.7: Estudio comparativo entre diferentes sistemas de reconocimiento de emociones. (Cuadro obtenido de la publicación [Cid et al., 2014])

de las AUs seleccionadas por el método dentro del FACS, así como del uso de información 3D y la limitación de usar cinco estados emocionales como salida del sistema.

3.5. Conclusiones

Las expresiones faciales constituyen una de las fuentes de información más importante a la hora de reconocer el estado emocional de una de las personas durante la comunicación. A lo largo de los últimos años, la comunidad científica ha tratado el tema del reconocimiento facial siguiendo diferentes técnicas y algoritmos, no sólo para la robótica y las interacciones IHR, sino también para todo lo que se refiere a interacciones humano-máquina.

En este capítulo se contribuye con dos sistemas de reconocimiento de expresiones faciales en tiempo real con la capacidad de reconocer el estado emocional por medio de las deformaciones de los músculos faciales del usuario. El primer sistema utiliza la información RGB-D desde un sensor $Kinect$ para implementar un modelo de malla 3D que permite seguir los elementos y deformaciones de la cara. El segundo sistema utiliza las cámaras RGB del robot para procesar la información visual mediante filtros morfológicos y de convolución, y finalmente un filtrado de $Gabor$ para extraer las características faciales del usuario, asociadas a los bordes de los elementos de la cara (cejas, ojos y boca, principalmente). Ambos sistemas presentan etapas y resultados similares, mejorando a otros métodos existentes en la literatura.

Los dos sistemas aquí presentados forman parte del sistema multimodal que será presentado en el Capítulo 6. La salida del sistema de reconocimiento facial conforma la entrada del sistema completo de estimación de la emoción humana, tan necesaria para una comunicación real entre un robot y una máquina, siguiendo el paradigma del lenguaje natural seguido en esta Tesis Doctoral.

Capítulo 4

Sistema de reconocimiento de emociones basado en el análisis del habla

4.1. Introducción

Para llevar a cabo interacciones entre un humano y un robot de una forma natural, el reconocimiento de emociones a partir de medios no invasivos se ha convertido en objeto de estudio por parte de numerosas investigaciones en los últimos años. Los actuales sistemas de reconocimiento de emociones presentan soluciones que basan su funcionamiento en el uso de una u otra fuente de información, tanto visual como auditiva, asociada con el lenguaje corporal, la voz y las propias expresiones faciales del hablante.

En los sistemas de reconocimiento de emociones más comunes, el uso de la voz y las expresiones faciales representan las soluciones más completas y robustas. Entre otros elementos propios de la dificultad del procesamiento de cada una de estas señales, la diferencia básica entre el procesamiento basado en voz y las expresiones faciales está relacionada con la cantidad de usuarios que se puede analizar en tiempo real, dado los largos tiempos de adquisición de datos necesarios en la captura de la información de voz y la propia dificultad de distinguir el número de usuarios. Un algoritmo basado en expresiones faciales, por ejemplo, necesita como máximo un número de $n\ frames$ y su evolución en el tiempo para estimar el estado emocional, normalmente menos de uno o dos segundos. Por el contrario, los sistemas basados en voz necesitan una sentencia o frase completa de varios segundos de duración para determinar el estado emocional, sin contar con otros factores como los fonemas, el idioma o el propio acento del comunicante. En el caso de múltiples usuarios, la diferencia se vuelve aún más notable debido a los diferentes tiempos de espera en la adquisición de datos y la cantidad de usuarios que serán analizados dentro de estos dos tipos sistemas (es decir, el visual y aquellos basados en voz).

El presente capítulo describe el sistema de reconocimiento de emociones humanas basado en voz propuesto en esta Tesis Doctoral, que hace uso de información de bajo nivel, y cuyas características, descritas con detalle a lo largo del texto, son el $Pitch$, la $Energía$ y $Tempo$. La elección de estas variables genera unos resultados con tasas de acierto similares e incluso mejores que estudios similares y con bajo coste computacional, como se verá reflejado en los resultados experimentales al final del capítulo. La solución propuesta presenta un método no invasivo en la detección de emociones que, junto con el reconocimiento de expresiones faciales descrito en el capítulo anterior, permite una comunicación similar a aquella realizada entre dos usuarios humanos.

4.2. Sistema de reconocimiento de emociones basado en el análisis del habla

4.2.1. Descripción del sistema

Como se ha comentado anteriormente, el sistema de reconocimiento de emociones presentado en este capítulo hace uso de características propias de la voz para estimar el estado emocional del usuario siguiendo un enfoque bayesiano. El escenario IHR para el sistema propuesto presenta al agente robótico interactuando de forma natural con un usuario humano no entrenado en un entorno controlado. Este agente tiene como finalidad analizar la mayor cantidad de información auditiva posible del usuario, a través de una interacción real, y entonces extraer el estado emocional del mismo. Por esta razón, el sistema de reconocimiento comienza con un método de adquisición de datos que permite la captura de la información verbal del $stream$ de audio adquirido por los sensores acústicos. Para obtener la información asociada a la voz humana, se utiliza una librería de procesamiento de audio, SoX, que implementa una serie de funciones para detectar y realzar la información relacionada con la voz, consiguiendo así una señal de audio que contiene la mayor cantidad de información verbal con una duración similar al tiempo correspondiente a la frase hablada por el usuario.

Tras este proceso de adquisición de datos, la señal es analizada para extraer diferentes variables asociadas a la prosodia de la voz humana que directamente son afectadas por los diferentes estados emocionales del usuario. Es en esta etapa donde se hace uso de las características de bajo nivel $Pitch$ (o frecuencia fundamental), $Energía$ y $Tempo$ (o velocidad de la voz humana, también conocida como *speech-rate*). Estas características son extraídas a través de diferentes métodos basados en el uso del dominio de la frecuencia y la cuantificación de la energía de una señal de audio parcial o completa, siendo utilizadas finalmente como entrada del clasificador bayesiano dinámico. En esta última etapa se estima el estado emocional asociado a las variables extraídas de la voz, dando lugar a los cinco posibles estados emocionales usados en esta Tesis Doctoral: felicidad, tristeza, miedo, enfado y el estado neutral. La metodología propuesta en este sistema no es nueva, y está presente en muchos trabajos en la literatura, como el descrito en [Cowie and Cornelius, 2003], donde se estudia la influencia de la intensidad y valencia de los estados emocionales en las características extraídas de la señal acústica.

La Figura 4.1 ilustra una visión general del sistema descrito en este trabajo. Las etapas en las que consta el sistema se resumen a continuación:

- *Detección de voz*: este proceso analiza el $stream$ del audio en una comunicación real entre el robot y el usuario a partir de los sensores acústicos del agente robótico. El $stream$ es procesado en tiempo real a través de una función para la detección de voz humana (***VAD***).

- *Extracción de características de la voz humana*: la información obtenida desde la etapa de adquisición permite seleccionar una serie de características necesarias para la estimación del estado emocional del usuario. En este caso, las variables acústicas están asociadas con características de bajo nivel, las cuales presentan una relación directa con los estados emocionales del usuario. En este trabajo se elige un número limitado de características de audio, de forma que se obtenga resultados en tiempo real y se limite la ambigüedad de los mismos. Finalmente, las variables obtenidas por medio de este proceso, F^A, serán utilizadas como entrada para el sistema de clasificación.

- *Reconocimiento de emociones*: este último proceso estima el estado emocional del usuario entre cinco posibles opciones establecidas para este sistema. Para ello utiliza una red bayesiana dinámica como clasificador, similar a la presentada en el capítulo 3.

Finalmente, el sistema propuesto fue implementado dentro de RoboComp, dando lugar a un componente software que se encarga de transferir la información obtenida a otros componentes del $framework$ (todo ello será descrito en detalle en el Apéndice C). A continuación se describen cada una de las fases anteriormente expuestas.

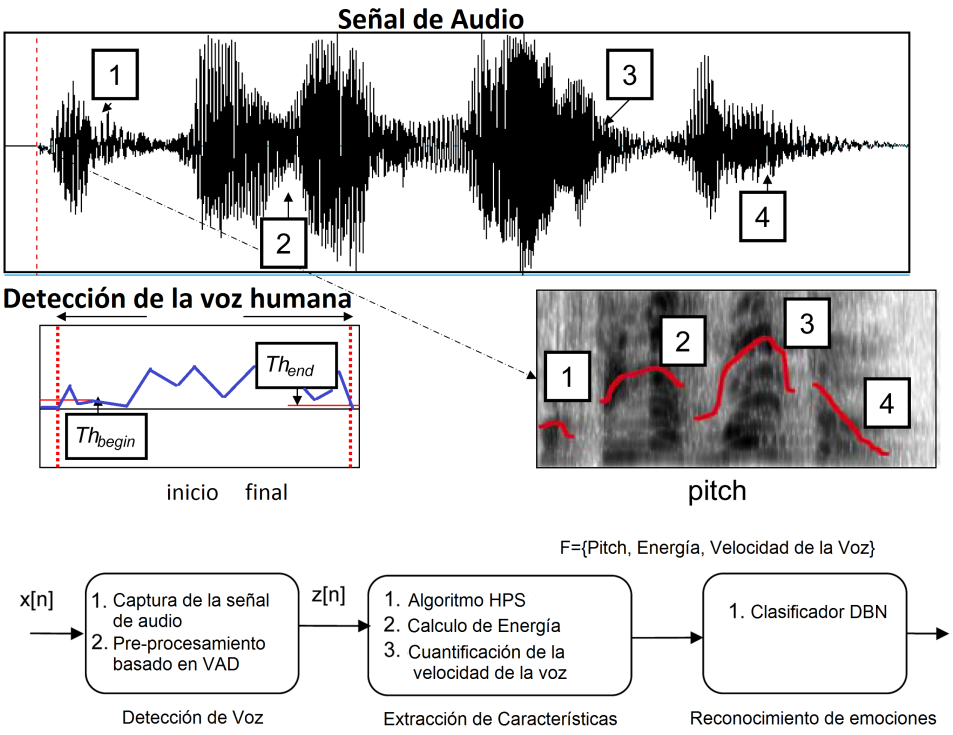


Figura 4.1: Vista general del sistema de reconocimiento de emociones basado en voz humana.

4.2.2. Detección de la voz humana

En primer lugar, la señal de audio adquirida por el sistema durante una IHR real (con frecuencia de muestreo, f_s, de 16 KHz) es procesada con el objetivo de detectar, en tiempo real, aquellas partes del $stream$ de audio donde el interlocutor está hablando con el robot. Para lograr esto, se ha utilizado la librería de audio SoX [C. Bagwell, 2014] que, entre otras funciones, permite la detección de silencios durante la comunicación, así como la detección de voz humana. La librería SoX se describe con detalle en el Apéndice A.3, y dentro de este sistema, permite ajustar los elementos para esta detección de voz y silencios de acuerdo con la propia sensibilidad del micrófono.

En particular, el algoritmo toma el $stream$ original y lo procesa con la función VAD ($Voice\ Activity\ Detection$) de la librería. Esta función basa su funcionamiento en la medición del $Cepstral$ de potencia (o $Cepstrum$ de potencia) [Childers et al., 1977], que permite la eliminación de los ruidos, los silencios o cualquier otro tipo de información que no esté relacionada

con la voz humana. Dada una señal en el tiempo, $x(t)$, correspondiente a una señal de audio adquirida por el robot, su $Cepstrum$ de potencia, $C(\tau)$ viene dada por la ecuación:

$$C(\tau) = \mathcal{F}^{-1}\left(\log(|\mathcal{F}(x(t))|^2)\right) \tag{4.1}$$

, donde $\mathcal{F}$ y $\mathcal{F}^-1$ representan la transformada de $Fourier$ directa e inversa, respectivamente.

Finalmente, la señal de audio compuesta por las tramas pre-procesadas por la función anterior, aquellas que contienen la información de la voz del usuario, se somete a un proceso de remuestreo ($resampling$) para convertir la frecuencia de muestreo f_s a 44.100 KHz, siendo este último, el $stream$ utilizado como información de entrada dentro de proceso de extracción de características acústicas.

En la Figura 4.2a se ilustra una señal de audio, de dos segundos y medio de duración, capturada por el sistema. Esta señal es procesada mediante la función VAD, realzando la información asociada a la voz humana, y limitando su duración a menos de un segundo (Figura 4.2c).

4.2.3. Extracción de características

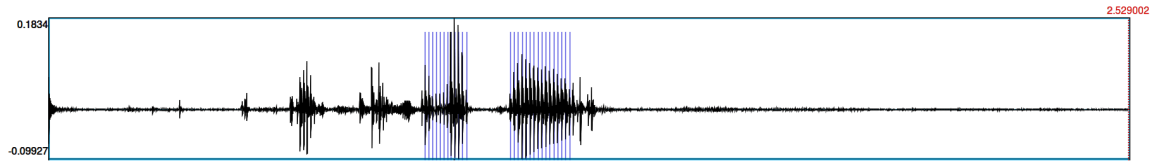
Tal y como se ha comentado al principio del capítulo, el proceso de extracción descrito en esta Tesis Doctoral se basa en el uso de características acústicas de bajo nivel, esto es, en el análisis de los elementos de la prosodia de la voz humana. El método propuesto sigue los pasos de otros algoritmos existentes en la literatura, donde el estudio detallado de la prosodia permite la extracción de elementos característicos de la voz, como queda reflejado en los trabajos [Nogueiras et al., 2001], [Prado et al., 2011], [Chen et al., 2012] y [Schuller et al., 2004] .

Así, dado el $stream$ de audio obtenido en la fase anterior, se extrae un conjunto de m características de la prosodia de la voz F^A=$\{f_i^A \mid i = 1..m\}$. Para la detección de la emoción durante una comunicación verbal, y según el trabajo descrito en [Cowie and Cornelius, 2003], el sistema propuesto toma tres elementos característicos de la voz humana (m = 3), el $Pitch$, la $Energía$ y el $Tempo$. En [Cowie and Cornelius, 2003] se estudiaba la relación entre las diferentes características del habla y los estados emocionales de un interlocutor, llegando a la conclusión de que muchos de los elementos de la prosodia son afectados por la intensidad y valencia de cada una de las emociones. Por citar un ejemplo, las emociones con alta intensidad presentan elevados valores de Energía, $Pitch$ y $Tempo$, mientras que las emociones con baja intensidad presentan unos valores inferiores en estas características.

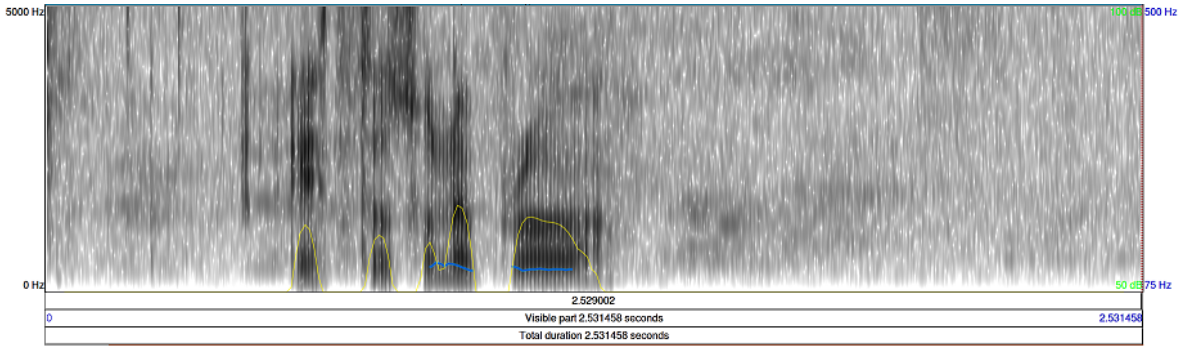
A continuación, se describen las características acústicas extraídas por el sistema:

√ ▪ ***Pitch*** : también llamada Frecuencia Fundamental, es la frecuencia de vibración de las cuerdas vocales que producen los sonidos. En los sistemas de reconocimiento, el rango del $Pitch$ es una característica clave que permite identificar no sólo el género de los usuarios, sino también el estado emocional por medio de la voz.

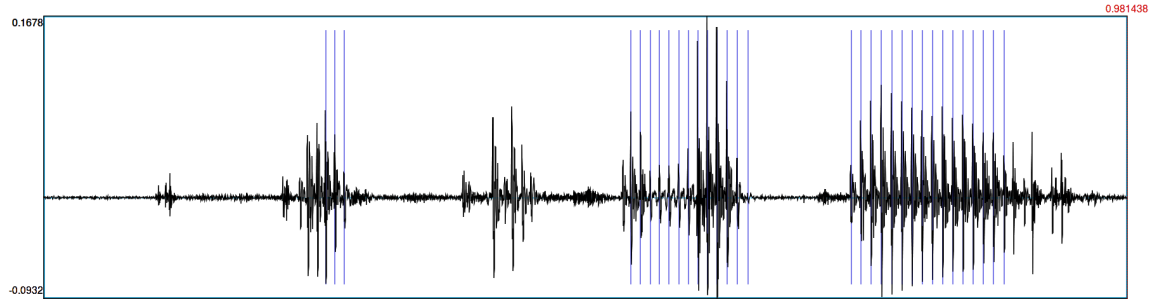
√ ▪ **Energía (*Energy*)**: es considerada como la distribución de los valores de amplitud de la señal en el tiempo. Dentro de la teoría de la señal, la energía es un factor determinante en el reconocimiento y generación de emociones por medio de la voz, dado que las emociones con una alta intensidad están asociadas a elevados valores de energía en la voz, mientras que las emociones con una baja intensidad muestran valores inferiores de energía.



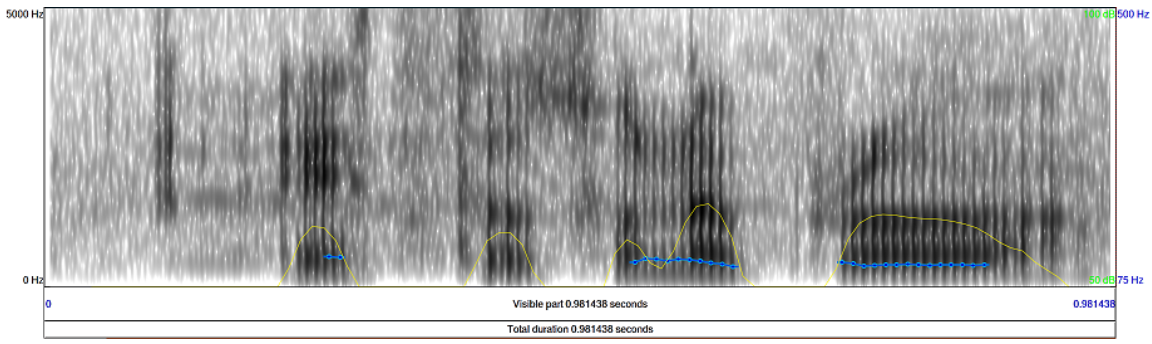
(a)



(b)



(c)



(d)

Figura 4.2: Señal de audio visualizada por el *toolkit* *Praat* [P. Boersma and D. Weenink, 2014], de la señal capturada por medio de la librería *SoX*; a) Señal original capturada del *stream*; b) *Pitch* (azul) y Energía (amarillo) de la señal original. c) Señal pre-procesada por medio de la función VAD; y d) *Pitch* (azul) y Energía (amarillo) de la señal procesada.

- **Velocidad de la voz (*Tempo*)**: Esta velocidad está asociada a la dicción o al número de palabras en un determinado período de tiempo. Dentro de los sistemas de reconocimiento, √

de forma similar a la Energía, la velocidad de la voz es una característica asociada directamente a la intensidad de las emociones humanas. Así, las emociones con alta intensidad presentan una elevada velocidad de la voz, y las emociones con una baja intensidad están asociados a una baja velocidad en la voz [Nogueiras et al., 2001].

Estas tres características son utilizadas como las variables de entradas de la red bayesiana que funciona como clasificador. En el Cuadro 4.1 se muestra la relación entre estas características acústicas y los estados emocionales. De la tabla se observa cómo las emociones de baja intensidad, por ejemplo, la tristeza y el estado neutral, presentan características similares. De igual forma, las emociones con una elevada intensidad como el enfado o el miedo, muestran también características comunes, pero con ligeras diferencias perceptibles para el usuario.

Emoción	*Pitch*	Energía	*Tempo*
Tristeza	Medianamente estrecho	Baja	Visiblemente lento
Felicidad	Muy amplio	Alta	Rápido o Lento
Miedo	Muy amplio	Normal	Muy rápido
Enfado	Muy amplio	Alta	Visiblemente rápido
Neutral	Muy estrecho	Normal	Lento

Cuadro 4.1: Relación entre las características acústicas y los diferentes estados emocionales de este sistema (Información recopilada principalmente desde [Sebe et al., 2005], [Haq and Jackson, 2010] y [Prado et al., 2011]).

Finalmente, en las siguientes sub-secciones se describen en detalle los métodos de extracción de cada una de estas características acústicas.

4.2.3.1. Pitch

El proceso seguido en este trabajo para calcular el rango de $Pitch$ está basado en el algo- √ ritmo ***HPS*** ($Harmonic\ Product\ Spectrum$) [Noll, 1970]. El proceso completo consta de las siguientes etapas:

1. Pre-procesamiento del $stream$

   √ En primer lugar, la cadena de audio es procesado con la función de ***Hann*** [Harris, 1978] (ver Ecuación 4.2) para crear ventanas de corta duración sobre la señal de entrada que permita dividirla en tramas. Para el sistema presentado, el instante aproximado de duración de cada trama T es calculado por medio de la Ecuación 4.3, usando un número de muestras N de 1024 (tamaño de cada trama) y una frecuencia de muestreo f_s de 44.100 KHz.

$$w(n) = 0{,}5 \cdot \left(1 - cos\left(\frac{2 \cdot \pi \cdot n}{N-1}\right)\right); \qquad 0 \leq n \leq N-1; \tag{4.2}$$

$$T = \frac{N}{F_S} = \frac{1024}{44100} = 0,023 \quad segundos \tag{4.3}$$

A continuación, se convierte la señal desde el dominio del tiempo al frecuencial (espectro en frecuencia), a través de la Transformada Rápida de $Fourier$ ***FFT*** ($Fast\ Fourier\ Transform$). En la implementación seguida en este trabajo, se utiliza el algoritmo $Cooley–Tukey$ [Cooley and Tukey, 1965], mediante la Transformada Discreta de $Fourier$, ***DFT*** ($Discrete\ Fourier\ Transform$) [Rockmore, 2000]. Sea $x(j)$ la señal en el dominio del tiempo, discreta y de longitud finita, su representación espectral $X(k)$ queda:

$$Si \qquad W_N = e^{-i\cdot\frac{2\cdot\pi}{N}} \qquad \rightarrow \qquad X(k) = \sum_{j=0}^{N-1} x(j)\cdot W_N^{k\cdot j}, \qquad k \in [0, N-1] \quad (4.4)$$

2. Algoritmo HPS

   El algoritmo HPS ($Harmonic\ Product\ Spectrum$) es un método muy común para la detección del $Pitch$ de una señal de audio [Schuller et al., 2004]. Este algoritmo procesa el espectro de una señal que consiste en una serie de picos con componentes armónicas múltiplos enteros de la frecuencia fundamental. Este método esté compuesto de dos etapas, tal y como se refleja en la Figura 4.3. En primer lugar se submuestrea (***downsampling***) la señal en frecuencia con diferentes factores enteros (f_i), de forma que el espectro frecuencial queda comprimido. Cada una de las señales obtenidas tras este primer proceso se compara con el pico de la frecuencia fundamental del espectro original, donde se puede ver que los picos armónicos se alinean (el primer pico en el espectro original coincide con el segundo pico en el espectro comprimido por un factor de dos, que coincide con el tercer pico en el espectro comprimido por un factor de tres, etc, como se ilustra en la Figura 4.3). A continuación, la última etapa realiza una multiplicación entre todos los espectros comprimidos y el espectro original, siendo el resultado de esta multiplicación, el pico máximo ($peak$) en la frecuencia fundamental de esa trama.

   Por último, a partir de la posición del valor máximo, se calcula la frecuencia fundamental para cada trama por medio de la Ecuación 4.5.

$$F = \frac{n \cdot F_S}{N} \quad (4.5)$$

   Para calcular el rango del $Pitch$ del audio completo se evalúan los valores máximos y mínimos de todas las tramas de la señal de audio. Esta diferencia entre el valor máximo y el mínimo medio es la característica acústica utilizada como variable de entrada en el sistema de clasificación.

4.2.3.2. Energía

La **Energía** de una señal de audio es una variable muy importante dentro de los sistemas de reconocimiento de emociones, debido en gran parte a su fuerte relación con el nivel de intensidad o excitación de las emociones de los usuarios [Ververidis and Kotropoulos, 2006]. El método propuesto para calcular esta característica comienza con un pre-procesamiento de la señal de entrada por medio de la función de $Hann$ [Harris, 1978], de forma idéntica a como fue descrita en la sub-sección 4.2.3.1 (Ecuaciones 4.2 y 4.3).

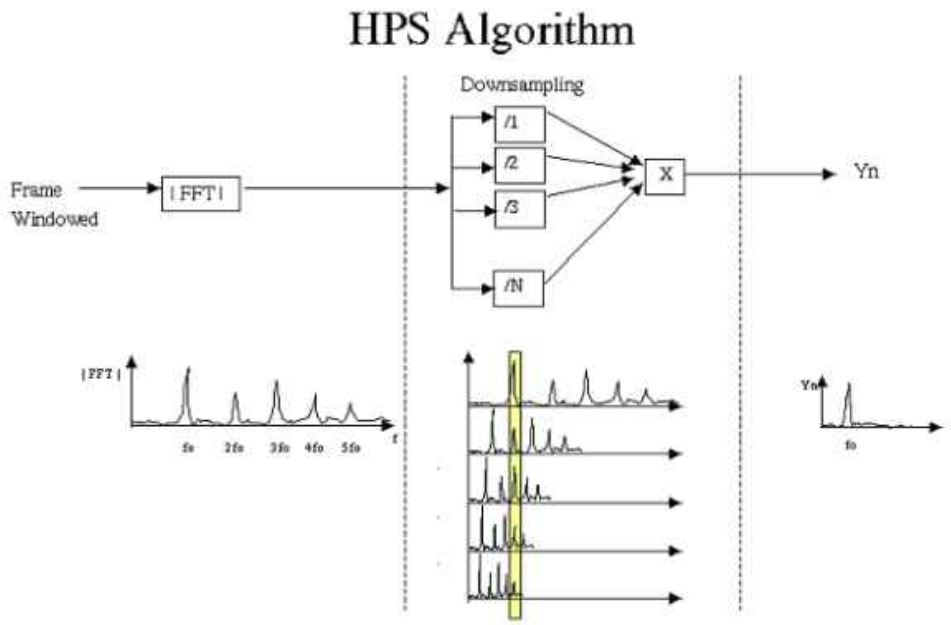


Figura 4.3: Algoritmo HPS, Figura obtenida desde [Cuadra et al., 2001].

A continuación, como la señal fue separada en tramas, se realiza el cálculo de la energía en cada una de ellas, según la Ecuación 4.6.

$$E = \frac{1}{N} \cdot \sum_{x=0}^{x=i} x[i]^2 \tag{4.6}$$

Siendo cada una de las energías de estas tramas la energía instantánea para un número especifico de muestras. No obstante, esta característica está asociada a la energía media de la señal, calculada a través de la división de la sumatoria de estas energías (en cada trama) por el número de tramas que conforman la señal, por medio de la Ecuación 4.7.

$$Energía\ \ Media = \frac{1}{t} \cdot \sum_{x=0}^{x=t} E_x, \qquad t\ =\ n^{o}\ de\ tramas\ de\ la\ señal \tag{4.7}$$

Finalmente, la energía media al procesar todas las tramas es la característica utilizada como variable de entrada dentro del proceso de clasificación.

4.2.3.3. Tempo

La característica acústica del $Tempo$ o la velocidad de la voz es normalmente evaluada por medio de la detección de los $beats$ que corresponden a palabras dentro de la señal. Sin embargo, el método utilizado en este sistema, propone estimar la velocidad de la voz por medio de una serie de multiplicaciones entre la señal de entrada y múltiples trenes de impulso con diferentes $Tempo$. El propósito de esta multiplicación es calcular los valores de las energías en cada caso, y analizar los resultados en búsqueda de la multiplicación que genere el máximo

valor de energía. La multiplicación que genere este valor máximo es aquella que contiene un tren de impulso con el $Tempo$ más cercano al de la señal de entrada.

A continuación, se describe en detalle las etapas que componen el procedimiento antes mencionado:

1. Primero, se convierte la señal al dominio frecuencial, realizando una transformada rápida de $Fourier$ mediante el uso de la DFT sobre la señal.

2. En segundo lugar, en el espectro de la frecuencia, la señal de entrada es multiplicada por múltiples trenes de impulso con diferentes velocidades de voz o $Tempo$, todos ellos con valores cercanos a la velocidad de la voz de un humano. La implementación de los diferentes trenes de impulso está basada en la ecuación:

$$Distancia\ \ entre\ \ impulsos = \frac{60}{BPM} \cdot F_s \tag{4.8}$$

3. Tercero, se analiza la cantidad de energía sobre la multiplicación entre la señal y cada tren de impulsos. Este cálculo utiliza la misma Ecuación 4.7.

4. Finalmente, la multiplicación que presente un valor máximo de energía tendrá un tren de impulso con el $Tempo$ más cercano en relación a la señal original de entrada. Esto se debe a que los trenes poseen un $Tempo$ en un rango similar al de la voz, lo cual permite una estimación de esta característica. El valor de este $Tempo$ es la característica utilizada como variable de entrada del clasificador.

4.2.4. Red bayesiana dinámica

En el sistema propuesto, como se ha comentado anteriormente, el uso de un limitado número de características acústicas reduce el coste computacional y la cantidad de variables utilizadas como entrada dentro del proceso de clasificación. Así, mediante un enfoque bayesiano similar al presentado en el Capítulo 3, es posible determinar el estado emocional del usuario dentro las siguientes posibles opciones: felicidad, miedo, tristeza, enfado y el estado neutral. El clasificador, implementado según una red dinámica bayesina, presenta un estructura de dos niveles que depende del tiempo, como se muestra en la Figura 4.4.

En este caso, el primer nivel contiene un simple nodo, AE ($AudioEmotion$), que representa a la variable asociada con los posibles estados emocionales resultantes del clasificador ($AE_{[Neutral]}$, $AE_{[Felicidad]}$, $AE_{[Enfado]}$, $AE_{[Tristeza]}$, $AE_{[Miedo]}$). Mientras, el segundo nivel de la red corresponde a tres nodos relacionados con cada una de las características de la prosodia de la voz extraídas en la sección anterior, y que son independientes entre sí. Aquí, PT, EN y TE representan los valores del $Pitch$, Energía y $Tempo$, respectivamente.

La red bayesiana propuesta para el reconocedor de emociones basada en voz presenta una estructura en común con el clasificador descrito en el Capítulo 3. Así, se comparten tanto las propiedades del modelo del clasificador, como el proceso de entrenamiento inicial y el propio umbral necesario para la convergencia de la red. La principal y única diferencia está asociada con el número de variables en el sistema, que modifica la cantidad de información necesaria durante el entrenamiento inicial del sistema y la fiabilidad de los resultados.

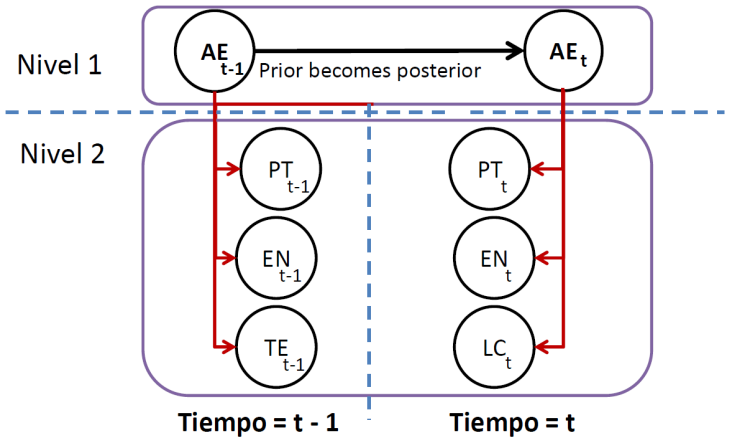


Figura 4.4: Red bayesiana dinámica, donde se muestran dos intervalos de tiempo (t-1, t).

La distribución de probabilidad conjunta asociada a este clasificador se calcula por medio de las variables PT, EN y TE de la señal de entrada. Considerando estas tres variables como independientes, se tiene:

$$\begin{aligned} P(AE, PT, EN, TE) &= P(PT, EN, TE \mid AE) \cdot P(AE) \\ &= P(PT \mid AE) \cdot P(EN \mid AE) \cdot P(TE \mid AE) \cdot P(AE) \end{aligned} \tag{4.9}$$

4.3. Resultados experimentales

Esta sección presenta la evaluación del sistema de reconocimiento descrito en este capítulo. Para ello se describen dos experimentos donde usuarios no entrenados expresan emociones mediante el uso de su voz. En la primera de las pruebas, diferentes personas interactúan con un robot real con unas condiciones del entorno limitada (por ejemplo, con ruido ambiente y baja calidad del audio). El segundo de los experimentos usa la base de datos $SAVEE$, presentada en [Haq and Jackson, 2010], y formada por un número reducido de usuarios en un entorno controlado (esto es, sin ruido y con audio de alta calidad). El algoritmo de reconocimiento ha sido desarrollado en C++, dentro del $framework$ RoboComp [Manso et al., 2010], como un componente más dentro de la arquitectura ($speechrecognitionComp$). El equipo donde se han realizado los experimentos es un ordenador con una CPU de 2.8 GHz Intel(R) Core(TM) i7 y 4 Gb de RAM funcionando en Linux.

El primer experimento está basado en un proceso completo de captura del audio proveniente de una interacción IHR y su posterior análisis y clasificación, según ha sido explicado en este Capítulo. El sistema de audio no profesional está compuesto de un micrófono Logitech y una tarjeta de audio para la captura ASUS Xonar DX PCI-Express. Para la prueba, se contó con un total de 22 participantes de diferentes edades, género y acento, todos ellos hispanohablantes. Con cada usuario se establecía una comunicación en español con el robot que constaba de diez frases, en orden aleatorio, con contenido emocional (en la forma, no en el contenido del mensaje). El entrenamiento de la red se hizo previamente, con otros usuarios realizando experimentos similares. Los resultados fueron analizados para cada participante, comparando la salida del clasificador para cada una de las sentencias en la conversación, con el contenido emocional de

la misma, conocido por el evaluador humano experto. Los resultados medios obtenidos para el conjunto de los participantes se representan en la matriz de confusión de la Tabla 4.2. Como puede observarse en esta tabla, el porcentaje de acierto para todos los estados emocionales es cercano o superior al setenta por ciento, valores más que aceptables para un sistema de reconocimiento basado en voz. Algunas emociones con una valencia negativa (tristeza y enfado) y el propio estado neutral presentan los mejores resultados en este sistema, superando el setenta por ciento de acierto. Como se desprende de la matriz de confusión, aparte de fallos en la clasificación (con valores en todos los casos inferior al cinco por ciento) existen errores en el sistema, donde no existe salida emocional o ésta es ambigua.

Test $P_{Speech1}$	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**87** %	0 %	0 %	0 %	2 %	11 %
Felicidad	0 %	**71** %	5 %	0 %	0 %	24 %
Miedo	2 %	2 %	**67** %	5 %	0 %	24 %
Enfado	2 %	0 %	5 %	**78** %	0 %	15 %
Neutral	5 %	0 %	0 %	0 %	**82** %	13 %

Cuadro 4.2: Resultados del sistema de reconocimiento de emociones basado en la voz humana, por medio de usuarios no entrenados en tiempo real (Cuadro obtenido de la publicación [Cid et al., 2014]).

El segundo de los experimentos consiste en evaluar el sistema de reconocimiento presentado por medio del uso de la base de datos $SAVEE$. En este caso se utilizan ficheros de audio de cuatro usuarios diferentes que expresan, en inglés, frases con contenido emocional (igual que en el caso anterior, en la forma de expresarlo, no en el propio mensaje). Esta base de datos es descrita en detalle en el Apéndice A.6. La evaluación seguida en este experimento sigue un procedimiento similar al anterior, de forma que se parte de las señales de audio del usuario para obtener una salida del clasificador bayesiano. El estado emocional del interlocutor obtenido tras su paso por el sistema es comparada, en cada sentencia, con la emoción inicial conocida por el evaluador experto. Para el entrenamiento de la red se ha utilizado la propia base de datos, usando para ello una décima parte de los archivos totales de las emociones.

La Figura 4.5a-c muestra la señal de dos usuarios (DC y KL) de la base de datos para un mismo estado emocional (en este caso, enfado). Esta señal es procesada por el sistema para obtener los valores de $Pitch$, Energía y $Tempo$ (Figura 4.5b-d) según se desprende de la descripción de cada una de las fases del sistema. El Cuadro 4.3 ilustra la matriz de confusión que resume los resultados del reconocedor. De la misma forma que en el experimento anterior, los estados emocionales con valencia negativa (tristeza y miedo), junto con el estado neutral, presentan tasas de acierto elevadas, superior al ochenta por ciento, con fallos en la clasificación por debajo del seis por ciento en todos los casos. Tal y como ocurría en el test anterior, los errores por ambigüedades o la no existencia de valores a la salida del reconocedor están acotados, y sólo en los estados emocionales de alta intensidad presentan un valor elevado (en torno al 25 %).

La Tabla 4.4 muestra un resumen de ambos experimentos, $P_{speech1}$ y $P_{speech2}$, respectivamente, donde se reflejan los porcentajes de acierto para cada uno de los estados emocionales. Tal y como se muestra en dicha tabla, los resultados en ambos experimentos siguen tendencia

Test $P_{Speech2}$	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**83** %	0 %	0 %	0 %	4 %	13 %
Felicidad	0 %	**76** %	3 %	0 %	0 %	21 %
Miedo	0 %	3 %	**81** %	4 %	0 %	12 %
Enfado	1 %	0 %	6 %	**67** %	0 %	26 %
Neutral	4 %	0 %	0 %	0 %	**89** %	7 %

Cuadro 4.3: Resultados de la segunda evaluación del sistema de reconocimiento de emociones basado en la voz humana,por medio de la base de datos $SAVEE$.

similar, y las diferencias entre uno y otro son mínimas y sin posibilidad de hacer conclusiones sobre las mismas. Sin embargo, es importante mencionar que estos resultados permiten validar este sistema, a pesar de que cada experimento utilizó diferentes fuentes de información de entrenamiento y entornos dispares.

Experimentos	Tristeza	Felicidad	Miedo	Enfado	Neutral
$P_{Speech1}$	**87** %	71 %	67 %	**78** %	82 %
$P_{Speech2}$	83 %	**76** %	**81** %	67 %	**89** %

Cuadro 4.4: Tabla comparativa entre los resultados del experimento 1 ($P_{Speech1}$) y el experimento 2 ($P_{Speech2}$).

Finalmente, se analizaron los errores en el sistema de reconocimiento para ambos experimentos que aparecen reflejados en la última columna de las tablas 4.2 y 4.3. Un desglose de estos errores se presenta en el Cuadro 4.5. En el mismo se observan los tres errores más frecuentes en el sistema propuesto en este capítulo. En primer lugar, un resultado erróneo en la clasificación, fruto de obtener una salida del estado emocional diferente al real. En segundo lugar, la ambigüedad en el resultado, esto es, un resultado oscilante en el estado emocional de salida. Finalmente, el error asociado a no obtener salida en el sistema, motivado por no superar el umbral establecido en el sistema para considerar un estado emocional. La mayor parte de estos errores tienen como origen la propia subjetividad de la información adquirida de cada uno de los estados emocionales en el entrenamiento inicial, la cual varía entre cada usuario y ocasiona lagunas o errores en la información de aprendizaje necesaria para la estimación de los estados emocionales. No obstante, en el caso especifico de los errores relacionados con la ambigüedad y los límites del umbral, estos se presentan en múltiples ocasiones como problemas del clasificador, pero en otras pruebas suele deberse a que el usuario expresa estados emocionales más cercanos a otras emociones que están fuera del rango utilizando en este sistema (por ejemplo, Sorpresa o Cansancio).

4.4. Conclusiones

La voz humana es una fuente de información para estimar el estado emocional de un interlocutor, no sólo por el contenido en sí del mensaje, si no también por los propios cambios que

Errores	Clasificación errónea	Ambigüedad	Bajo el umbral
$P_{Test(Speech1)}$	11 %	2 %	5 %
$P_{Test(Speech2)}$	13 %	1 %	2 %

Cuadro 4.5: Detalle de los errores (La información parcial del Cuadro fue obtenida de la publicación [Cid et al., 2014]).

ésta presenta cuando se expresa una idea con una emoción u otra. De forma totalmente inherente y no controlada por la persona, la voz modifica sus características en función del estado emocional. Es por ello que a lo largo de los años se han desarrollado técnicas de análisis de la voz para conseguir resultados en este ámbito.

En este capítulo se ha presentado un sistema de reconocimiento de emociones basado en características acústicas de bajo nivel de la voz humana durante una interacción. El análisis de la prosodia de la voz humana ha sido la base del sistema propuesto. En particular, se ha hecho uso del $Pitch$, de la Energía y del $Tempo$ como características de bajo nivel relacionadas con cada una de las emociones estudiada. Los valores de estas características conforman la entrada de un clasificador basado en una red dinámica bayesiana, que de nuevo busca la convergencia a una emoción según su evolución en instantes de tiempo consecutivos. El sistema de reconocimiento contribuye al estado del arte actual al proporcionar una solución completa y eficiente basada únicamente en la información verbal real, con interesantes resultados tanto para entornos controlados, es decir, sin ruido y con sistemas de captura profesionales [Sebe et al., 2005], como para ambientes no controlados.

El sistema descrito en este capítulo es utilizado como una entrada más dentro del sistema multimodal que se presenta en el Capítulo 6.

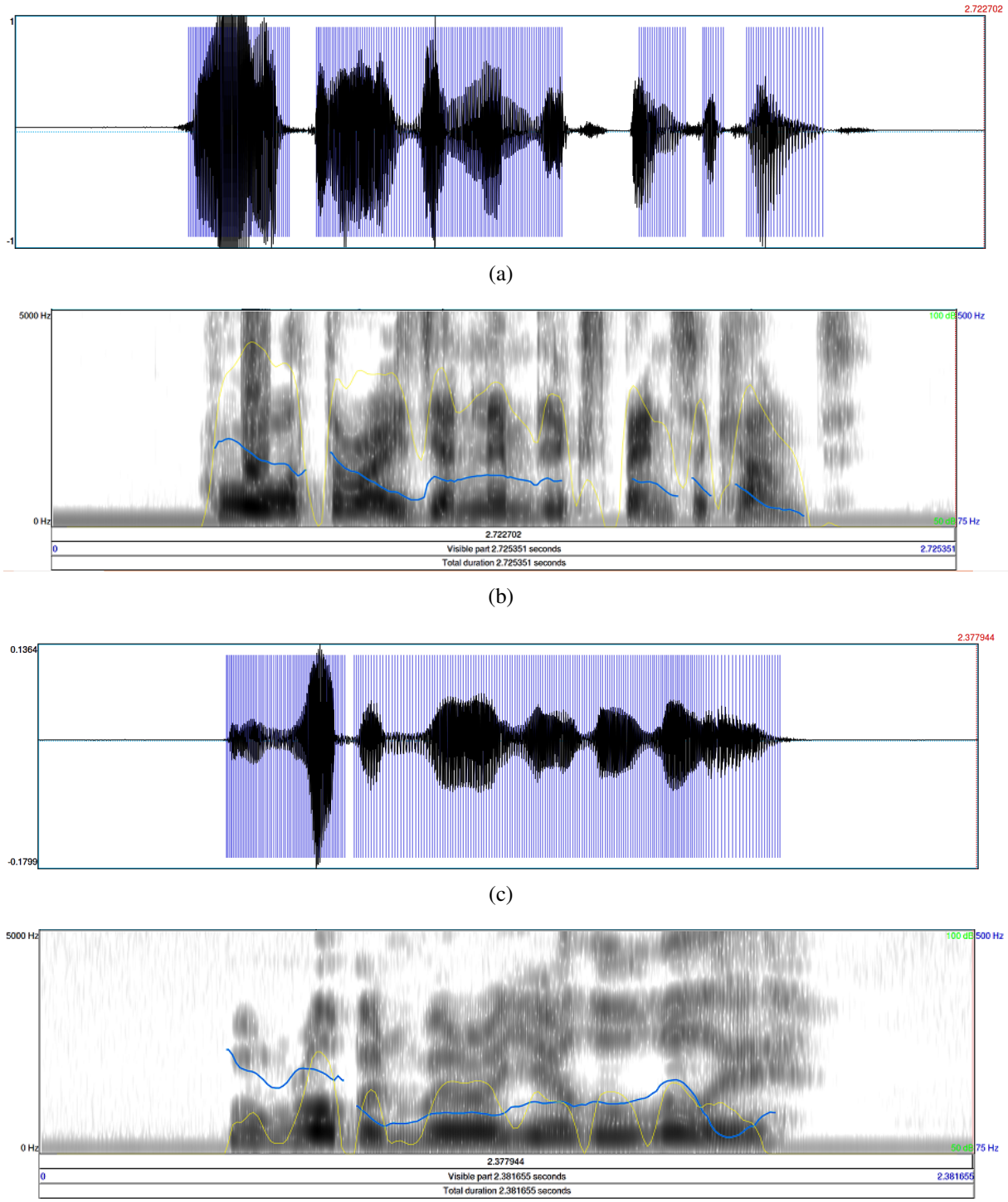


Figura 4.5: Señal de audio visualizada por medio del *toolkit* $Praat$ [P. Boersma and D. Weenink, 2014], obtenida por medio de la base de datos audiovisual $SAVEE$; a) Señal original desde el usuario DC que genera el estado emocional enfado; b) $Pitch$ (azul) y Energía (amarillo) desde la señal del usuario DC. c) Señal original desde el usuario KL que genera el estado emocional enfado; y d) $Pitch$ (azul) y Energía (amarillo) desde la señal del usuario KL.

Capítulo 5

Sistema de reconocimiento de emociones basado en el lenguaje corporal

5.1. Introducción

A lo largo de los capítulos previos se ha enfatizado el interés por conseguir interacciones humano-robots fluidas e intuitivas, lo más semejante posible a una comunicación real entre humanos en su vida cotidiana. Dos personas, durante una conversación, no sólo envían mensajes vocales y varían sus expresiones faciales según el contexto, sino que también gesticulan y modifican, por ejemplo, la posición de sus manos, brazos o cuello. El estado emocional de un interlocutor queda también reflejado tanto en las expresiones faciales como en la forma de expresar los mensajes (véase los capítulos 3 y 4), pero de una manera significativa aparece igualmente en el lenguaje corporal. La relevancia de la información visual (no-verbal), ya sean las expresiones faciales o el propio lenguaje corporal, se hace presente en múltiples estudios psicológicos que destacan cómo más de la mitad del mensaje enviado desde el locutor al receptor se realiza por medio de las expresiones corporales. En [Kurtenbach and Hulteen, 1992], los autores señalan que los gestos humanos pueden ser definidos como *un movimiento del cuerpo que contiene información emocional significativa*, siendo incorporados recientemente en múltiples sistemas de reconocimiento de emociones. Por ejemplo, si una persona se inclina hacia adelante y agita enérgicamente las manos podemos decir que se encuentra enfadada, mientras que si habla con las manos caídas y moviéndolas lentamente, esa persona probablemente esté triste.

Dentro del lenguaje corporal, la fuente de información más importante la constituyen los movimientos de la parte superior del cuerpo humano, tanto los brazos y manos, como el tronco y la propia cabeza. La mayoría de los trabajos en la literatura centran su esfuerzo en reconocer el esqueleto, y a partir de ahí, obtener características de alto nivel en el movimiento humano y su relación con las emociones. Existen varias teorías que estudian estas relaciones, siendo el análisis del movimiento emocional de Laban (***LMA***) una de las más importantes [Laban and Lawrence, 1947]. En ella, Laban otorga a cada movimiento emocional durante la danza cuatro categorías diferentes: $Cuerpo$, $Esfuerzo$, $Forma$ y $Espacio$. La primera categoría hace referencia a la relación de unas partes del cuerpo con otras. Por su parte, la categoría $Esfuerzo$, también denominada por Laban como dinámica del movimiento, se centra justo en las propiedades dinámicas del mismo (por ejemplo, el peso del individuo, aceleraciones, velocidades de las extremidades, etc). La categoría $Forma$ estudia, entre otras cosas, la posición que adopta el cuerpo durante el movimiento. Finalmente, la categoría $Espacio$ lo concibe a

partir del cuerpo de la persona que ejecuta el movimiento, estando delimitado por el radio de acción normal de cada uno de los miembros del cuerpo en su máxima extensión a partir del cuerpo inmóvil. En este marco teórico, Laban describía cómo debían ser estas categorías según qué emoción quiera expresarse.

Por su parte, desde el punto de vista de la metodología seguida en la literatura para el reconocimiento de emociones en base al estudio del lenguaje corporal, la mayor parte de los trabajos hace uso de secuencias consecutivas de imágenes RGB. A pesar de los múltiples problemas que conlleva esta fuente de información debido, entre otras, a que las condiciones de luz, el ruido y las sombras crean errores durante el proceso de extracción de características, durante años han sido las técnicas más utilizadas [Gonzalez-Sanchez and Puig, 2011], [Kessous et al., 2010], [Mancas et al., 2010]. Sin embargo, la aparición de nuevos sensores que proveen simultáneamente de información RGB y de profundidad en tiempo real ha permitido cambiar este enfoque, y de forma paulatina, diferentes grupos de investigación comienzan a utilizar información RGB-D para la extracción de características más completas que aquellas basadas únicamente en información de color.

En este capítulo se describe el sistema de reconocimiento de emociones con el que se contribuye en esta Tesis Doctoral. La elección de las características del lenguaje corporal usadas en el sistema propuesto ha sido realizada tras un estudio y revisión bibliográfica de cómo éstas quedan afectadas por el estado emocional del interlocutor. En este punto, destaca la relación directa de las características seleccionadas con la teoría de Laban y sus cuatro categorías. Por su parte, el sistema aquí descrito utiliza la información RGB-D adquirida por un sensor $Kinect$, así como un modelo interno de representación del usuario que es utilizado para el seguimiento del esqueleto durante la interacción. El método propuesto en este trabajo se basa en la detección de un conjunto de características dinámicas asociadas a los movimientos de la parte superior del cuerpo del usuario. El método extiende los sistemas presentados en [Kessous et al., 2010] y [Mancas et al., 2010], donde los autores utilizan técnicas similares para extraer diferentes características del lenguaje corporal pero a partir de información RGB.

5.2. Sistema de reconocimiento de emociones basado en el lenguaje corporal

5.2.1. Descripción del sistema

El sistema de reconocimiento de emociones descrito en este capítulo utiliza la información visual y de profundidad adquirida por el robot durante la interacción con un humano. Esta información es procesada para extraer un conjunto de elementos claves del lenguaje corporal que posteriormente permitan estimar el estado emocional del usuario. La Figura 5.1 ilustra una visión general del sistema propuesto. Como se observa en la figura, y siguiendo un esquema similar a los presentados en los capítulos precedentes, el método se divide en 3 etapas consecutivas:

- *Detección y seguimiento del esqueleto humano:* la primera etapa se encarga de la adquisición y procesamiento de los datos del sensor RGB-D. A la salida de esta etapa, el sistema devuelve en tiempo real la posición de las articulaciones principales del cuerpo del usuario, aquellas relacionados con la parte superior del mismo. Para ello se hace uso de la

librería OpenNI [OpenNi Organization, 2014], que permite, a partir de la información RGB-D, la detección y posterior seguimiento o $tracking$ del esqueleto del interlocutor.

- *Extracción de características:* esta etapa analiza los movimientos y posiciones 3D de las articulaciones de la parte superior del cuerpo (desde la cintura hacia arriba) durante el tiempo de captura, en búsqueda de características relacionadas o afectadas por los estados emocionales del usuario.
- *Red bayesiana dinámica:* La última etapa considera como entrada el conjunto de características extraídas en la etapa anterior para estimar el estado emocional del usuario. Al igual que el clasificador descrito en los capítulos 3 y 4, en este sistema se utiliza una red dinámica bayesiana, que considera no sólo el instante presente, sino la evolución temporal de las características de entrada.

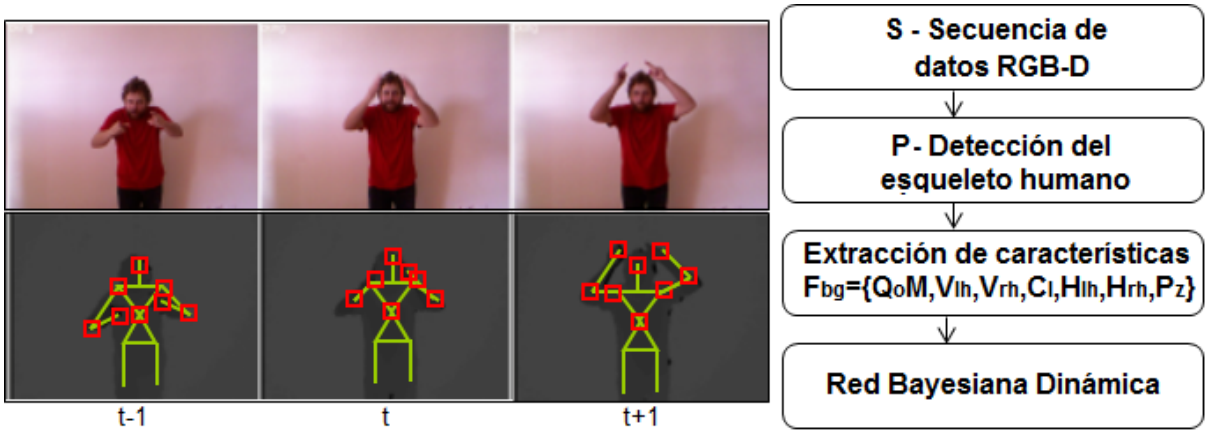


Figura 5.1: Vista general del sistema.

5.2.2. Detección y $tracking$ del esqueleto humano

La información adquirida por el sensor RGB-D es procesada por la librería OpenNI para extraer el esqueleto del usuario en tiempo real. El algoritmo utilizado implementa el método descrito en [Shotton et al., 2011], de forma que se consigue un conjunto de 31 puntos 3D asociados a las articulaciones principales del cuerpo humano. Del total de puntos devuelto por el método, sólo aquellos relacionados con la parte superior del esqueleto son utilizados, en concreto ocho articulaciones que se corresponden con la posición de la cabeza $H_{(x,y,z)}$, el hombro derecho $RS_{(x,y,z)}$ e izquierdo $LS_{(x,y,z)}$, el codo derecho $RE_{(x,y,z)}$ e izquierdo $LE_{(x,y,z)}$, la mano derecha $RH_{(x,y,z)}$ e izquierda $LH_{(x,y,z)}$ y el torso $T_{(x,y,z)}$. Así, el esqueleto humano está determinada por un vector P de 24 dimensiones. La Figura 5.2 muestra un esquema de las posiciones 3D de las articulaciones utilizadas en este trabajo, junto con el esqueleto detectado por el algoritmo.

5.2.3. Extracción de características

El conjunto de articulaciones obtenido en la etapa anterior es utilizado en esta fase para la extracción de características del lenguaje corporal que presenten una posible relación con

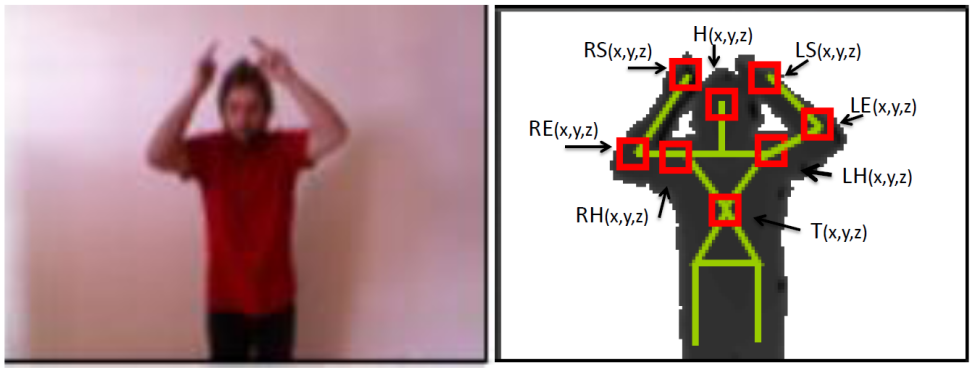


Figura 5.2: Articulaciones asociadas al modelo del esqueleto humano.

la emoción de la persona. En el enfoque presentado en este trabajo se ha utilizado el Análisis de movimiento de Laban (LMA) [Laban and Lawrence, 1947] y sus principales categorías ($Cuerpo$, $Esfuerzo$, $Forma$, y $Espacio$), dando como resultado la elección de siete características en el espacio tridimensional: las velocidades del movimiento de la mano derecha v_i^{rh} e izquierda v_i^{lh}, las alturas normalizadas de la mano izquierda H^{lh} y derecha H^{rh}, la cantidad de movimiento ***QoM***, el índice de contracción C_i y la proximidad P_z. Por un lado, QoM, v_i^{rh} y v_i^{lh} están relacionadas con la categoría $Esfuerzo$, al formar parte de lo que Laban considera como dinámica del movimiento. Por su parte, P_z está asociada al $Espacio$, al considerar la posición del usuario y de sus articulaciones para ejecutar el movimiento según su máxima extensión. Finalmente, C_i, H_{rh} y H_{lh} están relacionadas a la categoría $Forma$, al indicar, entre otras cosas, la posición que toma el cuerpo del usuario durante el movimiento. La salida del sistema es vector f_{bg} que define el movimiento del interlocutor. En la figura 5.3 se ilustran gráficamente estas variables.

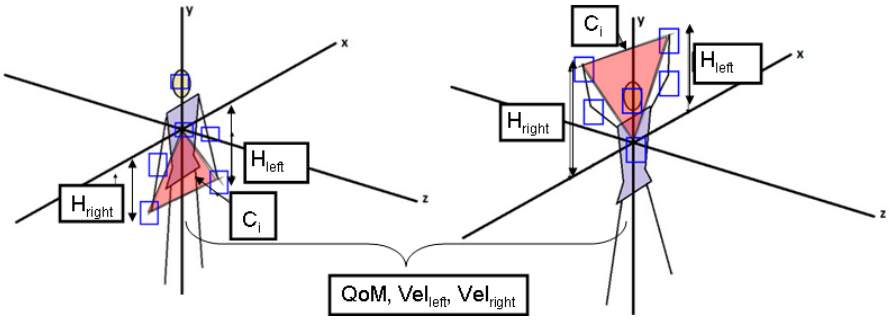


Figura 5.3: Representación gráfica del conjunto de características invariables extraídas desde la información del usuario (Figura obtenida de la publicación [Doblado et al., 2013]).

Dado un intervalo de N $frames$ en una secuencia S de imágenes RGB-D, el vector f_{bg} está formado por las siguientes características:

- **Velocidades del movimiento de las manos**.

  Cuando un humano expresa una idea u opinión en una comunicación, sus manos, en general, acompañan las palabras con movimientos, que varían su velocidad dependiendo del nivel de excitación o intensidad del usuario (mayor excitación es igual a mayor velocidad). Esta particularidad del lenguaje corporal es recogida en el sistema por medio de lo

que se ha denominado como la Velocidad del movimiento de las manos. Como su propio nombre indica, estas características corresponden a la velocidad del movimiento de cada una de las manos del interlocutor, siendo representada por las variables v_i^{rh} para la velocidad de la mano derecha y v_i^{lh} para la velocidad de la mano izquierda. El cálculo de esta características toma en consideración no sólo el instante actual, sino que requiere conocer lo N $frames$ previos. Así, dada la posición 3D de la mano izquierda en un instante de tiempo i, x_i^{lh}, y sea t el tiempo de adquisición de datos por parte del sensor RGB-D, la velocidad de la mano izquierda en ese instante de tiempo i queda como:

$$v_i^{lh} = \frac{1}{N} \cdot \sum_{k=i-N-1}^{i} \frac{(x_k^{lh} - x_{k-1}^{lh})}{t} \tag{5.1}$$

De este mismo modo, en el caso de la velocidad de movimiento de la mano derecha v_i^{rh}, se utiliza la Ecuación (5.1), utilizando ahora la posición 3D de la mano derecha x_i^{rh}.

- **Alturas normalizadas de la mano izquierda y derecha**

  El análisis del lenguaje corporal durante una interacción real demuestra que la altura de los brazos y de las manos es afectada por la intensidad de los estados emocionales del usuario. Una emoción con alta intensidad contiene una mayor cantidad de movimiento de los brazos desde el límite de altura superior al límite de altura inferior. En cambio, un estado con baja intensidad suele venir acompañado con una posición más o menos fija de los brazos y manos a una altura inferior. Así, el sistema presentado introduce como características las alturas normalizadas de las manos, representadas por las variables H_{lh} para la mano izquierda y H_{rh} para la mano derecha, las cuales, de forma similar a las características anteriormente mencionadas, presentan una dependencia con la posición 3D de la mano derecha x_i^{rh} e izquierda x_i^{lh}, respectivamente. Para estimar el valor de cada característica se realiza un cálculo inicial por medio de la coordenada y de los vectores de posición de cada mano, siendo finalmente normalizada en relación a la coordenada y del vector de posición del torso. El cálculo de esta variable toma un valor promedio de los N $frames$ anteriores al instante actual i. En la Figura 5.3 se muestran las alturas H_{lh} para la mano izquierda y H_{rh} para la mano derecha, como la distancia desde la coordenada y del centro del pecho (centro del sistema de referencia) con respecto a las coordenadas y de la posición de cada una de las manos.

- **Cantidad de movimiento**

  Similar al movimiento de las manos, el total del cuerpo humano es afectado por el estado emocional del interlocutor durante la interacción. En este sistema, QoM (*Quantity of Motion* es una característica descrita como la cantidad de movimiento que se puede detectar y cuantificar por medio de la información RGB-D adquirida desde la librería OpenNI. Estados emocionales de alta intensidad presentan, en general, valores altos de QoM, pues la mayor parte del cuerpo se encuentra en movimiento. Al contrario ocurre con estados emocionales de baja intensidad. Por este motivo, la estimación de QoM se realiza a través de la información relacionada con las posiciones 3D de las articulaciones de la parte superior del esqueleto durante los N $frames$ anteriores al instante de tiempo actual i. Así, el cálculo de QoM es definido por:

$$QoM_i = \frac{1}{N} \cdot \sum_{k=i-N-1}^{i} x_k^A - x_{k-1}^A \tag{5.2}$$

Donde, x_i^A es la posición 3D de las articulaciones en el instante de tiempo "i" siendo $A \in$ (*mano derecha, mano izquierda, codo izquierdo, codo derecho y cabeza)*. En el cálculo de esta variable no se tiene en cuenta la información del torso del esqueleto, que se supone que tiene un movimiento limitado.

- **Índice de Contracción**

  Esta variable cuantifica el grado de contracción del cuerpo humano durante la comunicación, tomando en consideración principalmente la relación entre el pecho y la posición de las manos. Estos elementos, es decir, el centro del torso, la posición de la mano derecha y la posición de la mano izquierda, conforman un triángulo cuya área es afectada directamente por el estado emocional del interlocutor durante la comunicación. Este triángulo es ilustrado en la Figura 5.3. Para estimar el índice de contracción C_i se utilizó la fórmula de Herón, definida por la siguiente ecuación:

$$C_i = \sqrt{s \cdot (s-u) \cdot (s-v) \cdot (s-w)} \tag{5.3}$$

  Donde, s representa el semi-perímetro del triángulo, obtenida por medio de la ecuación 5.4, y siendo u, v y w los lados del triángulo.

$$s = \frac{u+v+w}{2} \tag{5.4}$$

- **Proximidad**

  Esta característica está asociada a la dirección del movimiento del pecho con respecto a la posición 3D del sensor. Para calcular esta variable se normaliza la posición 3D del pecho por medio de la coordenada z, y tiene en cuenta la distancia del usuario al sensor, por medio de la siguiente ecuación:

$$Pz_i = \frac{z_i}{z_{ref}} \tag{5.5}$$

  En el caso que P_z sea un valor positivo, se considera que el usuario se acerca al sensor RGB-D en un instante de tiempo i.

Finalmente, estas 7 características son utilizadas en la siguiente etapa como variables de entrada en el proceso de clasificación. Cada una de estas variable será agrupada de acuerdo a un rango de posibles valores, el cual permitirá identificar las relaciones entre estas características y los estados emocionales. En el Cuadro 5.1 se describe esta relación. Como se observa en la tabla, existen estados que presentan valores similares en múltiples variables, como los estados asociados a emociones de baja intensidad (por ejemplo, la tristeza y el estado neutral) y aquellos otros asociados a los estados emocionales de alta intensidad y valencia negativa (miedo y enfado).

Variables	Felicidad	Enfado	Neutral	Tristeza	Miedo
Altura normalizada de las manos	Elevada	Media-Alta	Media-Baja	Baja	Media
Velocidad del movimiento de las manos	Elevada	Elevada	Baja	Muy baja	Elevada
QoM - Cantidad de movimiento	Elevada	Elevada	Media-Baja	Muy baja	Elevada
Proximidad P_z	1	< 1	1	1	> 1
Índice de Contracción C_i	Elevado	Bajo	Medio-Bajo	Medio	Bajo

Cuadro 5.1: Relación entre las características extraídas y los diferentes estados emocionales de este sistema.

5.2.4. Red bayesiana dinámica

La última etapa de este sistema es la encargada de estimar el estado emocional de la persona durante la comunicación de entre los cinco estados posibles definidos en esta Tesis Doctoral. Al igual que en los sistemas de reconocimiento de emociones basado en expresiones faciales y en el análisis del habla, esta etapa se implementa siguiendo un enfoque bayesiano y mediante el uso de una red dinámica. Como se ilustra en la Figura 5.4, esta red está compuesta por una estructura de dos niveles y una propiedad de dependencia del tiempo, donde la única variable del primer nivel cumple el rol de nodo padre para las variables del segundo nivel. Este primer nivel lo constituye la variable HE ($HE_{[Neutral]}$, $HE_{[Felicidad]}$, $HE_{[Tristeza]}$, $HE_{[Miedo]}$, $HE_{[Enfado]}$). El segundo nivel está formado por las variables obtenidas de las características del lenguaje corporal descritas en la subsección anterior:

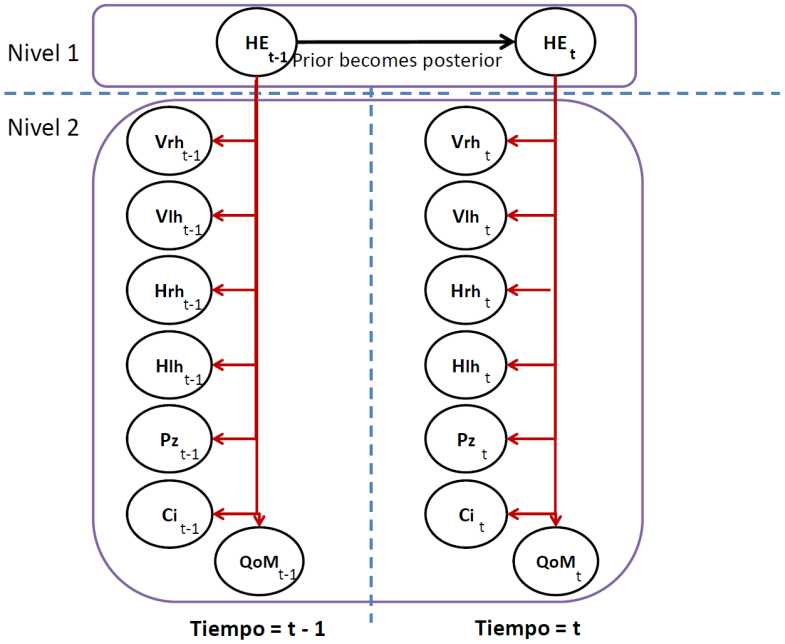


Figura 5.4: Red bayesiana dinámica, donde se muestra un intervalo de dos instantes de tiempo consecutivos (t-1, t).

- V_{rh}: Velocidad de movimiento de la mano derecha.

- V_{lh}: Velocidad de movimiento de la mano izquierda.
- H_{rh}: Altura normalizada de la mano derecha.
- H_{lh}: Altura normalizada de la mano izquierda.
- QoM: Cantidad de movimiento de los elementos de la parte superior del cuerpo.
- C_i: Índice de Contracción.
- P_z: Proximidad (distancia normalizada del centro del pecho a la posición 3D del sensor).

La red bayesiana se presenta con una estructura común al clasificador descrito en el Capítulo 3, con la cual comparte tanto las propiedades del modelo del clasificador, como el proceso de entrenamiento inicial e incluso el mismo umbral necesario para la convergencia de la red en el tiempo. La distribución de probabilidad conjunta asociada a las variables del segundo nivel de la red bayesiana, que son independientes entre ellas, se calcula mediante la ecuación 5.6.

$$\begin{aligned} P(HE, V_{lh}, V_{rh}, H_{lh}, H_{rh}, QoM, C_i, P_z) \\ = P(V_{lh}, V_{rh}, H_{lh}, H_{rh}, QoM, C_i, P_z \mid HE) \cdot P(HE) \\ = P(V_{lh} \mid HE) \cdot P(V_{rh} \mid HE) \cdot P(H_{lh} \mid HE) \cdot P(H_{rh} \mid HE) \\ \cdot P(QoM \mid HE) \cdot P(C_i \mid HE) \cdot P(P_z \mid HE) \cdot P(FE) \end{aligned} \tag{5.6}$$

5.3. Resultados experimentales

El proceso de evaluación descrito en esta sección tiene como objetivo verificar el rendimiento del sistema de reconocimiento, utilizando para ello usuarios no entrenados en un entorno no controlado. El sistema ha sido implementado en el componente software $bodyrecognitionComp$, del $framework$ RoboComp. Para estos experimentos se ha utilizado la base de datos de movimientos desarrollada en [Doblado et al., 2013], la cual contiene la información de 20 participantes (10 hombres y 10 mujeres) realizando diferentes gestos corporales afectivos, actuados y no actuados, para cada uno de los estados emocionales estudiados en este trabajo. La información es adquirida a partir de un sensor RGB-D $Microsoft\ Kinect$ y el equipo utilizado en un ordenador Intel Core i7 con 4Gb de RAM. En primer lugar, el sistema analiza el número de $frames\ N$ óptimo para el sistema, para ello realiza diferentes experimentos con gestos emocionales y analiza los resultados de las características extraídas según este parámetro. Una vez elegido este valor de N, se realizan los experimentos de clasificación necesarios para evaluar la precisión del sistema.

5.3.1. Estimación de parámetros del sistema

El valor del número de $frames\ N$ es crucial para una correcta detección de las características. Los movimientos ligados al lenguaje corporal están condicionados a una duración mínima de tiempo. Así, un valor de N demasiado bajo puede no dar información significativa, estando más relacionado con ruido asociado al movimiento. Por contra, un valor de N demasiado alto implicaría mucho tiempo dedicado a la detección de característica, dando lugar presumiblemente a resultados erróneos al estar los movimientos ligados a diferentes estados emocionales. Se

realizaron diferentes experimentos con valores de N (N = 5, N = 50 y N = 200) y se determinaron criterios de estabilidad para decidir el valor final elegido para el resto de experimentos de este capítulo. La Figura 5.5 muestra la evolución en el tiempo de las variables QoM y V_{rh} según estos valores de N para diferentes movimientos humanos, correspondiendo al color cian a N = 5, azul a N = 50 y rojo a N = 200. Como se observa en la figura, un valor de N elevado elimina todas posible información significativa, mientras que un valor demasiado pequeño introduce ruido en el cálculo de las variables, aparte es un valor poco significativo para detectar una emoción. Dado que el sistema está diseñado para trabajar en tiempo real, se ha elegido un valor de N = 50 para el resto de experimentos, de forma que el sistema procesa las características en aproximadamente dos segundos antes de obtener el estado emocional.

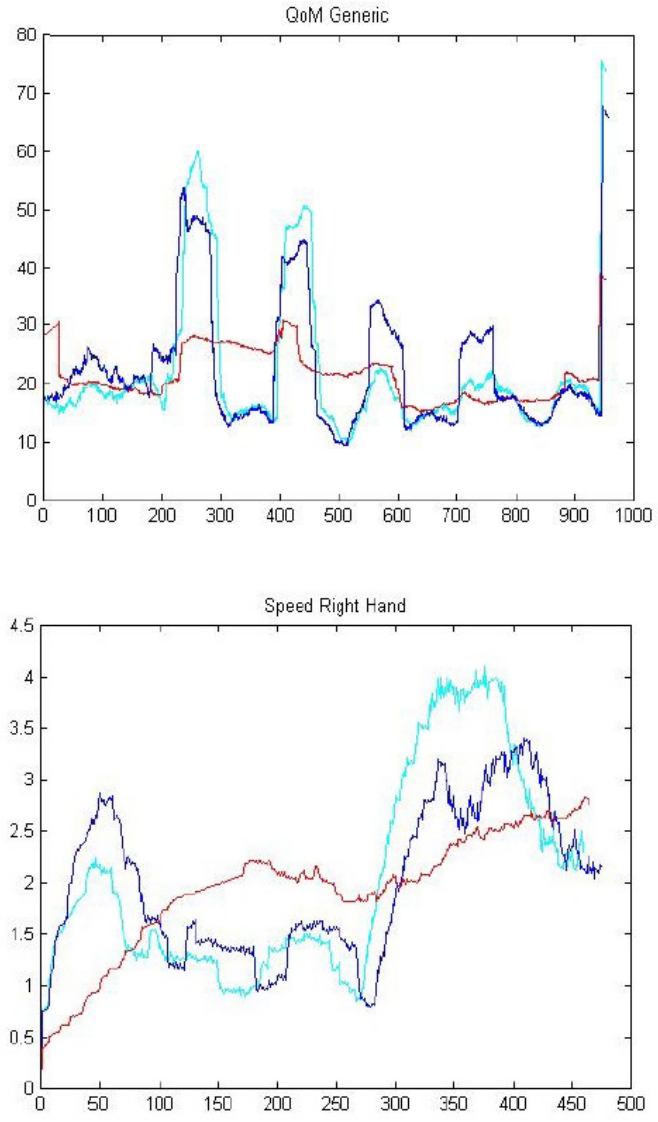


Figura 5.5: Evolución en el tiempo de las características QoM y V_{rh}, para diferentes valores de N. En cian, N = 5, azul N = 50 y rojo N = 200.

5.3.2. Evaluación del sistema

Como se comenta en el trabajo [Doblado et al., 2013], los sujetos de la base de datos fueron desarrollando los diferentes movimientos afectivos, en primer lugar de forma espontánea tras reaccionar a ciertos estímulos, y a continuación de manera actuada, una vez fueron enseñados a realizar determinados movimientos afectivos asociados a cada emoción. En la Figura 5.6 se muestran algunos de los gestos afectivos realizados por los usuarios de este experimento, siendo la información de la parte superior del cuerpo la procesada para la clasificación y la posterior estimación de los estados emocionales. Los resultados de este experimento se muestran en el cuadro 5.2, donde la precisión del sistema propuesto fue de un 73,58 %. Analizando la tabla anterior, puede comprobarse que los estados emocionales de tristeza y felicidad son los que presentaron los porcentajes de precisión más elevados, seguidos por el estado neutral y enfado, respectivamente. Por contra, se observó una baja precisión en el estado miedo, presumiblemente por las similitudes en este movimiento con el estado emocional de Enfado y los consecuentes errores en el clasificador.

Figura 5.6: Gestos corporales afectivos asociados a los cinco estados emocional estudiados en este trabajo (tristeza, neutral, miedo, enfado y felicidad, respectivamente) (Figura obtenida de la publicación [Doblado et al., 2013]).

El sistema propuesto ha sido comparado con sistemas de reconocimiento similares. En [Kessous et al., 2010], se describe un sistema de reconocimiento basado únicamente en información RGB capaz de reconocer hasta seis estados emocionales diferentes y en un sistema actuado (los usuarios conocían el tipo de movimiento efectivo a realizar). El trabajo analizado en [Savva et al., 2011] utiliza un sistema de captura de movimiento profesional 3D (***MCS***) para reconocer cuatro estados emocionales diferentes en una población de individuos que reproducen movimientos afectivos no actuados. El estudio comparativo llevado a cabo en este capítulo tiene en cuenta el número de emociones detectadas, la precisión en la detección de emociones y la existencia o no de información de profundidad en las características extraídas

Test P_{LC_E}	Enfado	Miedo	Tristeza	Felicidad	Neutral
Enfado	**56,7** %	29,9 %	0 %	13,3 %	0 %
Miedo	39,6 %	**46,9** %	0 %	10,2 %	3,3 %
Tristeza	0 %	0 %	**91,3** %	0 %	8,7 %
Felicidad	9,7 %	3,3 %	0 %	**89,7** %	0 %
Neutral	0 %	3,3 %	13,3 %	0 %	**83,33** %

Cuadro 5.2: Resultados del sistema de reconocimiento de emociones basado en el lenguaje corporal (Cuadro obtenido de la publicación [Doblado et al., 2013])

por el sistema. Un resumen de los resultados puede verse en la tabla 5.3. El hecho de usar información del sensor RGB-D y un $tracking$ en tiempo real del cuerpo humano permite una mayor precisión en la detección de emociones del sistema presentado en este capítulo respecto al trabajo descrito en [Kessous et al., 2010]. Por su parte, la selección de las características y la implementación del clasificador bayesiano incrementa las tasas de acierto del sistema completo en relación a ambos sistemas, incluso mejorando el sistema de cuatro emociones desarrollado en [Savva et al., 2011].

	Método Propuesto	Imágenes RGB	MCS
Precisión	**73,58** %	67,1 %	56,25 %
Nº de emociones	5	8	4
Inf. de profundidad	Si	No	Si

Cuadro 5.3: Cuadro comparativo entre diferentes métodos de reconocimiento de emociones.

5.4. Conclusiones

En una comunicación real, el cuerpo humano actúa según los estados emocionales de los interlocutores. Así, no es lo mismo expresar un mensaje si una persona se encuentra tensa o enfadada, que si el mismo mensaje se transmite por una persona que se encuentra relajada o triste. La comunidad científica trabaja en los últimos años en este campo, dada la importancia que presenta en el desarrollo de IRH afectivas.

En este capítulo se ha presentado un sistema para el reconocimiento de gestos humanos afectivos, el cual está basado en las características extraídas analizando la parte superior del cuerpo humano por medio de la información RGB-D y un seguimiento del esqueleto. Este sistema tiene como objetivo estimar cinco posibles estados emocionales a partir de la información afectiva que existe en un humano durante una conversación. El método propuesto analiza y cuantifica los movimientos del modelo 3D del esqueleto a lo largo de un período específico de tiempo. Se presenta en este capítulo como contribución principal un conjunto de características 3D asociadas al movimiento humano, según la teoría de análisis de movimiento Laban. El uso de estas características y el modelo de sistema similar a los descritos en capítulos anteriores - uso de la red dinámica bayesiana como clasificador - han permitido obtener unos resultados

experimentales adecuados para un sistema de este tipo, mejorando incluso a aquellos extraídos de sistemas similares existentes en la literatura.

Capítulo 6

Sistema multimodal para el reconocimiento de emociones

6.1. Introducción

En los capítulos precedentes se ha destacado la importancia de la interacción humano-robot en el campo de la robótica social, donde ambos interlocutores - robot y humano - son capaces de entablar una conversación de forma natural, acompañando a las palabras con movimientos e información afectiva. En los últimos años han evolucionado las metodologías existentes para la detección del estado emocional de los usuarios. En un primer lugar fueron métodos simples, basados en el análisis de un canal de información (voz o audio, por ejemplo). Sin embargo, estos métodos se vuelven más robustos si en el resultado final se tiene en consideración diferentes canales de información afectiva y todos aportan resultados significativos. Estas métodos, conocidos como sistemas de reconocimiento de emociones multimodales, utilizan diferentes modos o fuentes de información, tan dispares como el análisis de señales de voz, gestos corporales, expresiones faciales o incluso señales eléctricas [Sebe et al., 2005].

Estos sistemas multimodales basan su funcionamiento en la hipótesis de que una emoción humana no se expresa de una única forma, si no que gran parte de la información emotiva es expresada en diferentes modalidades, ya sean visuales o auditivas. Así, las interacciones afectivas multimodales se presentan como una solución para una comunicación mucho más realista y cercana, donde un robot puede reconocer la información emocional desde múltiples enfoques o modalidades de entrada. Diferentes estudios en la literatura han avanzado en este concepto, muchos de ellos analizando información facial y auditiva y desarrollando simples metodologías para fusionar los datos de cada subsistema [Kessous et al., 2010], [Caridakis et al., 2010]. En la mayor parte de estos enfoques cada modalidad introduce información redundante en el sistema de reconocimiento de emociones, algo que lejos de ser un problema en una interacción real, resulta útil al reducir los errores asociados al ruido o las posibles oclusiones [Prado, 2012].

No obstante, este enfoque multimodal presenta una serie de problemas en su implementación práctica, debido en gran medida a que el reconocimiento de las emociones humanas por medio de múltiples enfoques se realiza en base a diferentes tiempos de respuesta, que no tienen por qué estar sincronizados entre sí (es decir, cada modo extrae información del estado emocional en un instante de tiempo determinado). Esta falta de sincronización y la dificultad que conlleva la implementación de la misma en el sistema completo al ser fuentes independientes, obliga a la existencia de una modalidad predominante, la cual mantiene como salida del siste-

ma completo la estimación ofrecida por este único modo, y sólo en determinados momentos, aquellos en los que existe una estimación en alguno de los otros sistemas, se realiza la fusión de las fuentes de información. Este concepto o estrategia basado en una modalidad predominante está presente en múltiples trabajos en la literatura, como los presentados en [Sebe et al., 2005] y [Jaimes and Sebe, 2005].

El presente capítulo describe el sistema multimodal para el reconocimiento de emociones humanas basado en un enfoque multimodal. El sistema completo integra, en un primer momento, todas las modalidades para el reconocimiento de emociones descrito en esta Tesis Doctoral, la expresión facial, descrita en el Capítulo 3, la información emocional de la voz del interlocutor, descrito en el Capítulo 4, y finalmente, el lenguaje corporal, presentado en el Capítulo 5. La integración en el sistema completo de cada una de las fuentes de información se lleva a cabo siguiendo una estrategia de fusión basada en un clasificador bayesiano dinámico, donde cada modo es analizado individualmente y una vez realizada la estimación, se fusionan los resultados en un nivel de decisión (es decir, la información se integra a partir de las modalidades individuales, después de haber sido interpretada por cada uno de los clasificadores). Como se explica en este mismo capítulo, a nivel de experimentación y dada las complicaciones que presentaba el sistema final al integrar la información corporal en la estimación de la emoción humana, se ha decidido por integrar únicamente dos fuentes de información en el sistema evaluado, las expresiones faciales y la señal de audio.

6.2. Sistema multimodal para el reconocimiento de emociones

6.2.1. Descripción del sistema

La Fig. 6.1 muestra una visión general del sistema propuesto. Como puede verse en la figura, el sistema completo integra información procedente de los tres sistemas descritos en esta Tesis Doctoral. La información del sensor RGB-D es analizada de forma independiente por el sistema de reconocimiento de emociones basado en expresiones faciales y en el lenguaje corporal, respectivamente. A su vez, la voz del interlocutor es analizado por el reconocedor de emociones basado en el análisis del habla. Los sistemas se han descrito con detalle en los capítulos precedentes, y basan su funcionamiento en un clasificador bayesiano. Un bloque *Control de tiempo* determina el instante en el que los clasificadores ofrecen a su salida la emoción detectada. Finalmente, los datos son fusionados de nuevo haciendo uso de un clasificador bayesiano que, dado los niveles probabilísticos a la entrada, ofrece un único valor de salida que determina la emoción detectada por el sistema completo. El módulo encargado de realizar la fusión es el *Nivel de decisión*. La Figura 6.2 ilustra la red bayesiana en la fusión de los datos. En esta figura, UE_t representa el primer nivel de la red bayesiana, y se corresponde con los posibles estados emocionales del interlocutor en un instante de tiempo t. En el nivel 2 de esta red dinámica se encuentran los resultados de cada una de los subsistemas descritos previamente, FE_t, AE_t y HE_t, que se relacionan con las expresiones faciales, el análisis del habla y los gestos afectivos durante la interacción.

A continuación se describen con más detalle tanto el bloque de Control de tiempo como el Nivel de decisión del sistema propuesto.

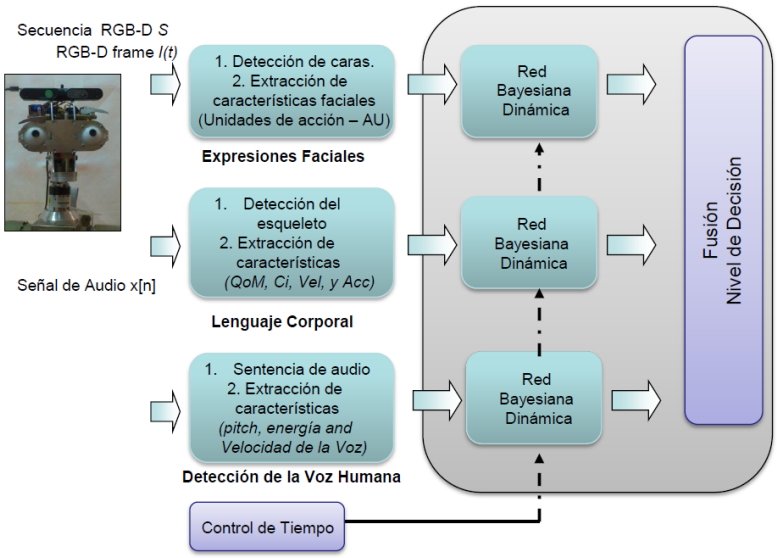


Figura 6.1: Visión general del sistema de reconocimiento multimodal.

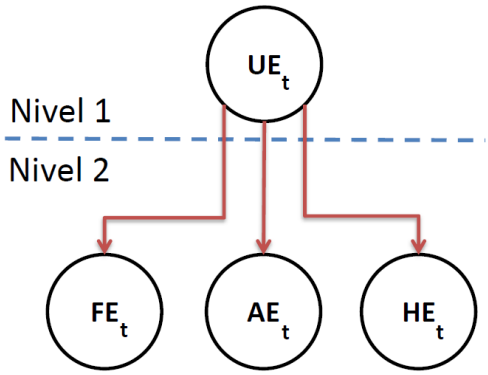


Figura 6.2: Red bayesiana basada en un enfoque multimodal global, por medio de expresiones faciales, voz humana y lenguaje corporal.

6.2.2. Control de tiempo

La principal dificultad de los sistemas multimodales radica en cómo realizar la sincronización entre los distintos bloques que conforman el reconocedor. La detección de una expresión facial no tiene por qué coincidir en tiempo con una expresión corporal, y mucho menos con la información emotiva que puede ser extraída del análisis del habla. El bloque Control de tiempo sincroniza los resultados de cada una de las redes bayesianas, pero, en lugar de dar a la salida de cada subsistema una estimación de la emoción humana en un mismo instante de tiempo, el sistema propuesto utiliza el reconocedor de emociones basado en el análisis de la expresión facial como modalidad predominante en el sistema completo [Sebe et al., 2005]. Así, sólo cuando existe información de audio o expresiones corporales durante la interacción, el Control de tiempo es el módulo encargado de que estos datos sean fusionado en el clasificador dinámico final.

La Figura 6.3 muestra el funcionamiento del módulo de Control de tiempo descrito en esta sección. En este ejemplo concreto, para facilitar su comprensión, se utiliza el sistema multi-

modal simplificado, basado únicamente en las expresiones faciales y en el análisis de la voz humana. Cuando únicamente disponemos de información facial, la salida del sistema completo se corresponde con la salida de esta única modalidad. En el caso de disponer de información de audio, el resultado del reconocedor de emociones completo, gracias al módulo de Control de tiempo, toma los datos facilitados por esta modalidad para estimar el resultado final (en color rojo en la Figura 6.3).

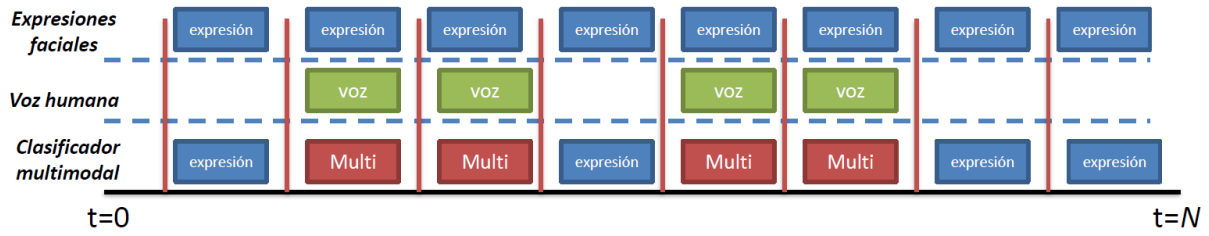


Figura 6.3: Comportamiento en el tiempo del sistema propuesto. El sistema predominante se corresponde con la salida del reconocedor de expresiones faciales. Sólo cuando hay resultados de audio se fusionan los datos en una única salida del sistema multimodal.

6.2.3. Nivel de decisión

El nivel de decisión implementa el proceso de clasificación final. Para ello analiza la información de salida de las redes bayesianas de cada una de las modalidades para estimar el estado emocional del usuario durante la interacción. Para el sistema completo, considerando las tres modalidades de información, este proceso de clasificación final se puede representar por medio de una red bayesiana de tres niveles, como se ilustra en la Figura 6.4. En esta figura, el nodo UE_t cumple el rol de nodo padre de aquellos nodos del segundo nivel FE_t, AE_t y HE_t, y representa el resultado del clasificador. Los valores de FE_t, AE_t y HE_t representan las salidas de los reconocedores de emociones basados en expresiones faciales, en el análisis de audio y en gestos corporales, respectivamente.

Para estimar el estado emocional en un instante de tiempo t, UE_t, se utilizan los datos de los nodos del segundo nivel de la red, FE_t, AE_t y HE_t, independientes entre sí. El cálculo de la distribución conjunta asociada a esta fusión bayesiana, se describe por medio de la Ecuación 6.1:

$$\begin{aligned} P(UE, FE, AE, HE) &= P(FE, AE, HE \mid UE) \cdot P(UE) \\ &= P(FE \mid UE) \cdot P(AE \mid UE) \cdot P(HE \mid UE) \cdot P(UE) \end{aligned} \tag{6.1}$$

A continuación, aplicando la regla de Bayes, obtenemos el $posterior$, que nos permite obtener el resultado del sistema completo.

$$P(UE \mid FE, AE, HE) = \frac{P(FE \mid UE) \cdot P(AE \mid UE) \cdot P(HE \mid UE) \cdot P(UE)}{P(FE, AE, HE)} \tag{6.2}$$

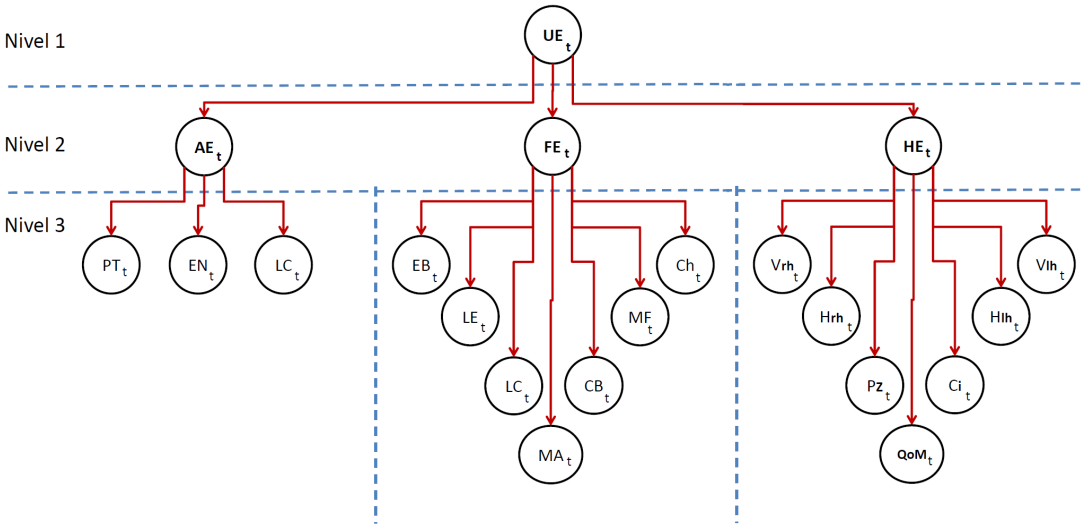


Figura 6.4: Red bayesiana basada en el enfoque multimodal descrito en este trabajo.

6.3. Resultados experimentales

La evaluación del sistema multimodal descrito en este capítulo se centra en el análisis de la precisión y robustez de los resultados obtenidos. Además, se estudian las principales diferencias respecto al uso de una única modalidad en la estimación de la emoción humana. El uso del lenguaje corporal como fuente de información en el sistema de detección de la emoción descrito en el Capítulo 5, presenta una serie de problemas asociados, en gran medida, a los tiempos de respuesta del clasificador y a la ambigüedad en el reconocimiento de muchos gestos corporales, imposibles de extraer el inicio y fin de los mismos durante una interacción real. Todo ello hace su uso inviable en el sistema completo, aún utilizando una modalidad predominante, como fue comprobado en los diferentes experimentos llevados a cabo a lo largo de las pruebas del sistema completo. Así, a pesar de que una tercera modalidad permitiría reducir ambigüedades y errores en la estimación de la emoción dada por el sistema completo, se decidió desarrollar a nivel teórico el sistema completo pero a nivel práctico implementar únicamente el sistema con dos modos, expresiones faciales y el análisis afectivo del habla. La Figura 6.5 muestra el sistema simplificado usado en los experimentos, donde sólo la información relacionada a la voz humana y las expresiones faciales es tenida en cuenta. La red bayesiana utilizada a Nivel de decisión se muestra en la Figura 6.6.

La primera parte de esta sección resume las principales características de los sistemas presentados en los Capítulos 3 y 4. A continuación, se analizan los resultados entregados por el sistema multimodal, haciendo énfasis en las principales diferencias y beneficios de estos resultados. El algoritmo del sistema presentado fue desarrollado en C++, e implementado dentro de RoboComp por medio del componente $multimodalrecognitionComp$. La fuente de información audio-visual para este experimento fue proporcionada por medio de la base de datos $SAVEE$, descrita en el Apéndice A.6.

El Cuadro 6.1 muestra los resultados obtenidos con el sistema de reconocimiento facial, tal y como fue recogida en el Capítulo 3. Dado que la base de datos utilizada sólo presenta información RGB, los datos reflejan los resultados del sistema de reconocimiento de expresiones faciales que hace uso del filtro de $Gabor$ para la detección de características. En la tabla,

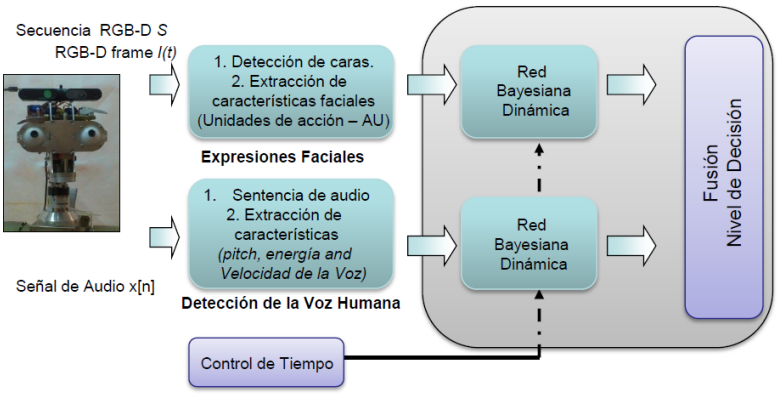


Figura 6.5: Visión general del sistema de reconocimiento multimodal basado en dos enfoques.

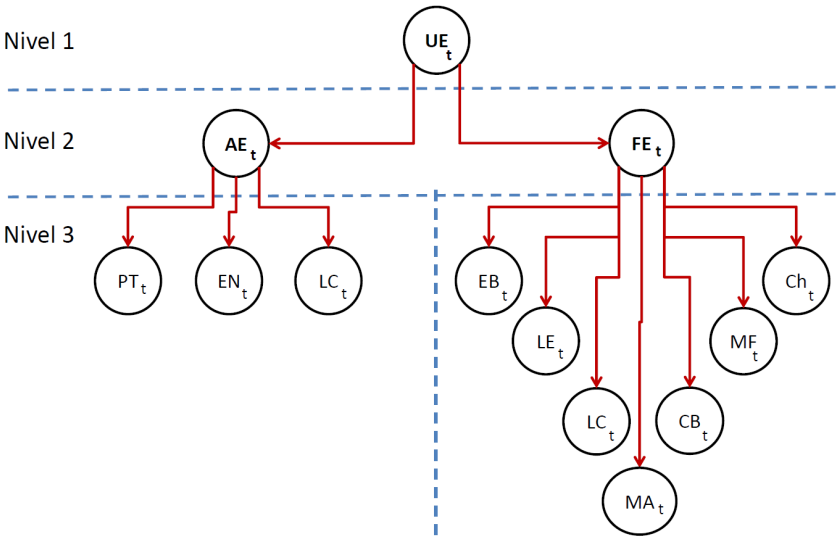


Figura 6.6: Red bayesiana basada en un enfoque multimodal reducido (bimodal) para los experimentos, por medio de expresiones faciales y la voz humana.

se comprueba cómo las expresiones con valencia negativa y alta intensidad (miedo y enfado), presentan los peores resultados.

En el Cuadro 6.2 se ilustran los resultados del sistema de reconocimiento de emociones basado en la voz humana. Este sistema, descrito con detalle en el Capítulo 4, extrae las características acústicas de la prosodia de la voz para estimar cada uno de los estados emocionales. Del análisis de la tabla se observan cómo los estados con baja intensidad (tristeza y neutral, en este caso), presentan los mejores resultados con respecto a los estados de alta intensidad (felicidad, miedo o enfado). Del análisis conjunto de las tablas 6.1 y 6.2, se demuestra un bajo rendimiento del segundo sistema en comparación con el sistema basado únicamente en expresiones faciales. Además, existe un elevado número de errores en la estimación, relacionados en gran medida a fallos en la clasificación.

En el análisis conjunto se observa igualmente que el sistema de reconocimiento basado en expresiones faciales, no sólo analiza la información en cada instante de tiempo, sino que también presenta los mejores resultados con la menor cantidad de errores asociados al clasificador,

Test $P_{Facial2}$	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**96** %	0 %	0 %	0 %	1 %	3 %
Felicidad	0 %	**98** %	0 %	0 %	0 %	2 %
Miedo	1 %	3 %	**89** %	0 %	0 %	7 %
Enfado	2 %	0 %	0 %	**93** %	0 %	5 %
Neutral	1 %	0 %	0 %	0 %	**97** %	2 %

Cuadro 6.1: Resultados del sistema de reconocimiento de expresiones faciales basado en filtrado de $Gabor$. Estos resultados se corresponden con la ejecución del algoritmo sobre la base de datos $SAVEE$, tal y como se describe en el Capítulo 3.

Test $P_{Speech2}$	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**83** %	0 %	0 %	0 %	4 %	13 %
Felicidad	0 %	**76** %	3 %	0 %	0 %	21 %
Miedo	0 %	3 %	**81** %	4 %	0 %	12 %
Enfado	1 %	0 %	6 %	**67** %	0 %	26 %
Neutral	4 %	0 %	0 %	0 %	**89** %	7 %

Cuadro 6.2: Resultados del sistema de reconocimiento de emociones basado en la voz humana, usando la base de datos $SAVEE$. Estos datos están descritos en el Capítulo 4.

como se muestra en la Tabla 6.3. Esto demuestra la importancia y las ventajas del reconocimiento de emociones basado en expresiones faciales, y de ahí su uso en el sistema multimodal propuesto como el método predominante dentro del sistema de reconocimiento de emociones.

Errores	Clasificación errónea	Ambigüedad	Bajo el umbral
$P_{Test(facial)}$	2 %	1 %	1 %
$P_{Test(Speech)}$	11 %	2 %	5 %

Cuadro 6.3: Detalle de los errores, entre el sistemas de reconocimiento de emociones basado en expresiones faciales, el sistema basado en la voz humana.

Una vez analizado los resultados individuales de cada subsistema, se utiliza la misma base de datos $SAVEE$ para probar el enfoque multimodal del sistema propuesto. La tabla 6.4 resume los datos obtenidos en el experimento. Un análisis detallado de este cuadro verifica las mejoras en el rendimiento y precisión del sistema final, aumentando las tasas de éxito en la matriz de confusión presentada. Los posibles problemas del sistema de reconocimiento basado en voz para emociones de alta intensidad son corregidos en el sistema multimodal, donde el uso de un método predominante con mayor precisión solventa las ambigüedades. Por su parte, el sistema completo permite mejorar los resultados para el resto de emociones, dando lugar a un sistema más fiable en la estimación de emociones. A pesar de todo, los efectos del sistema predominante son bastante claros en los resultados, y como se comprueba en la tabla 6.4 el sistema multimodal posee la misma tendencia a presentar problemas con los estados emocionales con valencia negativa, pero de alta intensidad (enfado y miedo).

Test $P_{Speech2}$	Tristeza	Felicidad	Miedo	Enfado	Neutral	Errores
Tristeza	**98** %	0 %	0 %	0 %	0 %	2 %
Felicidad	0 %	**99** %	0 %	0 %	0 %	1 %
Miedo	0 %	2 %	**91** %	0 %	0 %	7 %
Enfado	2 %	0 %	0 %	**95** %	0 %	3 %
Neutral	0 %	0 %	0 %	0 %	**98** %	2 %

Cuadro 6.4: Resultados de la evaluación del sistema de reconocimiento basado en un enfoque multimodal usando la base de datos $SAVEE$.

6.4. Conclusiones

En este capítulo, se describe el sistema de reconocimiento de emociones basado en un enfoque multimodal, el cual utiliza la información desde diferentes modalidades propias del lenguaje natural. Por un lado, la información visual es utilizada de forma independiente para el reconocimiento de emociones basado en expresiones faciales y en el análisis de movimientos afectivos. Por otro lado, la modalidad auditiva presenta un reconocimiento de emociones basado en la extracción de elementos de la prosodia de la voz humana. Todos los sistemas individuales son fusionados por el sistema final siguiendo un esquema clásico, basado en un modo predominante. Esta metodología permite una salida continua en el tiempo, y sólo cuando existe una estimación en cualquier otro subsistema, ésta es integrada en el sistema multimodal. El modo predominante en este caso lo constituye el reconocimiento de expresiones faciales, con salida continua en el tiempo y tasas de precisión aceptables. El sistema multimodal se ha probado sobre bases de datos bi-modales, con una mejora sustancial en el comportamiento final del mismo.

Capítulo 7

Sistema de imitación del lenguaje natural para robot sociales

7.1. Introducción

La imitación puede considerarse como una de las primeras formas de comunicación entre individuos. Imitar, para un ser vivo, es crucial no sólo para establecer comportamientos sociales, sino también para su propia supervivencia. Si un individuo ha sido capaz de llegar a una edad adulta ha sido porque, entre otras cosas, supo cómo mantenerse a salvo de todos los peligros a los que podía enfrentarse en su crecimiento. Siguiendo este esquema, si cualquier otro individuo de su especie quiere sobrevivir, sería lógico imitar el comportamiento del primero para llegar a este mismo fin. En un ámbito social, como es el caso en el que se investiga en esta Tesis Doctoral, la imitación ayuda a que exista una mayor integridad dentro de un grupo. El individuo aprende viendo cómo se hacen las cosas, escuchando cómo se dicen las cosas, y se relaciona con otros individuos según el resto lo hacen. Estas teorías psicológicas acerca del rol que cumple la imitación en el comportamiento de los individuos ha sido utilizada en las últimas décadas en el desarrollo de la robótica social y es el marco de trabajo que se presenta en este capítulo.

En primer lugar, muchos estudios describen la importancia del uso de robots con forma similar a la humana [Mori, 1970], [Picard, 2000], ya sea mediante un diseño antropomórfico o zoomórfico. El primero hace referencia a robots con forma similar a la del ser humano, con los rasgos característicos que nos hacen diferentes, como puede ser el disponer de ojos, párpados, boca o la propia forma de la cara. Los robots zoomórficos presentan formas en su diseño parecida a animales, también con los elementos físicos que lo caracterizan y diferencian de otras especies. Este tipo de diseños, en general, permite mejorar la empatía y la atención dentro de una interacción afectiva [Cid et al., 2014], [Cid et al., 2013c]. Cuánto de parecido puede llegar a ser un robot y el nivel de aceptación del mismo por un humano ha sido objeto de estudios de diferentes teorías. El principio de la misma, conocido como el valle inquietante de los robots, del inglés *Uncanny valley*, es que cuando los robots antropomórficos miran y actúan casi como un ser humano real, pueden causar una respuesta de rechazo entre los observadores.

En este contexto se plantea como hipótesis que un robot capaz de generar e imitar emociones, siguiendo un lenguaje natural, facilita una interacción entre ambos interlocutores, mejorando la aceptación por parte del humano. Esta hipótesis abre diferentes líneas de trabajo, entre ellas, cómo la propia forma del robot puede llegar a ser determinante para imitar, desde las expresiones faciales del interlocutor, hasta el lenguaje corporal [Calderita et al., 2011],

[Ge et al., 2008]. Junto con la forma del robot, es necesario estudiar cómo los grados de libertad del diseño facilita una comunicación más natural con un interlocutor humano, o cómo la capacidad de síntesis de voz con componentes afectivos puede llegar a mejorar esta interacción respecto a sólo usar información visual [Rybski et al., 2007], [Aly and Tapus, 2011].

No obstante, para percibir e intercambiar información de forma similar a los humanos es necesario que el robot disponga de una serie de sensores (acústicos y visuales) que permitan el uso de métodos no invasivos en la interacción. Para la detección de la emoción humana, como se ha visto en capítulos anteriores, el uso de sensores RGB o RGB-D posibilita una fuente de información de la que extraer una estimación a partir de la información facial y del lenguaje corporal. A su vez, disponer de micrófonos para capturar el habla del interlocutor y altavoces para poder comunicarse, permite al robot intercambiar mensajes y realimentarse del contenido de la conversación. Para esto último es necesario dotar al robot con mecanismos para el reconocimiento automático del habla (ASR) [Anderson and Kewley-Port, 1995] y para la generación sintética de habla a partir de texto (TTS) [Moberg, 2007].

En relación a cómo afecta los mensajes verbales generados por un robot en una conversación afectiva real, se han realizado diferentes estudios que demuestra que si el robot habla de una forma monótona, esto es, sin ningún cambio en el $pitch$ o en el énfasis, no existe ninguna mejora significativa con respecto al uso únicamente de la información visual [J.Cahn, 1990]. Del resultado de estos estudios se concluye que cambios en la prosodia de la voz, para añadir información afectiva, influyen directamente en una comunicación verbal, permitiendo a los robots interactuar con los humanos de forma natural mediante el uso de mensajes verbales. A pesar de lo anterior, sólo el uso de la información acústica o visual, de forma independiente, en una interacción presenta una notable diferencia con respecto a la comunicación basada en el lenguaje natural, donde, en realidad, se utilizan de forma coordinada ambos enfoques (visual y acústico). En una conversación, el habla del robot ha de estar acompañado de movimientos, no sólo asociados a la sincronización de la boca con el mensaje generado, si no también de otros elementos del robot (cuello, ojos, etc). Esto último se conoce como el efecto $McGurk$ [Chen and Rao, 1998], una idea explotada dentro de la IHR [Oh et al., 2010] [Hara et al., 1997].

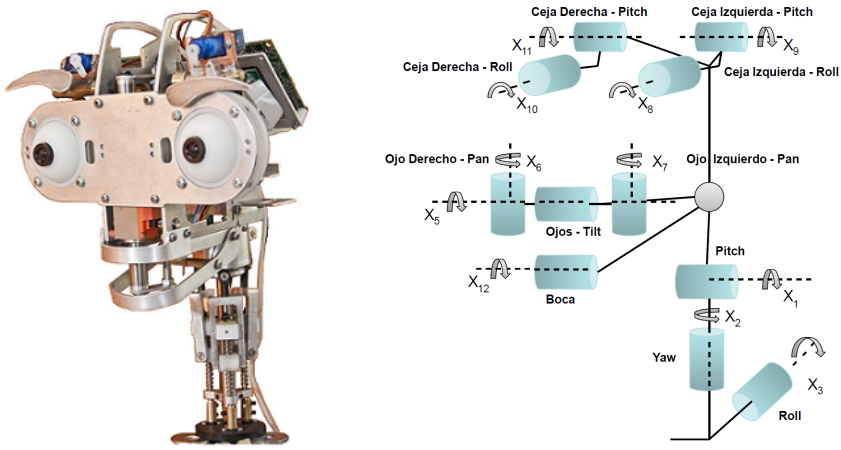
En este capítulo se explica el sistema de imitación de emociones propuesto en esta Tesis Doctoral. Se describe la plataforma robótica usada para corroborar la hipótesis planteada, así como el mecanismo en sí de imitación. La metodología seguida permite extender y exportar los rasgos emocionales obtenidos por el sistema a cualquier otro diseño antropomórfico o zoomórfico, incluso a cualquier otro diseño que pueda expresar cambios faciales - por pocos que sean - o sintetizar audio. A su vez, se describen los métodos ASR y TTS utilizados en este trabajo, que sirven de preámbulo para la descripción del mecanismo de sincronización del habla para el robot presentado.

7.2. Muecas: una cabeza robótica expresiva

En esta sección se describe brevemente la cabeza robótica Muecas, diseñada de forma conjunta por el grupo de Robótica y Visión Artificial RoboLab, de la Universidad de Extremadura, y la empresa extremeña IADEX. Este robot tiene como principal objetivo mantener una interacción IHR natural, donde el robot pueda expresar emociones durante la comunicación. Para ello, en el diseño de Muecas se analizaron trabajos similares existente en la literatura, así como estudios relativos a la importancia de la apariencia final de un robot para conseguir una mayor

aceptación y empatía por parte de los usuarios con los que interactuará. Un resumen de robots similares y sus principales características fue presentado en el trabajo [Cid et al., 2014].

El diseño final de Muecas presenta 12 grados de libertad que se distribuyen en sus elementos principales, tal y como se recoge en la Figura 7.1, y que le permiten expresar emociones de forma similar a como lo hacen los humanos. A su vez, el robot está equipado con diversos elementos hardware que le permiten interactuar con el entorno y con otros usuarios. Por un lado, Muecas está equipado con un sistema de audio (micrófono y altavoz), sistema inercial (brújula, giroscopio y acelerómetro) así como de sistemas de adquisición de vídeo y profundidad (cámaras estéreo y sensor RGB-D). Por otro lado, los módulos de software se compone de diferentes subsistemas para el reconocimiento de la emoción humana y la imitación, y también para la generación y recepción de mensajes de audio. Todo ello integrado dentro del $framework$ RoboComp. Una descripción más detallada de cada uno de los elementos hardware del robot, así como del software integrado puede encontrarse en [Cid et al., 2014].



(a) (b)

Figura 7.1: a) Cabeza Robótica Muecas; b) Cadena Cinemática de la cabeza robótica Muecas, donde X_i representa cada elemento móvil y grado de libertad. (Figura obtenida de la publicación [Cid et al., 2014])

7.3. Sistema de imitación en Interacciones Humano-Robot

En esta sección se describen aquellos métodos implementados para la imitación de expresiones faciales y del lenguaje corporal del interlocutor por parte del robot. En particular, ambos sistemas de imitación trabajan actualmente sobre la cabeza robótica Muecas, si bien pueden ser exportados a cualquier robot con forma similar o que al menos sea capaz de generar emociones. Para este fin, el sistema completo de imitación con el que se contribuye en esta Tesis Doctoral presenta un modelo interno de representación de los estados emocionales y posiciones del usuario, de forma que la imitación puede ser utilizada con cualquier otra plataforma.

El sistema de imitación consta de dos subsistemas que trabajan en paralelo, tal y como aparece reflejado en la Figura 7.2. El primero intenta recrear la información emocional que puede transferir el usuario de forma directa, mediante el uso de sus expresiones faciales (enmarcado

en color rojo en la Figura 7.2). El segundo sistema incluye al anterior información relativa a la posición y orientación de la cabeza del usuario, de forma que aparte de la propia expresión emocional, considera los diferentes actuadores de la cabeza robótica y la malla $Candide$-3 para generar una lista de movimientos basados en el lenguaje natural, que pueden ser imitados por la cabeza robótica Muecas durante una interacción afectiva (color cian en la Figura 7.2).

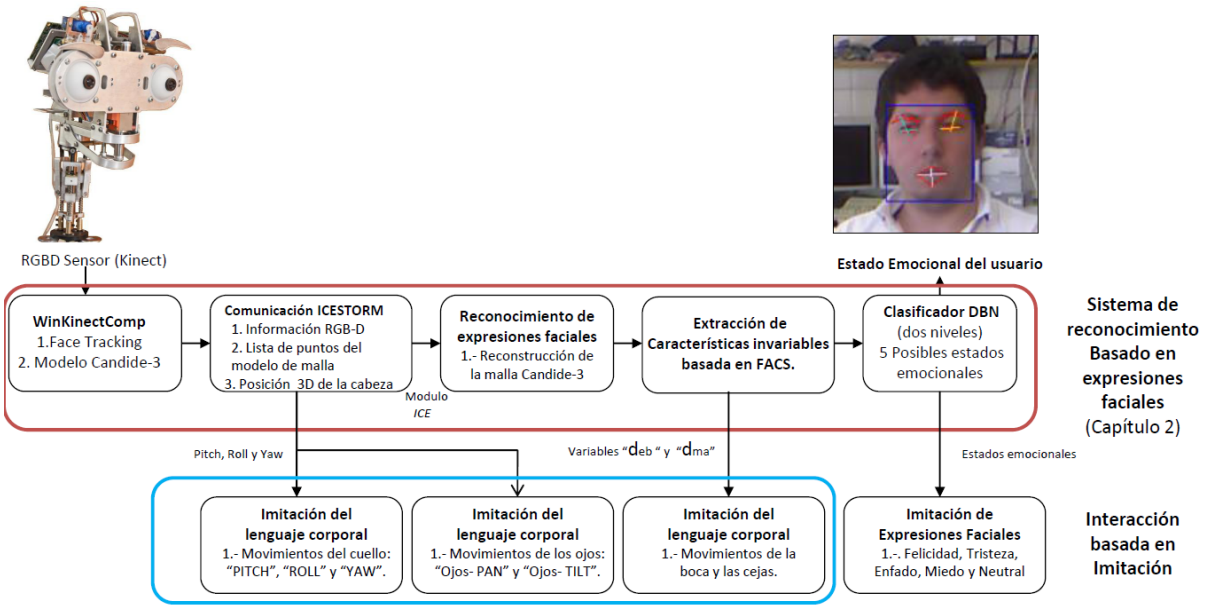


Figura 7.2: Visión general de los procesos de imitación de las expresiones faciales y del lenguaje corporal

A continuación, se introduce el modelo de representación utilizado en el sistema de imitación, así como la descripción de cada uno de los subsistemas de los que consta.

7.3.1. Modelo de representación del estado emocional del usuario

En los últimos años, el uso de modelos de representación interna por parte del robot ha sido adoptado por varias arquitecturas cognitivas. Estos sistemas se encargan de construir modelos selectivos del entorno de trabajo del robot, de los usuarios con los que interactúa o incluso del robot mismo. A través de estos modelos, el robot puede, antes de realizar una determinada acción, realizar simulaciones internas que permitan anticipar el resultado de una acción futura. Por ejemplo, un robot que navega en un entorno desconocido modela su entorno conforme va desplazándose por el mismo, de forma que en un futuro pueda calcular una ruta hacia un punto ya visitado de forma segura. En el caso de interacciones entre un humano y robot, disponer de un modelo de representación del estado emocional permite al robot adelantar el resultado de futuras acciones, si estas pueden llevarse a cabo sin colisiones o si el fin es el esperado.

Siguiendo este enfoque, la cabeza robótica Muecas está representada internamente por un modelo virtual, un avatar, que se compone no sólo del estado emocional y posición del usuario con el que se interactúa y que se pretende imitar, si no también de las características físicas del robot (la cabeza Muecas, en este caso) y sus limitaciones en el movimiento (modelo de malla, los límites de giro o la velocidad máxima, entre otros). De esta forma. la cabeza robótica Muecas es representada internamente según $M_{\{robot\}} =$

$\{(m_0, p_0), (m_1, p_1), ...(m_5, p_5), x, f, c\}$, donde m_i representa el posible estado emocional del usuario, $m_i \in \{Felicidad, Tristeza, Enfado, Miedo, Neutral\}$, p_i es la probabilidad de cada uno de los estado emocionales, $0 \leq p_i \leq 1$ y $\sum_{p_i} = 1$, x es el vector 6-D de la posición de la cabeza (incluyendo la orientación según los tres ejes), f es el conjunto de las características físicas de la cabeza (esto es, el modelo de malla) y c se define como las restricciones físicas. Cada vez que se detecta una nueva emoción del usuario se actualiza la representación interna del robot, M_{robot}.

Dado este modelo, cualquier otro robot partícipe de una IHR puede utilizar el sistema aquí presentado. Únicamente ha de ser capaz de incluir su propia representación interna dentro del modelo y las restricciones físicas asociadas a su diseño.

7.3.2. Sistema de imitación de expresiones faciales

Como se ha visto a lo largo de este trabajo, durante una IHR, la expresión facial se convierte en una de las fuentes de información emocional más robusta y precisa. La imitación de los gestos expresivos realizados por el usuario permiten a un agente robótico expresar información emocional en una comunicación no verbal, de forma intuitiva y amigable. Por este motivo, se ha desarrollado un método de imitación de expresiones faciales, implementado dentro del $framework$ RoboComp, a través del componente software $imitationComp$. Este componente, aparte de tener comunicación con los sistemas de reconocimiento de emociones descritos en los capítulos precedentes, controla los movimientos de cada uno de los elementos móviles de la cabeza robótica Muecas.

El proceso de imitación propuesto requiere solamente de la información relacionada con el estado emocional del interlocutor. En el caso que nos ocupa, esta información es facilitada por el sistema de reconocimiento de emociones basado en expresiones faciales, descrito en el Capítulo 3. Este sistema estaba basada en el uso de las Unidades de Acción, definidas en el *Facial Action Code System*, de forma que cada emoción tenía asociada un conjunto de AUs. El Cuadro 7.1 resume las AUs asociadas a cada uno de los estados emocionales con los que se trabaja en esta Tesis Doctoral. Como puede observarse en el mismo, salvo el estado Neutral, todos los demás estados presentan una combinación de, mínimo, tres AUs para detectar la emoción.

Cada uno de los estados emocionales tiene asociado, para esta cabeza robótica particular, un conjunto de movimientos asociados a los componentes hardware. La tabla 7.1 muestra cuáles son los movimientos de los distintos elementos móviles de Muecas para expresar una determinada emoción. El hecho de estar basado en el sistema FACS y en el uso de las AUs hace el sistema de imitación fácilmente reproducible en cualquier otro robot capaz de expresar emociones, simplemente modificando este $mapping$ desde las AUs a los elementos móviles.

La figura 7.3 resume el procedimiento descrito en esta sección. En primer lugar el usuario expresa una emoción a través de una expresión facial (Figura 7.3a). El sistema de reconocimiento de emociones detecta la misma y comunica el estado del usuario al robot. Antes de llevar a cabo el movimiento mecánico de la cabeza Muecas, como se muestra en la Figura 7.3c, el modelo virtual de la cabeza robótica es actualizado y se comprueban posibles limitaciones dada las restricciones del modelo (Figura 7.3b).

Estado Emocional	Unidades de Acción - AUs	Componentes móviles de *Muecas*
Neutral	-	-
Felicidad	AU6 - AU12 - AU25	Cejas - Párpados - Ojos - Boca
Tristeza	AU1 - AU4 - AU15	Cejas - Párpados - Ojos
Miedo	AU1 - AU4 - AU20 - AU25	Cejas- Párpados
Enojo	AU4 - AU23 - AU24	Cejas- Párpados

Cuadro 7.1: Unidades de Acción asociadas con cada uno de los estados emocionales. Estas AUs tienen su equivalente directo en los elementos móviles de la cabeza robótica Muecas.

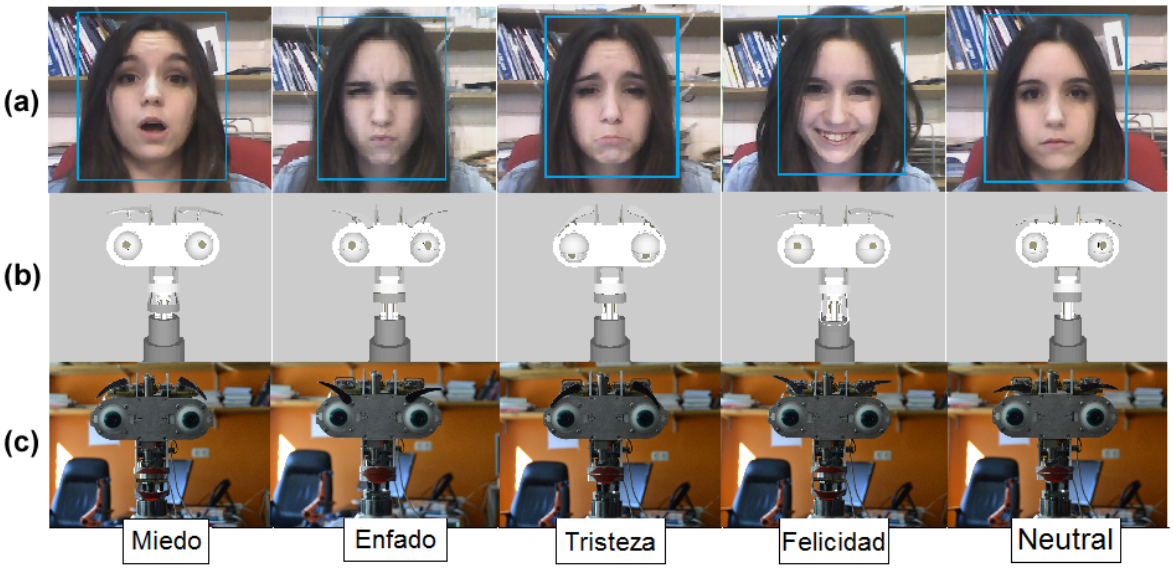


Figura 7.3: a) Imagen del usuario del sistema de reconocimiento de expresiones faciales del Capítulo 3; b) Representación Virtual del proceso de imitación de expresiones faciales; y c) Imitación de expresiones faciales por medio del agente robótico Muecas. (Figura obtenida de la publicación [Cid et al., 2014])

7.3.3. Imitación del lenguaje corporal

En este caso, la imitación del lenguaje corporal por medio del robot se realiza una vez se dispone de una estimación de la posición del usuario (posición de la cara, y la orientación de la misma en el espacio), y un seguimiento de los movimientos de los elementos de la cara. Por este motivo, se hace uso de la información del modelo de malla $Candide$-3 obtenida por medio del componente $WinKinectComp$, como se describe en el Capítulo 3. El uso de este componente permite, a la vez que reconstruir el modelo de malla del usuario, disponer de la información entregada por la librería *Kinect for Windows SDK* relacionada con la posición y orientación del mismo respecto al sensor RGB-D.

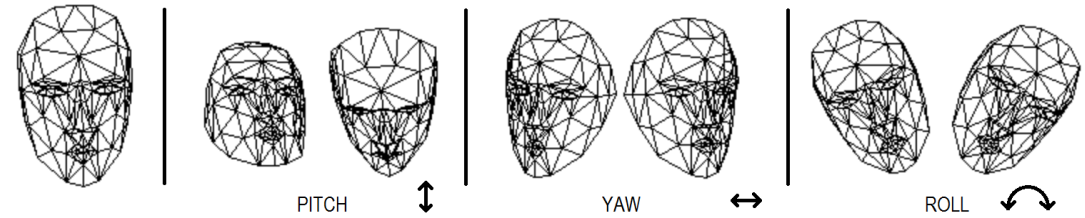
De igual forma que se hizo con el sistema anterior, cada uno de los movimientos detectados por el sistema de reconocimiento es transformado en un conjunto de AUs según el sistema FACS, como se muestra en el Cuadro 7.2. Estos AUs, genéricos, son posteriormente transformados en movimientos reales de la cabeza robótica Muecas, como se observa en la tabla. El

Movimiento	AUs	Movimientos de Muecas
Yaw	AU51-AU52	Cabeza girada a la izquierda - Cabeza girada a la derecha
Pitch	AU53-AU54	Cabeza inclinada hacia arriba - Cabeza inclinada hacia abajo
Roll	AU55-AU56	Cabeza inclinada hacia la izquierda - Cabeza inclinada hacia la derecha
Cejas	AU1-AU4	Cejas en una posición elevada - Cejas en una posición baja
Boca	AU24-AU25	Boca cerrada - Boca abierta
Ojos - pan	AU61-AU62	Ojos girados a la izquierda - Ojos girados a la derecha
Ojos - tilt	AU63-AU64	Ojos mirando hacia arriba - Ojos mirando hacia abajo

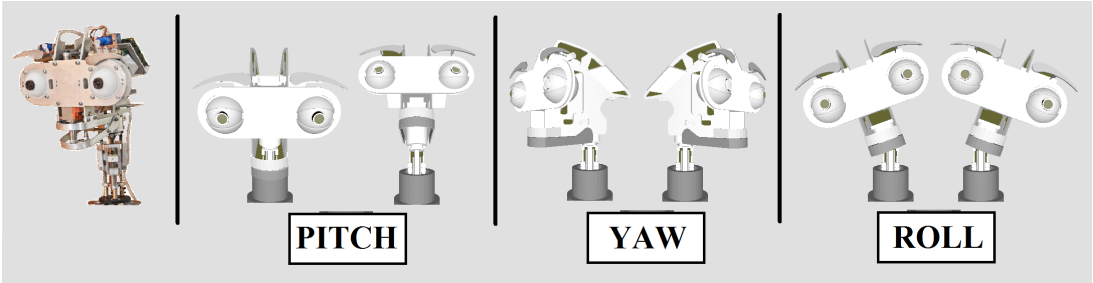
Cuadro 7.2: Movimientos de la cabeza robótica *Muecas*, y las AUs obtenidas por medio del algoritmo de seguimiento de la cara.

proceso de imitación se divide en tres fases, donde cada una de ellas realiza el cálculo de los movimientos de los distintos elementos de la cabeza del usuario. Por un lado, el componente $WinKinectComp$ transfiere directamente los movimientos de la cabeza de la persona, esto es, el $Pitch$, $Roll$ y el Yaw, lo que permite calcular directamente la posición de los motores del agente robótico para imitarlos. En la Figura 7.4 se observan los grados de libertad de la cabeza robótica, obtenidas del modelo de malla $Candide$-3, y en la Figura 7.5 se ilustra la estructura mecánica y los motores encargados de estos movimientos. Por otro lado, la segunda parte está relacionada con el movimiento de las cejas y la boca del usuario, por medio de la información de las distancias euclídeas de los nodos de la malla $Candice$-3: d_{eb} y d_{ma} (Ver Figuras 7.6b y 7.6d), las cuales se adquirieron desde el proceso de extracción de características faciales descrito en el Capítulo 3. En la Figura 7.6 se ilustran los motores de la cabeza robótica Muecas encargados del movimiento de las cejas y boca, y los nodos de la malla utilizados para estimar las posiciones de estos motores en cada movimiento. Finalmente, la última fase se refiere al seguimiento de la posición del usuario durante la interacción. Si bien no es exactamente una imitación en sí misma, es importante destacar la necesidad de llevar a cabo esta acción no solo para empatizar durante la comunicación, si no también para mantener en todo momento información RGB del interlocutor, considerando que existen otros procesos utilizando la información visual de las cámaras RGB localizadas en el globo ocular de Muecas. Conocido los movimientos del $Pitch$ y el Yaw del usuario, el sistema ajusta la posición 3D de los ojos de acuerdo con esta información. En la Figura 7.7 se ilustra los movimientos $Tilt$ y el Pan de los ojos en el agente robótico durante el seguimiento de la posición del usuario.

Finalmente, se presenta a continuación una serie de experimentos encaminados a evaluar el sistema de reconocimiento e imitación descrito. La metodología seguida consiste en una secuencia de 120 repeticiones de los movimientos $Pitch$, Yaw y $Roll$ por parte de un grupo de veinte usuarios, de diferente género, edad y rasgos faciales. De la misma forma, se realizaron la misma cantidad de movimientos de las cejas y los ojos, que consistían en su apertura y cierre de manera exagerada. Cada uno de estos movimientos era imitado por la la cabeza robótica Muecas. Un observador experto evaluaba la tasa de éxito del experimento, considerando como acierto aquellos movimientos realizados por el robot que se correspondían con su equivalente humano. En el cuadro 10.4 se muestran los resultados de las estimación e imitación del lenguaje corporal del usuario, P_{users}. Como se observa en la tabla, los movimientos asociados al $Roll$, $Pitch$ y Yaw presentan los mejores resultados en cuanto a acierto, superiores todos al noventa



(a)



(b)

Figura 7.4: Grados de libertad; a) $Pitch$, Yaw y $Roll$ del modelo $Candide$-3; y b) $Pitch$, Yaw y $Roll$ de la cabeza Robótica Muecas

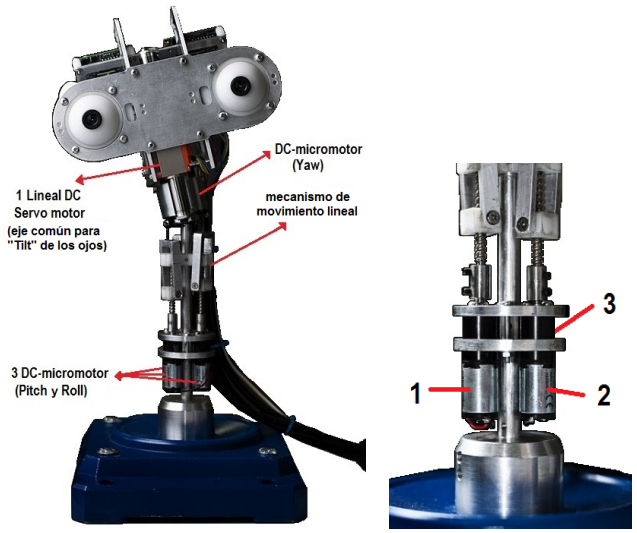
Test	Porcentaje de correcta estimación e imitación de los movimientos, P_{users}
Pitch	93 %
Roll	98 %
Yaw	95 %
Mov. de las cejas	82 %
Mov. de la boca	63 %

Cuadro 7.3: Resultados del sistema de imitación de movimientos basado en Unidades de Acción AUs, realizada por medio de la cabeza robótica Muecas (Cuadro obtenido parcialmente de la publicación [Cid et al., 2014]).

por ciento de acierto. En cambio, la imitación de los movimientos de la boca del usuario tiene un porcentaje de acierto cercano al sesenta y cinco por ciento, un valor esperado dada las limitaciones en la extracción del modelo $Candide$-3 bajo ciertas condiciones (analizadas en el Capítulo 3) y su fuerte dependencia al mismo.

7.4. Interacción basada en la voz

Igual que la imitación del lenguaje corporal o la propia emoción humana, es importante dotar al robot con la capacidad de transmitir y recibir información durante la comunicación. El desarrollo de sistemas que permiten una interacción humano-robot cada vez más fluida y expresiva, a la vez que empática, es la base del éxito para el desarrollo de la robótica social. Por este motivo, es importante que los robots puedan transmitir, recibir y llegado el caso, retroalimentar información específica acerca del usuario por medio de una interacción verbal en



(a) (b)

Figura 7.5: Cabeza Robótica Muecas; a) Imagen de los motores asociados a los movimientos del cuello; y b) Imagen de los motores encargados de los movimientos $Pitch$ y $Roll$ (Figura obtenida de la publicación [Cid et al., 2014]).

tiempo real.

Existen diferentes sistemas que permiten transmitir y reconocer mensajes por medio de la voz en la literatura. Estos métodos se conocen como sistemas de reconocimiento del habla ASR (del inglés *Automatic Speech Recognition*) y los sistemas de generación de voz a partir de texto ***TTS*** (*Text-to-Speech*). El primero de ellos permite recibir información verbal al robot, traduciendo a cadena de texto el mensaje percibido por los sensores de audio (micrófonos). El segundo dota al robot con la capacidad de transmitir mensajes, de forma que a partir de texto se envía un mensaje sonoro a través del sensor (altavoz).

Desde el punto de vista de la interacción afectiva, sin embargo, sólo los sistemas TTS presentan la posibilidad de manipulación y de esta forma adaptarse al contexto de la comunicación (expresando diferentes tipos de estados emocionales por medio de la voz, por ejemplo). Por un lado, se puede manipular la prosodia del sistema TTS para generar contenido emocional a través de frases controladas previamente [J.Cahn, 1990]. Por otro lado, dentro del propio lenguaje natural, el proceso del habla se presenta en sí mismo como una percepción multimodal, es decir, no sólo transmitimos el mensaje verbal, sino que también éste viene acompañado de gesticulaciones y expresiones faciales, lo que se conoce como el efecto $McGurck$ [Chen and Rao, 1998]. Este efecto realza la importancia de un método o algoritmo que genere movimientos en el robot durante una interacción afectiva, tanto de la propia boca conforme se habla, como de movimientos del cuello y rostro.

En esta sección se describen los algoritmos TTS y ASR usados en el robot Muecas, que forman parte del $framework$ RoboComp. A su vez, se describe el algoritmo de sincronización desarrollado en esta Tesis Doctoral, y con el que se contribuye en el actual estado del arte. La Figura 7.8 ilustra una visión general del sistema. A continuación, se describirán en detalle cada uno de los módulos del mismo.

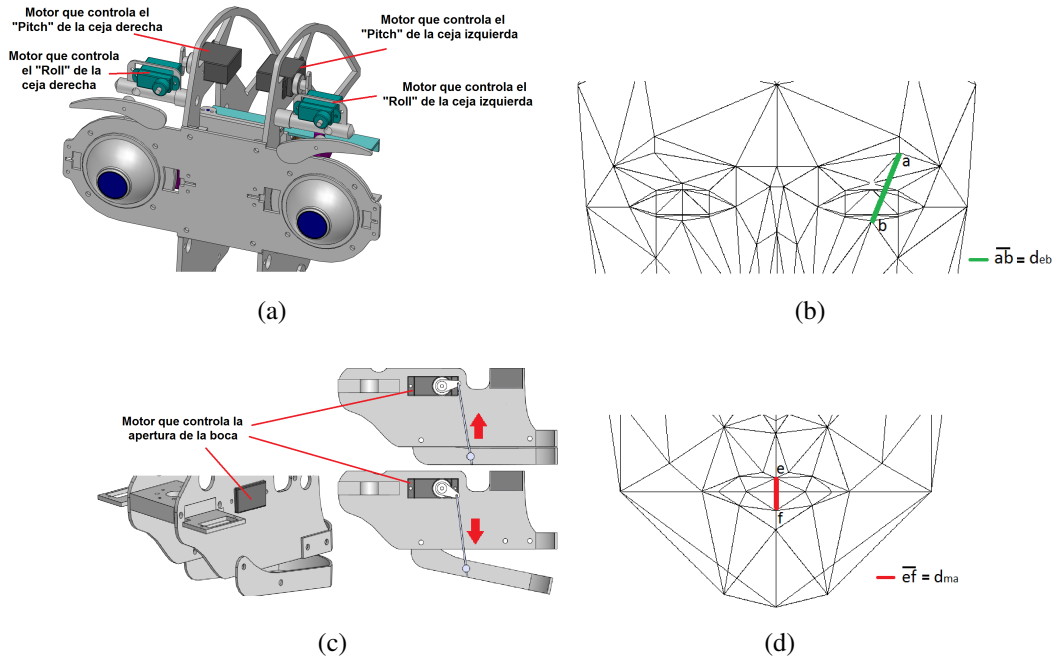


Figura 7.6: movimiento de la cabeza robótica Muecas; a) Movimiento de las cejas; y b) Apertura de la boca. (Figura obtenida de la publicación [Cid et al., 2014]).

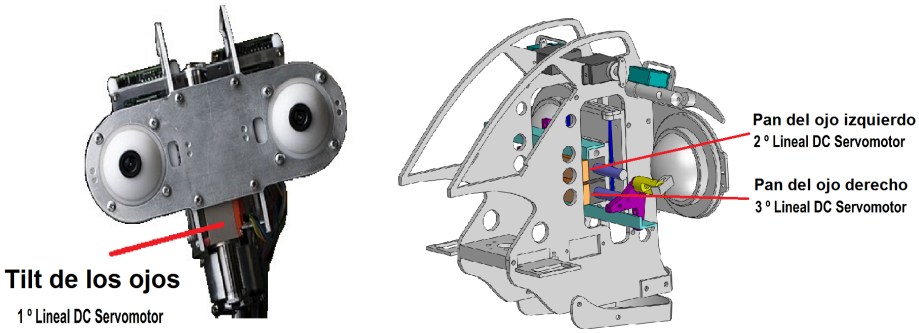
7.4.1. Sistemas ASR

Los sistemas de reconocimiento automático de voz (ASR) [Anderson and Kewley-Port, 1995], realizan una función de reconocimiento del contenido del mensaje transmitido por la voz de un interlocutor humano. El uso de este sistema dentro de la robótica social se presenta como un proceso fundamental para interactuar e intercambiar información con los usuarios de forma no invasiva, sin necesidad de utilizar la información visual. Dentro de esta Tesis doctoral, el reconocimiento del contenido del mensaje permite obtener una realimentación de información del usuario de manera descriptiva y objetiva, no subjetiva como lo era la información emocional. Así, disponer de un sistema ASR dentro de la arquitectura cognitiva de Muecas permite usar la información del usuario en algoritmos de aprendizaje o interacción afectiva, como se verá en capítulos posteriores.

El funcionamiento de los sistemas ASR requiere la adquisición externa de la información acústica, directamente a través del micrófono con el que el robot está equipado, o bien, como es el caso más normal, por medio de un fichero de audio con características específicas. En el caso del trabajo presentado, se utiliza la librería SoX para las fases previas de adquisición y procesado de la señal de audio original. Esta librería ya fue utilizada para el reconocimiento de emociones basado en voz, descrito en el Capítulo 4, y se presenta como Apéndice A.3 en este documento. El componente $speechGoogleComp$ es el encargado de realizar la implementación en C++ del sistema ASR dentro del $framework$ RoboComp.

En el caso del sistema ASR desarrollado por $Google$, que es el sistema usado en la arquitectura de Muecas, el audio generado por el interlocutor es presentado mediante el formato $.flac$ con una frecuencia de muestreo F_s de 16 $Khz.$, por medio de la siguiente linea de comando:

wget -U 'Mozilla/5.0' –post-file outvad.flac –header="Content-Type: audio/x-flac; rate=16000" -O - 'http://www.google.com/speech-api/v2/recognize?output=json&lang=es_ES&key=KEY' > speech.json



(a) (b)

Figura 7.7: Movimientos de la cabeza robótica Muecas; a) $Tilt$ común de los ojos; y b) Pan individual de cada globo ocular. (Figura obtenida de la publicación [Cid et al., 2014]).

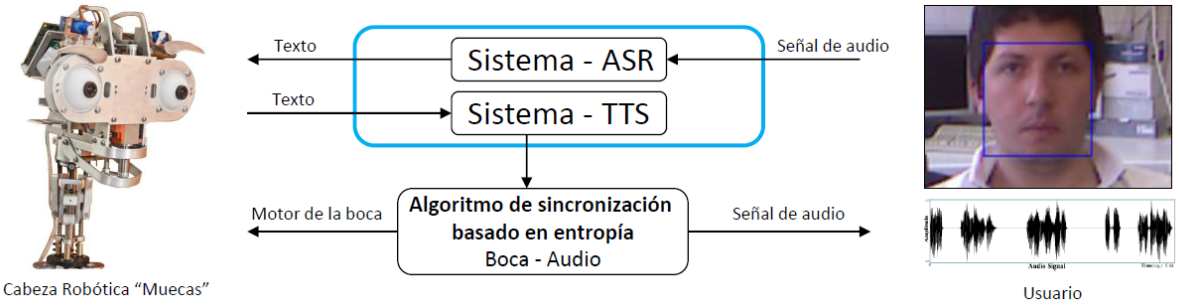


Figura 7.8: Visión general de los procesos de interacción basados en la voz

A continuación, se analiza la información del fichero de salida del sistema, en este caso, $speech.json$, el cual contiene las diferentes opciones con respecto al contenido del mensaje en el audio y una probabilidad de exactitud asociada a la primera opción.

{**"result":[**{**"alternative":[**{**"transcript":"opción1","confidence":0.94677681**},
{**"transcript":"opción2"**}],**"final":true**}],**"result_index":0**}

7.4.2. Sistemas TTS

Los sistemas de $Text\text{-}To\text{-}Speech$ (TTS) [Moberg, 2007] son programas de síntesis de voz que permiten convertir un texto escrito, en un mensaje verbal por medio de una voz sintética. Existen un gran número de TTS, comerciales y gratuitos con algún tipo de licencia, y una gran versatilidad en cuanto a sus principales características se refiere. En esta Tesis Doctoral, la capacidad de generar audio, simulando la voz del robot, facilita la interacción directa con el usuario durante la interacción, formulando preguntas o respondiendo según la situación. Además, se hace uso del sistema TTS para añadir información emocional a la comunicación, modificando parámetros característicos del habla del robot para así mostrar estados emocionales concretos.

En esta sección se introducen dos de los algoritmos TTS utilizados en este trabajo. Por un lado, se describe el TTS de $Google$, una versión gratuita multilenguaje y que permite, con una conexión a internet, la generación de voz por parte del robot de una forma rápida y con gran aceptación por parte de usuarios no entrenados. A su vez, se presenta el TTS de $Verbio$,

un software comercial muy utilizado para diferentes aplicaciones y que entre otras cualidades, facilita la modificación de parámetros internos del software para expresar emociones. Estos dos algoritmos han sido probados con diferentes usuarios, evaluando su aceptación según diversos parámetros [Cid et al., 2011].

Ambos sistemas TTS han sido integrados dentro de la cabeza robótica Muecas, como dos componentes del $framework$ RoboComp. A continuación se describen el funcionamiento de los mismos.

1. Google TTS

   El sistema TTS desarrollado por $Google$ presenta una gran naturalidad con múltiples idiomas. Para su funcionamiento precisa del servicio de internet de $Google$ a tal efecto, lo que puede llegar a suponer ciertos problemas en algunas situaciones. La latencia, con una conexión normal, es más que aceptable, de forma que su uso correcto está garantizado para interacciones reales. El componente $speechGoogleComp$ implementa la llamada al algoritmo TTS $Google$ dentro del $framework$ RoboComp, ejecutando la siguiente línea de comando:

   wget -q -U Mozilla -O audio.mp3
   "http://translate.google.com/translate_tts?ie=UTF-8&tl=es-ES&q=Agregar+Texto+Aqui"

   Como se observa en la expresión anterior, en la línea se especifica el idioma por medio del parámetro $es - ES$ (Español) y se incorpora el texto a convertir en un mensaje de audio por medio de una voz sintética. Además, esta línea genera el fichero de salida de audio $audio.mp3$, que será aquel que se reproduzca a través de los altavoces del robot. Para ello, en el sistema implementado, se realiza una llamada al reproductor $Mplayer$ (Ver. Apéndice A.5).

2. Verbio TTS

   El sistema TTS de $Verbio$ [Verbio Technologies, 2014], es un software comercial para la generación de audio multilenguaje, cuya principal ventaja respecto a otros sistemas similares, aparte de la calidad y naturalidad de la voz sintetizada, es la posibilidad de manipular la salida del audio por medio de cambios en los elementos de la prosodia. Así, $Verbio$ permite el uso de etiquetas que cambian algunos de estos parámetros, como el énfasis, la intensidad, el tono, la velocidad o la energía, entre otros, lo que facilita la síntesis de voz con carga emocional. A continuación aparecen dos frases y su correspondiente llamada al TTS, donde se incluyen modificaciones de elementos de la prosodia:

   no estoy seguro de esto, ¡pero! se ve bastante bien por aquí

   $"<$ prosody pitch=\" $+ 92\,\% \setminus " > no\ estoy\ seguro\ de\ esto < /prosody >$ $,< break\ strength = \setminus "400ms \setminus " >< prosody\ rate = \setminus "x - slow \setminus " >< prosody\ volume = \setminus "loud \setminus " >!`Pero! < /prosody >,< emphasis\ level = \setminus "strong \setminus " > se\ ve\ bastante\ bien\ por\ aquí. < /emphasis >< /prosody >"$

   No logro entenderte, puedes hablar mucho más lento?

”no logro entenderte <break strength=\”800ms\” >, ?‘$puedes$ < $prosodyrate$ = \”$x-slow$\” > $hablar$ < $breakstrength$ = \”400ms\” > $mucho$ < $breakstrength$ = \”500ms\” > mas < $breakstrength$ = \”500ms\” > $lento$?. < /$prosody$ >″

Este segundo sistema ha sido incluido dentro del $framework$ RoboComp, implementándose para tal efecto el componente $SpeechVerbioComp$. Este componente recibe la información del texto y la información emocional (si la hubiera), y genera un archivo de salida de audio $salida.ogg$ que será posteriormente reproducido por el robot, al igual que ocurría con el TTS de $Google$.

7.4.3. Lenguaje corporal en el uso de la voz

Dado que la interacción humano-robot requiere de una comunicación cercana, cada vez se hace más importante el reconocimiento e imitación de los diferentes elementos que conforma el lenguaje natural. Así, incluso durante una comunicación verbal, donde gran parte del contenido del mensaje va en la voz, el uso eficiente de gestos, movimientos labiales o la propia expresividad del rostro del robot, permite solucionar incertidumbres y afectar la percepción del usuario. En particular, la percepción del habla durante una comunicación entre humanos es multimodal, y depende fuertemente de la información visual a partir del movimiento realizado por la boca con cada una de las palabras. Este fenómeno es conocido en la literatura como el efecto $McGurk$ [Chen and Rao, 1998], y la base del sistema presentado en esta sección.

El efecto $McGurk$ dentro de una IHR insta a que los robots sociales estén dotados de algoritmos de sincronización entre el movimiento de la boca y la voz sintetizada, de forma que se permita mejorar el nivel de atención, naturalidad y cercanía, reduciendo la brecha comunicacional entre humanos y robot [Oh et al., 2010][Hara et al., 1997]. Por este motivo, en esta Tesis Doctoral se presenta un algoritmo capaz de sincronizar el movimiento de una boca robótica con el audio sintético generado por un sistema TTS. Este método se basa en el cálculo de la entropía de la señal de audio, a partir del fichero de salida del sistema TTS, y de forma que la apertura de la boca robótica se corresponde directamente con este valor. El método propuesto es independiente del sistema TTS, pues utiliza directamente la señal generada, sin ruido.

El algoritmo de sincronización consta de tres fases consecutivas. En primer lugar, se realiza un procesado de la señal de audio original, dividiendo la señal completa en ventanas del mismo tamaño. A continuación, se calcula el valor de la entropía en cada una de estas ventanas y finalmente, se realiza la sincronización en sí con la boca del robot. Sea una señal de audio de entrada $X(t)$, procedente del sistema TTS, con una frecuencia de muestreo F_s de 16 Khz y duración T, $X(t) = [0, ..., F_s \cdot T - 1]$, el proceso completo se describe a continuación:

1. **Procesamiento de la señal de audio:** este procesado consta de dos pasos principales. Por un lado, esta etapa realiza el cálculo del valor absoluto de la señal, $V(i) = |X(t)|$, para, a continuación generar ventanas o tramas sobre el vector $V(i)$. El tamaño de estas tramas (N_{trama}) corresponde al número de muestras calculadas por medio de la frecuencia de muestreo F_s y el periodo estimado de cada trama (t_{trama}), según la Ecuación 7.1. Este periodo se fija a una décima de segundo, al ser éste el tiempo medio aproximado de duración de un fonema y el limite del tiempo de respuesta de los motores de la boca del robot Muecas. La Figura 7.9 ilustra los diferentes pasos de esta etapa. En 7.9a, se muestra la imagen original, que es procesada para calcular el valor absoluto 7.9b y posteriormente

enventanada 7.9c. Las muestras asociadas a cada ventana son posteriormente utilizadas para calcular la entropía.

$$N_{trama} = F_s \cdot t_{trama} \tag{7.1}$$

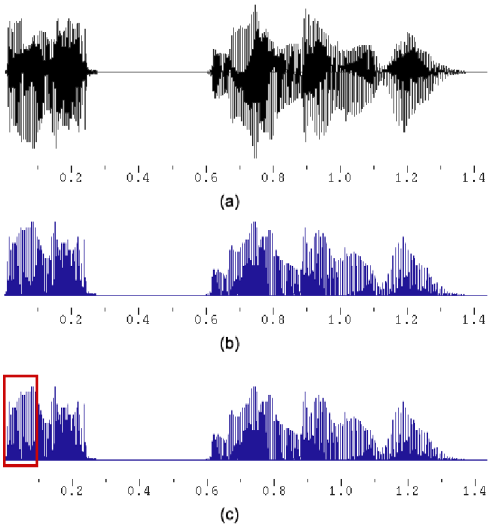


Figura 7.9: Representación gráfica del procesado de la señal en la primera etapa del sistema; a) Imagen de la señal de audio original; b) Valor absoluto de la señal de audio; y c) Imagen de las ventanas de tiempo utilizadas en este estudio. (Figura obtenida de la publicación [Cid et al., 2012]).

2. **Cuantificación de la señal de audio:** esta etapa tiene como objetivo el cálculo del nivel de entropía de la señal de audio obtenida de la fase previa. La entropía puede definirse como la cantidad de información existente en una señal medida en $bits$. Sea $V_N(i)$ las muestras de la señal de audio original, sin ruido, de una ventana de tamaño N_{trama}, el valor de entropía para esta ventana viene dada por la Ecuación 7.2, donde $P(v_i)$ es la probabilidad de encontrar la medida en cada trama, y x_i es la n_{th} medida para cada muestra.

$$H(X) = -\sum_{i=1}^{N_{trama}} P(v_i) \log_2 P(v_i) \tag{7.2}$$

3. **Sincronización:** en esta etapa final se realiza la sincronización de los movimientos de la boca con respecto al sonido generado por sistema TTS. La posición enviada al motor del robot es directamente proporcional al nivel de entropía calculada en la etapa anterior, incluyendo un factor de corrección dependiente de los valores de la prosodia del TTS [Cid et al., 2012].

$$Ángulo \propto Entropía \tag{7.3}$$

El funcionamiento de este algoritmo sólo produce un leve retardo en la generación y procesamiento del fichero de salida, inferior al tiempo de comunicación entre el ordenador

y los motores del robot. Por este motivo, se incluye una función que considera este retardo en la comunicación antes de reproducir el sonido por medio de $Mplayer$ (Ver. Apéndice A.5).

7.4.3.1. Evaluación del algoritmo de sincronización

Los robots sociales, en IHR complejas, requieren de una serie de características que permitan mejorar la comunicación por medio de elementos del lenguaje natural. Para evaluar las posibles mejoras durante una interacción real por medio de un algoritmo de sincronización, es necesario cuantificar aspectos tan subjetivos como la atención, el atractivo y la comprensión en la comunicación. Estos aspectos comúnmente son evaluados mediante métodos basados en encuestas, adquiriendo y analizando la percepción y opinión del usuario.

Por este motivo, se decidió evaluar el método propuesto por medio de las impresiones del usuario, de acuerdo a varias consideraciones: i) Cómo influye el algoritmo de sincronización en la percepción directa del habla en una cabeza robótica o simulada virtualmente; ii) cómo se ve afectado el rendimiento y la influencia de una voz sintética generada por medio de diferentes sistemas TTS en la percepción de un usuario, si se utiliza o no se utiliza el algoritmo de sincronización; iii) qué impacto y diferencias presentan algunos algoritmos de sincronización con respecto al descrito en este capítulo; y iv) cómo el lenguaje corporal mejora o afecta la comunicación de conceptos como las emociones en interacciones con humanos. Así, cada uno de estos tópicos fue evaluado mediante diferentes estudios comparativos, que se describen en las siguientes secciones.

En relación a las condiciones para cada estudio, se utilizó el componente $mouthComp$ del $framework$ RoboComp, que implementa el algoritmo de sincronización y se comunica con la cabeza robótica Muecas. Con respecto a los participantes, se seleccionaron 15 personas mediante entrevista personal, con diferentes niveles de conocimiento acerca de la robótica. Por sectores, cinco personas presentaban un nivel alto de conocimiento en robots, cuatro personas poseían un nivel moderado o intermedio y seis personas demostraron un nivel mínimo o bajo de conocimiento sobre la robótica.

A continuación, a partir de los diferentes aspectos a evaluar para cada estudio comparativo, se definen las siguientes preguntas de la encuesta:

- **A)** *Comportamiento natural* - ¿Parece la boca moverse de forma natural?
- **B)** *Expresividad* - ¿La boca le resulta expresiva?
- **C)** *Capacidad para atraer y mantener la atención* - ¿La boca captura su atención?
- **D)** *La comprensión del mensaje* - ¿La boca directa o indirectamente, ayuda a entender el mensaje?

Para evaluar estas preguntas, cada respuesta del usuario está condicionada a una escala lineal de 1–5, donde 1 es el nivel más bajo, y 5 el nivel más alto. También, se presentan una gran cantidad de sentencias predeterminadas que son generadas por medio de un sistema TTS, siendo cada sentencia diferente a la anterior pero manteniendo un contexto uniforme en la interacción.

Finalmente, es importante mencionar que los experimentos relacionados con la cabeza robótica Muecas fueron realizados a continuación de la fecha de publicación de [Cid et al., 2012]. Por lo cual, estos resultados son datos analizados únicamente en esta Tesis Doctoral.

7.4.3.2. Estudio comparativo de diferentes bocas robóticas

Inicialmente se estudia la influencia del algoritmo de sincronización propuesto aplicado a diferentes robots. Así, se ha implementado el método en diferentes agentes robóticos diseñados para la interacción humano-robot (Figura 7.10). Estos robots pueden ser diferenciados sean agentes virtuales o físicos. En el primer caso, se refiere a los modelos animados del robot Ursus (Figura 7.10a), un robot diseñado por el Laboratorio de Robótica y Visión Artificial RoboLab, con forma de oso, y usado en terapias de rehabilitación, y un modelo de un robot simple que posee una boca robótica implementada por medio de una matriz de LEDs y cuyo diseño está basado en [Lee et al., 2009] (Figura 7.10b). El diseño de ambos modelos fue realizado por medio de *3D Studio Max*. En el caso del modelo virtual de Ursus, se incluyen todos los grados de libertad del robot real, de forma que pueda imitar completamente la cadena cinemática durante la sincronización. En el caso de la boca implementada con una matriz de LEDs (21x3), ésta se ilumina desde dentro hacia los extremos de acuerdo al nivel de apertura entregado por el algoritmo. En cuanto a los agentes físicos se encuentra el robot social Ursus [Mejías et al., 2013] y la cabeza robótica Muecas [Cid et al., 2014], que son ilustrados en la Figura 7.10c y 7.10d, respectivamente.

(a) (b) (c) (d)

Figura 7.10: Bocas utilizadas en el estudio comparativo de la Sección 7.4.3.2; a) Modelo animado del Robot Ursus; b) Modelo animado de un robot con una boca basada en LEDs; c) Robot Ursus; y d) Cabeza robótica Muecas. (Figuras obtenidas parcialmente de la publicación [Cid et al., 2012])

La evaluación se lleva a cabo por los usuarios, siguiendo el método de encuesta descrito anteriormente, donde los participantes, sin interferencia externa evalúan los diferentes aspectos en una interacción con un agente robótico. Este proceso comienza con un primer escenario, donde los usuarios están frente a una pantalla que intenta recrear las cabezas de los agentes virtuales a un tamaño cercano al real. Mientras, en el segundo escenario, los usuarios están frente a frente a la misma altura con los agentes físicos. Las comunicaciones e interacciones solo fueron secuencias predeterminadas de mensajes generados por el sistema TTS.

En el Cuadro 7.4 se muestran los resultados derivados de la percepción de los usuarios con respecto al funcionamiento del algoritmo. Se aprecia en el mismo que las bocas robóticas de los agentes físicos presentan los mejores resultados en general, pero principalmente en aspectos como la naturalidad y la capacidad de atraer la atención, debido a que al humano percibe directamente el movimiento de los elementos móviles. En cambio, los agentes virtuales presentan buenos resultados con respecto a la expresividad y la respuesta en la comprensión del

Bocas	Preguntas			
	A	B	C	D
Boca robótica animada de un modelo virtual	67 %	68 %	69 %	78 %
Boca basada en LEDs de un modelo Virtual	42 %	46 %	59 %	73 %
Boca robótica del agente Ursus	74 %	66 %	74 %	64 %
Boca robótica del agente Muecas	81 %	73 %	75 %	66 %

Cuadro 7.4: Comparativa con respecto al uso del algoritmo de sincronización con diferentes bocas robóticas. (Cuadro obtenido parcialmente de la publicación [Cid et al., 2012])

mensaje verbal sintético. No obstante, fue la cabeza robótica Muecas quien demostró el mejor rendimiento en casi todos los aspectos de la evaluación gracias a su diseño antropomórfico.

7.4.3.3. Estudio comparativo de los diferentes sistemas TTS

Este estudio comparativo tiene como objetivo cuantificar los cambios y efectos en la percepción del usuario causados por el uso del algoritmo de sincronización con respecto a diferentes sistemas TTS del mercado. Durante la evaluación se utilizaran cinco sistemas TTS, como son $Verbio$ (*Verbio Technology*), $Festival$ (*Univ. of Edimburg*), $Acapela$ (*Group Acapela*), $Ivona$ (*Ivona Software*), y $GoogleTTS$ (*Google*). Para la evaluación propuesta en esta sección, cada uno de estos sistemas generaba un fichero de salida de audio con su respectiva frecuencia de muestreo, como se muestra en el Cuadro 7.5. No obstante, el uso del algoritmo de sincronización descrito en esta Tesis Doctoral, especifica que esta adaptado a sistemas TTS que generen un fichero de salida con una frecuencia de 16 Khz, por lo que previamente fue necesario un remuestreo para ajustar a esta frecuencia las cadenas de audio a transmitir.

Sistema TTS	Verbio	Festival	Ivona	Acapela	GoogleTTS
Frecuencia de muestreo F_s	16 Khz	44 Khz	22 Khz	22 Khz	16 Khz

Cuadro 7.5: Sistemas TTS ($Text\text{-}to\text{-}Speech$) utilizados en el estudio comparativo de la Sección 7.4.3.3 (Cuadro obtenido parcialmente de la publicación [Cid et al., 2012])

La primera parte de esta evaluación es realizada como un estudio comparativo, que tiene como objetivo ser utilizado como base para estimar si el uso del algoritmo de sincronización para bocas robóticas afecta positiva o negativamente la percepción de los usuarios en una interacción. Esto se debe a que esta parte de la evaluación global sólo comprueba, para cada sistema TTS, los aspectos básicos descritos en la Sección 7.4.3.1. En el Cuadro 7.6 se muestran los resultados de la evaluación para cada sistema TTS, que demuestran como $Acapela$ y $GoogleTTS$ presentan los mejores resultados en gran parte de los aspectos a considerar, tales como la naturalidad o la capacidad para atraer la atención.

En segundo lugar, se trata de evaluar cada uno de los algoritmos TTS con el sistema de sincronización descrito en esta sección y utilizando un único robot, en este caso Muecas. En el Cuadro 7.7 se ilustran los resultados de estos experimentos, donde se comprueba como nuevamente los sistemas $GoogleTTS$ y $Acapela$ presentan los mejores resultados en todos los índices

TTS	Preguntas			
	A	B	C	D
Verbio	52 %	46 %	52 %	72 %
Festival	60 %	56 %	52 %	80 %
Acapela	68 %	72 %	68 %	72 %
Ivona	64 %	60 %	56 %	68 %
GoogleTTS	76 %	64 %	73 %	85 %

Cuadro 7.6: Comparativa de los diferentes sistemas TTS (Cuadro obtenido parcialmente de la publicación [Cid et al., 2012])

TTS/Muecas	Preguntas			
	A	B	C	D
Verbio	80 %	72 %	78 %	66 %
Festival	58 %	49 %	65 %	63 %
Acapela	84 %	80 %	85 %	78 %
Ivona	72 %	62 %	74 %	62 %
GoogleTTS	92 %	78 %	82 %	80 %

Cuadro 7.7: Comparativa de los diferentes sistemas TTS utilizados en la evaluación del algoritmo de sincronización con la cabeza robótica Muecas.

de evaluación. Por su parte, tanto $Festival$ como $Ivona$ consiguen los peores resultados, principalmente en aspectos como la expresividad o la comprensión del mensaje.

Finalmente, al comparar los resultados de la primera y segunda parte de esta evaluación (Cuadro 7.6 y Cuadro 7.7), se determinó que $GoogleTTS$ y $Acapela$ presentan los mejores resultados independiente del uso del algoritmo de sincronización, principalmente, en aspectos como la naturalidad y la capacidad para atraer la atención. Mientras, el sistema $Festival$ presenta bajos resultados en ambos estudios. Como conclusión final se puede demostrar fácilmente cómo el uso del algoritmo de sincronización a través de un agente robótico antropomórfico, en este caso Muecas, permite mejorar sustancialmente la percepción de la información por parte del usuario.

7.4.3.4. Estudio comparativo de los diferentes algoritmos de sincronización

Una parte esencial dentro de un estudio comparativo es la evaluación del algoritmo propuesto con respecto a otros métodos existentes en la literatura. El primero de los métodos se basa en analizar la señal de audio en busca de un instante de tiempo que contenga voz humana, para generar un nivel aleatorio de apertura de la boca. El segundo, controla los niveles de apertura de la boca a través de un pulso binario que se activa cuando detecta voz humana, alternando entonces entre abierta (1) y cerrada (0) mientras dure la conversación. Cuando no existe voz, esto es, se detecta silencio o ruido ambiente, se mantiene cerrada (0). La frecuencia entre pulso a 1 y 0 durante la conversación lo fija el tiempo de respuesta de los motores de la boca, según la expresión ($5\ x\ T_{Respuesta}$). La evaluación y comparación de estos métodos se lleva a cabo de la

Algoritmos de sincronización	Preguntas			
	A	B	C	D
Entropía	84 %	78 %	80 %	72 %
Aleatorio	58 %	40 %	54 %	42 %
Binario	60 %	52 %	45 %	56 %

Cuadro 7.8: Comparativa entre diferentes algoritmos de sincronización

misma forma que fue explicada en la Sección 7.4.3.1. El Cuadro 7.8 resume los resultados obtenidos en este experimento. De su análisis se desprende cómo el algoritmo de sincronización propuesto en esta Tesis presenta una mejor experiencia para el usuario a través de la percepción de aspectos relevantes del dialogo en una comunicación, tales como la naturalidad o la expresividad.

Por último, los resultados del estudio comparativo fueron descritos en el Cuadro 7.8, que demuestran como este tipo de algoritmo de sincronización propuesto en esta Tesis, representa una mejor experiencia para el usuario a través de la percepción de aspectos relevantes del dialogo en una comunicación, tales como la naturalidad o la expresividad.

7.4.3.5. Estudio comparativo del uso del lenguaje corporal

Este último estudio comparativo evalúa el impacto del lenguaje corporal en una comunicación humano-robot. Para lograr este objetivo se modifico el mensaje verbal sintético generado por medio del sistema TTS $Verbio$ para simular información emocional, gracias a cambios específicos en los parámetros de la prosodia, especialmente el $Pitch$ y el énfasis, entre otros. Dentro de esta evaluación, los cambios en la prosodia se realizaron a nivel de palabras y en algunos casos a nivel de frases, sin afectar al algoritmo de sincronización que trabaja sólo con el fichero de audio de salida del sistema TTS.

El proceso de evaluación se divide en dos experimentos que siguen el sistema de encuestas descrita en la Sección 7.4.3.1. El primer experimento cuantifica la percepción de los usuarios, a través del uso de mensajes verbales con información emocional, el algoritmo de sincronización y el movimiento de la boca de la cabeza robótica Muecas. Mientras, el segundo test evalúa los efectos en la percepción de los usuarios del uso de mensajes verbales con información emocional, el algoritmo de sincronización, el movimiento de la boca de la cabeza robótica Muecas, y los múltiples movimientos de los elementos de la cara de la cabeza, como las cejas, los movimientos del cuello ($Pitch$, $Roll$ y Yaw) y el movimiento de los ojos ($Tilt$ y Pan), entre otros. Para ambos experimentos las condiciones son similares, en ambos se generaran mensajes verbales con información emocional relacionada a los estados emocionales estudiados en esta Tesis Doctoral.

En el Cuadro 7.9, los resultados experimentales demuestran como el uso del lenguaje corporal con mensajes verbales que incluyen información emocional, presentan mejores resultados con respecto al uso únicamente de los mismo mensajes verbales y el movimiento de la boca. De este modo, los resultados experimentales comprueban la importancia del lenguaje corporal dentro de la comunicación e interacción entre humanos y robots, apoyando fuertemente la teoría del lenguaje natural basado en lenguaje corporal y voz para una comunicación similar a la humana.

Lenguaje Corporal	Preguntas			
	A	B	C	D
Voz y boca	60 %	76 %	72 %	80 %
Voz y Movimiento	71 %	78 %	84 %	92 %

Cuadro 7.9: Comparativa con respecto al uso del lenguaje corporal en la cabeza robótica Muecas.

7.5. Conclusiones

La constante evolución y desarrollo de los robots sociales requiere de soluciones que permitan mejorar el nivel de interacción, empatía y naturalidad. Todo ello por medio de sistemas basados en el lenguaje natural de los humanos, y en los elementos característicos de la comunicación verbal y no-verbal. Para lograr esto, se implementaron una serie de sistemas de imitación que permiten al robot intercambiar información emocional durante la interacción a través de sus movimientos y expresiones faciales, siendo acompañada por mensajes verbales que permiten, a su vez, la incorporación de información emocional por medio del cambio en parámetros de la prosodia en los sistemas de generación de voz sintética.

Aun así, el objetivo de este capítulo no es sólo imitar algunos patrones del comportamiento del usuario, sino aprender qué movimientos están asociados a cada emoción, dependiendo siempre de las capacidades y limitaciones físicas del robot. Por todo lo cual, la imitación de múltiples movimientos de la cara del usuario y las expresiones faciales sólo se presentan como una forma de utilizar de forma adecuada las características antropomórficas de la plataforma robótica Muecas en una interacción afectiva. En el caso de la información relacionada a la voz, el propósito del sistema de percepción del habla multimodal no es retroalimentar la interacción, sino permitir a un robot integrarse a través de una comunicación con diferentes tipos de usuarios, por medio de un comportamiento que busca obtener un elevado nivel de atención y empatía.

Los sistemas presentados fueron evaluados en este capítulo, tanto el sistema de imitación como el propio lenguaje corporal generado por el robot durante la comunicación. En este segundo caso, la evaluación se llevó a cabo mediante un estudio comparativo amplio que comprobó como la percepción de los usuarios mejoraba al implementarse un movimiento de la boca al hablar, y añadir información emocional tanto en la expresividad del robot como en el propio mensaje generado. Con este capítulo se demuestra cómo factores que son importantes durante una IHR, como la naturalidad o la expresividad, pueden ser percibidos de forma más simple y eficiente por el usuario, a través de estos algoritmos.

Parte II

Affordances emocionales

Capítulo 8

Estado del arte

8.1. Affordances

El concepto de $Affordances$ fue descrito por el psicólogo J.J. Gibson [Gibson, 1979] en su teoría ecológica, y representa un nuevo enfoque relacionado con la percepción del entorno. En su teoría, Gibson afirmaba que estas $affordances$ describen todas las posibilidades de acción que son materialmente posibles para un objeto, independiente de la capacidad del animal para percibir esta posibilidad, es decir, según palabras del propio autor *"the affordances of the environment are what it offers the animal, what it provides or furnishes, either for good or ill"*. En la literatura no existe una descripción o traducción directa de la palabra $affordances$ en sí, dando lugar a diferentes traducciones como *oportunidades dadas por el entorno*, *oportunidades ambientales*, o $permisividad$. Todas estas definiciones tratan de describir las idea original de Gibson, en pocas palabras, aquellas relaciones entre el ambiente y el usuario (animal o humano), que existen de forma natural.

El término $affordances$ ha sido analizado por varios autores a lo largo de los años, concluyendo en su mayor parte, que los conceptos dentro de esta teoría están basados sólo en un enfoque ecológico, normalmente relacionados a la percepción del entorno, la física ecológica o la relación entre percepción y acciones en seres vivos. La relevancia de estos conceptos dentro de los estudios de las $affordances$ se observa más claramente en trabajos como el descrito en [McGrenere and Ho, 2000], donde se explica cómo Gibson consideraba la importancia del entorno de los animales en todos sus estudios. Según los trabajos de Gibson, al analizar la percepción visual de los animales de forma aislada, se llegaba a un entendimiento falso del entorno, dado que, siempre según su teoría, la percepción por parte de los animales presenta un factor genético que le permitía obtener información estructurada desde su entorno natural, ya sea a nivel de medios, superficies, sustancias o a nivel de partículas y átomos, siendo imprescindible para su propia supervivencia. Esta información del entorno es adquirida por medio de una interacción directa y continua con su medio ambiente a través de acciones complejas. En el caso de la percepción de las $affordances$ en el entorno, Gibson plantea el término *Percepción Directa* [McGrenere and Ho, 2000], que se realiza cuando existe una $affordances$ y una información exacta del entorno que especifica la existencia de esa única $affordances$. Por ejemplo, alguien concluye que puede caminar por un puente sobre un río si observa que este puente está formado por tablas rígidas de madera. Aquí la $affordances$ sería la *posibilidad de caminar encima* y es percibida sin necesidad de observadores externos, únicamente a través de los sentidos, capaces de extraer la información relevante del entorno, y de la propia experiencia del animal

(*information pickup Theory*). Sin embargo, es importante clarificar el hecho que la percepción directa por medio de las experiencias o conocimiento pasado del usuario sólo está asociada a la percepción de las $affordances$, y no permite determinar la existencia de éstas.

Considerando todo lo descrito anteriormente, es posible citar tres aspectos relevantes de las $affordances$, como:

1. Las $affordances$ están limitadas por las capacidades o habilidades del usuario.

2. La existencia de una $affordances$ no depende de la capacidad para percibirla.

3. Las $affordances$ disponibles, que están limitadas por las capacidades del usuario, no cambian, aunque varíe el efecto que uno desee obtener.

En resumen, estos aspectos dan a entender que las $affordances$ dependen de las capacidades físicas para realizar una acción por parte de un usuario, sin tomar en consideración las experiencias o conocimiento previo de este. Además, es posible afirmar que cada $affordances$ existente que esté asociada a las capacidades del usuario, es invariante a la capacidad de percepción del usuario.

8.1.1. Affordances en la interacción humano computador

El concepto clásico de $affordances$ ha sido influenciado e introducido a otros campos, como la Interacción Humano-Computador (IHC), por investigadores como D. A. Norman a través de su libro *The Design of Everyday Things* [Norman, 2002]. En su libro, Norman discute el rol de la información perceptual dentro del concepto original planteado por Gibson, definiendo las $affordances$ ahora como "*...cada una de las posibles acciones asociadas a cada objeto del entorno, siendo limitadas por la capacidad para percibir las acciones disponibles por parte del observador, por medio de los atributos que componen cada objeto*". Así, en el enfoque desarrollado por Norman, se da prioridad al diseño de los objetos desde su forma y de las propiedades que los componen, que son denominados $atributos$ en la literatura.

La relevancia de estas propiedades dentro del concepto de $affordances$ es descrita por Norman en [Norman, 2002], donde "*...el término affordances se refiere a las propiedades percibidas y reales de las cosas, principalmente aquellas propiedades fundamentales que determinan cómo se podría utilizar posiblemente una cosa. Una silla ofrece (es para) apoyo y, por lo tanto, nos permite sentarnos. Una silla también se puede llevar,*". Esto se puede analizar desde el punto de vista de un diseñador, donde lo importante es si el usuario percibe que alguna acción es posible (o en el caso contrario, cuando no percibe ninguna $affordances$). Sin embargo, dentro de esta definición existe la condición de que las pistas visuales o atributos de los objetos sean compatibles con la capacidad de percepción de los usuarios, expresando que los atributos y las posibles acciones del usuario deben ser percibidas correctamente. Esto abre una discusión sobre cómo reconocer la relación entre los atributos y las acciones con respecto a las capacidades físicas del usuario, describiendo que es el aprendizaje previo quien le permite especificar qué acción debe tomar de acuerdo a cada objeto (con propiedades únicas). Esto es descrito por Norman de la siguiente forma: "*... Yo creo que las affordances resultan de la interpretación mental de las cosas, en base a nuestro conocimiento y experiencia pasada aplicada a nuestra percepción de las cosas que nos rodean* [Norman, 2002].

De forma similar, Gibson, en su teoría ecológica, consideraba la importancia de la percepción de las propiedades de los elementos o del entorno, como se recoge en "*... la composición y el diseño de superficies constituyen lo que ellos ofrecen. Si es así, para percibirlas es percibir lo que ellos ofrecen*" [Gibson, 1979]. No obstante, a diferencia de Norman, no se consideraban imágenes mentales, ni se hacía distinción entre las $affordances$ de un objeto de acuerdo a sus atributos. En su teoría, Gibson da más importancia a las capacidades de acción del usuario, mientras que para Norman son más importantes las capacidades de percepción y de representación mental del usuario. Por ejemplo, en el caso de dos objetos similares en forma y color como una manzana y una pelota roja, si utilizamos el primer principio (Gibson), un usuario o un robot pueden morder la pelota y lanzar la manzana, porque es posible dentro de sus capacidades. Mientras que en el segundo principio (Norman), el usuario debería morder la manzana y lanzar la pelota, debido a que percibe estas acciones a través de la información adquirida anteriormente, ya sea porque asocie algunos atributos del objeto con información de aprendizaje por imitación, o por experiencias pasadas. En la Figura 8.1 se ilustra una representación visual de este ejemplo desde los diferentes enfoques relacionados a las $affordances$.

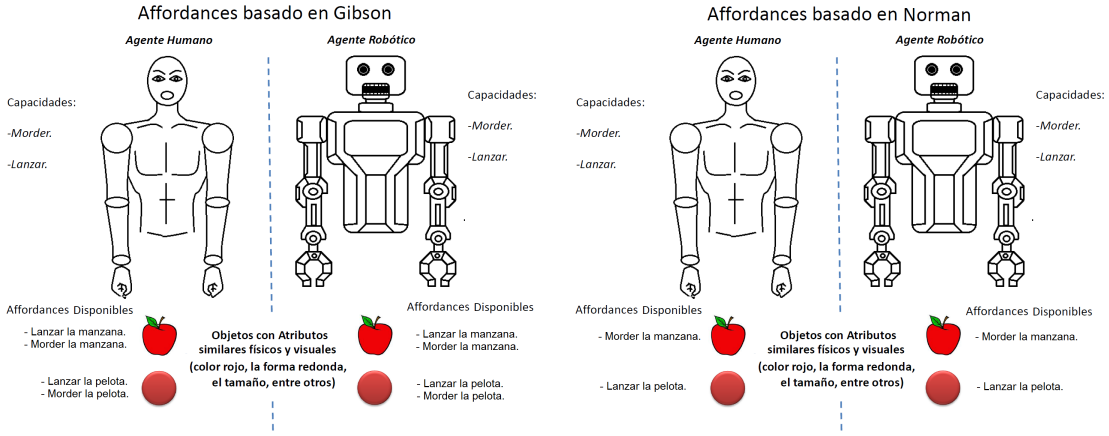
Esta clara prioridad relacionada a la percepción de las propiedades de los objetos es lo que genera la mayor parte de las diferencias entre las definiciones de Gibson y Norman, que son descritas en detalle en el Cuadro 8.1.

Affordances de Gibson	Affordances de Norman
Las posibilidades de acción en el medio ambiente están relacionadas con las capacidades de acción de un usuario.	Las propiedades percibidas pueden no existir realmente.
Son independientes de la experiencia, el conocimiento, la cultura, o la capacidad de percibir del usuario.	Sugerencias o pistas acerca de cómo utilizar las propiedades.
Su existencia es binaria - una affordances existe o no existe.	Puede depender de la experiencia, el conocimiento, o la cultura del usuario.
	Puede hacer una acción difícil o fácil.

Cuadro 8.1: Comparación entre las $affordances$ definidas por Gibson y Norman (Cuadro obtenido de la publicación [McGrenere and Ho, 2000])

En la literatura relacionada con sistemas basados en $affordances$, principalmente dentro de la IHC, el uso del concepto de $affordances$ basado en percepción introducido por Norman se presenta como el más utilizado en comparación al concepto original definido por Gibson. No obstante, a pesar de lo extendido y la buena acogida por parte de la comunidad científica acerca de la definición de $affordances$ por parte de Norman, en [Norman, 1999], Norman reconoce la ambigüedad asociada a su concepto, y decide diferenciarse del concepto de $affordances\ reales$ descrito por Gibson (que considera las oportunidades del entorno y no la percepción del usuario), al definir su concepto de $affordances$ como una extensión del concepto clásico denominado: $affordances\ perceptuales$ o $perceptibles$.

Finalmente, es importante mencionar que a pesar de que el termino $affordances$ es muy conocido y utilizado dentro de la IHC, no es un concepto totalmente comprendido. Esto genera extensiones o versiones propias del concepto original, que en muchos casos difiere de los estudios de Norman y Gibson. Un ejemplo claro de una extensión de este concepto, es el libro



(a) (b)

Figura 8.1: Representación de los diferentes enfoques de $affordances$ dentro de un ejemplo práctico: a) $affordances$ según Gibson; y b) $affordances$ según Norman.

Emotional Design: Why we love (or hate) everyday things [Norman, 2004], donde Norman habla sobre las relaciones existentes entre los objetos y las emociones. Esto deja abierto un campo para el uso de la información emocional en sistemas basados en $affordances$, que es el tema principal dentro de esta Tesis Doctoral.

8.1.2. Formalización del concepto de affordances

Como se mencionó anteriormente, el concepto de $affordances$ descrito en los trabajos de J.J. Gibson ha sido analizado y extendido en múltiples estudios, debido a que posee una gran ambigüedad que ha resultado en una gran variedad de aplicaciones de este concepto en diferentes campos de investigación. No obstante, el problema de esta variedad está relacionado a que no hay un entendimiento común del término $affordances$ dentro de la comunidad científica. Por ejemplo, aunque muchos investigadores estudian y citan el concepto desde los trabajos de Gibson, finalmente utilizan un concepto más acotado como es el descrito por Norman. Debido a este problema, Norman, dentro de sus últimos trabajos, analizó esta ambigüedad desde su definición de $affordances$, y se replanteó describirla más como una extensión del concepto original de Gibson, que denomina $affordances\ perceptivas$, donde explica que desde el punto de vista del diseño, importa mucho más lo que las personas perciban que lo que es real, debido a que dentro del diseño, las $affordances$ pueden ser reales o percibidas, e incluso ambas.

No obstante, esta diferencia entre las definiciones de Gibson y Norman, que cada vez se hizo más clara, causó que otros autores como Gaver [Gaver, 1991] propusieran un concepto más fiel al original, siendo considerado uno de los primeros trabajos que esperaban formalizar el concepto de $affordances$ dentro de la interacción Humano-Computador (IHC). En [Gaver, 1991], se describe este concepto de $affordances$ como: "*posibles acciones ofrecidas por un objeto o entorno, siendo estas acciones existentes sean o no percibidas por el usuario*", separando la percepción de las $affordances$ dentro de la interacción humano-computador. Al describir esta definición, Gaver comparte el mismo punto de vista relacionado a la forma y las propiedades de los elementos que Norman, pero difiere con respecto a qué información se relaciona a las $affordances$. En [McGrenere and Ho, 2000] se analiza esta diferencia, al describir que traba-

jos como los realizados por Norman o similares relacionan las $affordances$ directamente a la acción, mientras Gaver considera las $affordances$ como las propiedades (o atributos) del mundo (u objetos) que sugieren o hacen posible una acción, utilizando el término $diseño$ para especificar la información que se relaciona a las $affordances$.

Además de lo anterior, uno de los enfoques más relevantes dentro del concepto de Gaver, es la distinción de las $affordances$ desde el punto de vista de la información perceptible que esté relacionada con la misma, como se ilustra en la Figura 8.2. Gaver plantea una fuerte distinción en cuatro tipos específicos, basándose en la capacidad de realizar las acciones y percibirlas:

1. $Affordances$ **perceptibles:** se describen como $affordances$ que son perceptibles por el usuario. Estas se definen de forma similar a las $affordances$ descritas por Norman, dependiente de la capacidad para percibir una acción a través del conocimiento previo del agente, ya sea por medio de aprendizaje de experiencias o el contexto de la situación y siempre dependiente de cada usuario.

2. $Affordances$ **ocultas:** son $affordances$ que no son percibidas visualmente, pero sí pueden ser inferidas por medio del conocimiento y experiencia previa del usuario.

3. $Affordances$ **falsas:** se describen como las posibles acciones percibidas en el entorno, pero que no pueden ser llevadas a cabo por el usuario.

4. **Rechazo correcto:** está relacionada al hecho que dentro del entorno no existe ninguna $affordances$, ni errores de percepción para identificarla por parte del usuario.

Dentro de estos tipos de $affordances$, las ocultas y las falsas presentan una fuerte dependencia con la capacidad para percibir o inferir la existencia (o no existencia) de una $affordances$ por parte del usuario, ya sea por medio de la información visual o desde otro tipo de evidencia, como su propia experiencia. En el caso de las $affordances$ perceptibles, Gaver exploró diferentes aspectos relacionados a acciones complejas que requieran determinadas acciones o condiciones previas antes de ser consideradas una $affordances$. Introduce entonces dos nuevas definiciones:

1. $Affordances$ **secuenciales:** son $affordances$ donde una acción puede llevar a demostrar otras posibles acciones que puedan seguirle.

2. $Affordances$ **Anidadas:** también conocidas como *affordances agrupadas en el espacio*, son $affordances$ que, combinadas, revelan una determinada acción asociada a un propósito.

Para ser más específico, en las $affordances$ secuenciales se describen casos donde una $affordances$ perceptible se da a conocer después de un tiempo, constituyendo una serie de posibles acciones sobre el objeto. En cambio, el concepto de $affordances$ anidadas suele ser explicado mediante un ejemplo descrito en [Gaver, 1991], "*Una puerta solo puede sugerir una $affordances$ para su manipulación debido a su separación parcial de la pared, pero no especifica qué tipo de manipulación sería la más eficaz. Sólo viendo la $affordances$ de tirar del mango como anidado dentro de una $affordances$ de tirar de la puerta, puede abrirse la puerta al ser una $affordances$ perceptible*".

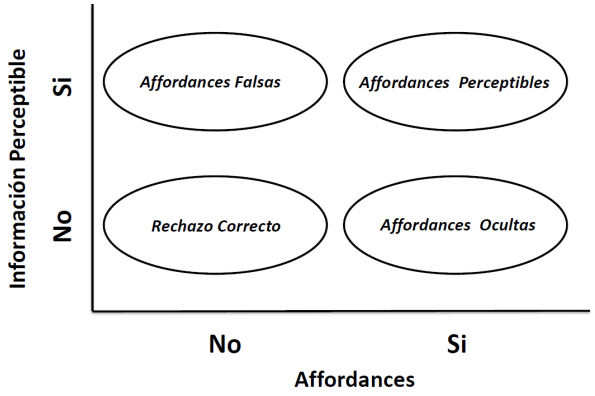


Figura 8.2: En [Gaver, 1991] se dividen las $affordances$ de acuerdo a la información adquirida de ellas, permitiendo distinciones entre correcto rechazo y $affordances$ percibidas, con respecto a $affordances$ falsas y ocultas (Figura adaptada de la publicación [Gaver, 1991]).

A pesar que el concepto descrito por Gaver se presenta como una formalización compleja y amplia de las $affordances$, no ha tenido una buena acogida dentro de la comunidad científica. Esto puede ser causado por la gran cantidad de publicaciones que extienden el concepto original de $affordances$ descrito por Gibson, y los múltiples trabajos que se basan en un concepto ambiguo de $affordances$, los cuales han generado estudios dedicados totalmente a formalizar y clarificar un concepto común dentro de la literatura. Específicamente, es posible analizar revisiones como [Sahin et al., 2007], donde se citan a autores con formalizaciones relevantes del termino $affordances$, tales como: Turvey [Turvey, 1992], Chemerro [Chemerro, 2003], Stoffregen [Stoffregen, 2003] o Steedman [Steedman, 2002]. Entre estos autores, se decidió analizar brevemente los puntos de vista acerca de cómo formalizar el concepto de $affordances$, a través de los estudios de Turvey y Chemerro, esperando que ayuden a clarificar y hacer más conciso el entendimiento de este término.

- **Formalización de Turvey:** siendo considerada una de las principales formalizaciones del concepto de $affordances$, Turvey define las $affordances$ como: "*Una clase particular de disposición, cuyo complemento es una propiedad disposicional de un organismo*". Dando a entender que las $affordances$ son $disposiciones\ en\ el\ ambiente$, y definiendo sus disposiciones complemento como las $efectividades$ del organismo. En este concepto, Turvey explícitamente asigna un factor relevante a las condiciones o atributos del entorno o ambientales, junto a las capacidades de los animales sobre cómo se generan y actualizan las $affordances$.

- **Formalización de Chemerro:** en [Chemerro, 2003], se define un concepto similar a Stoffregen [Stoffregen, 2003], que describe a las $affordances$ como: "*Una relación entre las capacidades de un organismo y las características del entorno*". Esta definición toma en consideración un enfoque que relaciona, mediante una interacción, las $propiedades\ del\ organismo$ y las $propiedades\ del\ entorno$, donde el comportamiento del organismo es proporcionado o afectado por el entorno. Esto rechaza otros enfoques que relacionan las $affordances$ únicamente con las propiedades del entorno.

Finalmente, se presenta un ejemplo de la ambigüedad en el significado del término $affordances$, que se puede ver reflejado en trabajos como el realizado por L.C. Vaughan

[Vaughan, 1997]. En este estudio, Vaughan analiza las $affordances$ desde el punto de vista en que las propiedades de un objeto o material entregan la información sobre ese objeto o material, basado en el libro de Gibson - $Un\ enfoque\ ecológico\ a\ la\ percepción\ visual$ -, que describe "*... la composición y el diseño de superficies constituyen lo que ellos ofrecen. Si es así, para percibirlas es percibir lo que ellos ofrecen. Esta es una hipótesis radical, ya que implica que los* $valores$ *y* $significados$ *de las cosas en el medio ambiente pueden ser percibidas directamente*" [Gibson, 1979]. Este enfoque, desde un punto de vista más amplio, describe qué factores, como el movimiento de un animal, entregan una gran cantidad de información, suficiente para reconocer las motivaciones, acciones y las emociones del individuo. Este análisis de las $affordances$ permite proyectar las emociones o intensiones humanas sobre un objeto basado en los movimientos.No obstante, aunque este trabajo está basado en la teoría de $affordances$, tiende a diferenciarse claramente del término definido por Gibson y Norman.

8.1.3. Aprendizaje basado en affordances

Dentro de la literatura, es posible identificar múltiples autores que hablan sobre mecanismos de aprendizaje basados en $affordances$ (*Learning of Affordances*). Entre estos, uno de los más significativos es el realizado por E. J. Gibson, quien estudia el desarrollo del aprendizaje en los niños desde el punto de vista de las $affordances$. En sus estudios, E. J. Gibson define al aprendizaje como "*..descubrir características distintivas y propiedades invariantes de las cosas y los eventos*" [Gibson, 2000] o descrito de otra forma "*..descubrir la información que especifica una* $affordances$" [Gibson, 2003], [Sahin et al., 2007]. En esta definición se da a entender que el proceso de aprendizaje está asociado a la capacidad de percepción del usuario, lo que denomina en su trabajo como $aprendizaje\ perceptivo$. Este tipo de aprendizaje no se relaciona con las respuestas a un estímulo, sino a la información percibida del medio ambiente, y cómo cada usuario es capaz de descubrir y comprender solamente una información critica o mínima asociada a las propiedades especificas de los elementos. Este proceso es denominado por la autora como $diferenciación$, el cual es definido como "*..la reducción desde una gran variedad de información (perceptual) a la mínima óptima que especifica las* $affordances$ *de un evento u objeto*" [Gibson, 2003].

Este tipo de aprendizaje perceptivo se realiza desde que los usuarios son bebés. Un recién nacido utiliza actividades de exploración (por ejemplo, escuchar o morder los objetos a su alcance) para adquirir información perceptiva acerca de los cambios en el entorno que sean resultados de sus actividades o acciones [Gibson, 2000]. Además, durante el desarrollo de los niños, las actividades de exploración se vuelven cada vez mejor planeadas y controladas, en busca de un objetivo final. La importancia de esta exploración por parte de los humanos, se debe a que la exploración autónoma tiene un papel importante en la teoría de Piaget sobre el desarrollo de los niños [Piaget, 1952], que describe cómo los comportamientos inteligentes se desarrollan por primera vez en el proceso de interacción con los objetos y el entorno.

Con respecto a los trabajos de J. J. Gibson en el campo del aprendizaje, en [Sahin et al., 2007] se hace referencia al hecho que E. J. Gibson declaró en una entrevista ([Szokolszky, 2003]), que J. J. Gibson no demuestra interés en este campo, sino que se centra solamente en la percepción. No obstante, en sus estudios el propio Gibson sugiere que algunas $affordances$ se aprenden en la infancia cuando un niño experimenta e interactúa con objetos externos [Gibson, 1979].

8.1.4. Affordances en la robótica

Desde su introducción en la interacción humano-computador por D.A. Norman, las $affordances$ han sido un campo de investigación en ascenso en las últimas décadas. Este concepto ha sido incorporado en múltiples áreas, desde el diseño de interfaces gráficas hasta la robótica. Dentro de esta última, las $affordances$ han tenido una muy buena acogida para su aplicación en sistemas robóticos autónomos, donde estudios como [Sahin et al., 2007] discuten los puntos en común entre la teoría de Gibson y la robótica basada en comportamientos, dado que ambos eran principios que iban en contra de los paradigmas predominantes y favorecían el modelado y la inferencia. No obstante, a pesar de poseer líneas de pensamiento similares y estar basados en una interacción directa entre un agente y su entorno, Gibson no concuerda con el uso de modelos del entorno y procesos de inferencia, al considerarlos innecesarios en una interacción directa. Esto se refleja en el hecho que las $affordances$ sólo necesitan información específica y relevante del entorno para ser percibidas, lo que comúnmente se conoce como $percepción\ directa$ dentro de los estudios de Gibson. Mientras, en la robótica sucede un hecho similar, debido a que cada actuador está relacionado a un sensor que le entrega la información necesaria para cada acción, lo cual causa que el robot se especialice en una información específica del entorno. Esta información específica del entorno asociada a una acción en la robótica, es lo que puede considerarse un equivalente del concepto de $percepción\ directa$ descrita por Gibson, y que tampoco necesita un proceso de modelado ni inferencia. Finalmente, es esta relación entre la percepción directa y las acciones en la robótica, lo que convierte a las $affordances$ en una de las soluciones más utilizadas para el desarrollo de sistemas cognitivos de bajo nivel y procesos de aprendizaje de alto nivel.

En los procesos de aprendizaje basados en $affordances$, es posible determinar a través de la literatura varios aspectos principales con respecto a su aplicación en la robótica. Por ejemplo, en [Sahin et al., 2007] se encuentra una revisión de algunos estudios que se agrupan por dos enfoques relevantes. Por un lado, el primer enfoque está relacionado con utilizar las $affordances$ en el aprendizaje del resultado de ciertas acciones en determinadas condiciones o entornos, el cual es representado por trabajos como [Fitzpatrick et al., 2003], [Stoytchev, 2005], entre otros. Por otro lado, los segundos métodos están asociados a utilizar las $affordances$ en sistemas de aprendizaje de propiedades invariantes del entorno, que ofrezcan ciertos comportamientos específicos a un agente robótico, siendo representado por trabajos como [Cos-Aguilera et al., 2003] o [MacDorman, 2000]

Entre los trabajos citados, en [Fitzpatrick et al., 2003] se estudia el aprendizaje de objetos a través de un agente robótico, que debe aprender qué acciones están relacionadas a cada objeto ($affordances$) y observar o cuantificar los efectos en el entorno. No obstante, como en otros estudios similares, no se crean relaciones entre las atributos (visuales en este caso) de los objetos y las $affordances$, que es el proceso contrario de lo que se espera lograr en esta Tesis Doctoral. Mientras, en [Cos-Aguilera et al., 2003] se analiza el uso de las $affordances$ dentro de un modelo que integra la selección del comportamiento y el aprendizaje de $affordances$ a través de la relación entre las características percibidas de los objetos y las consecuencias en la realización de una acción sobre un objeto, donde las consecuencias son monitorizadas de acuerdo a las metas u objetivos internos de un robot autónomo.

Con respecto a lo descrito anteriormente, en esta Tesis doctoral se decidió adoptar un concepto distinto que fusione los aspectos y características más importantes de ambos enfoques. En este trabajo se toma como base estudios que implementen sistemas de aprendizajes, como

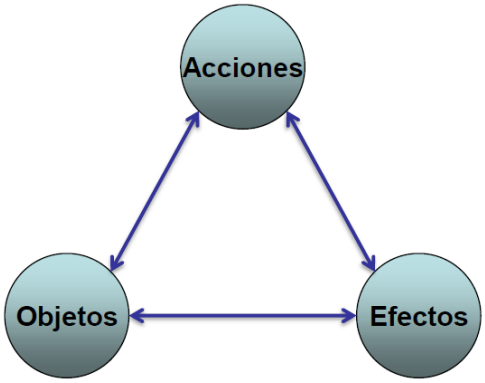


Figura 8.3: Representación del concepto de $affordances$.

Entradas	Salidas	Función
(O, A)	E	Predecir el efecto de un objeto o acción
(O, E)	A	Reconocimiento y planificación de una acción
(A, E)	O	Reconocimiento y selección de un objeto

Cuadro 8.2: Las $affordances$ representan un relación entre Objetos, Acciónes y Efectos (Información obtenida de la publicación [Montesano et al., 2008])

el descrito en [Montesano et al., 2008], donde se potencia el uso de las $affordances$ como una herramienta de aprendizaje, basada en el concepto de que cada acción está relacionada a las propiedades (atributos) de los objetos con los que interactúa, y los efectos de cada acción son cuantificables y predecibles. En la Figura 8.3 y el Cuadro 8.2, se ilustra este concepto que demuestra cómo se puede planificar una acción por medio de los objetos existentes y el reconocimiento de una acción específica para obtener un efecto esperado. En términos técnicos, esta relación de las $affordances$ es lo que permite a un agente robótico reconocer los elementos del entorno y realizar una acción en respuesta, ya sea a través de las diferentes articulaciones o motores del robot, o planificando o prediciendo un efecto. Esto último es lo que lleve al robot a seleccionar una acción a través de un repetitivo proceso de aprendizaje que le permita llevar a cabo un objetivo dentro de una interacción.

Finalmente, es posible describir algunos estudios que demuestran la versatilidad y dinamismo de las $affordances$ en la robótica, ya sea en estudios relacionados a la predicción de las $affordances$ en objetos para tareas de manipulación [Hermans et al., 2011] basado en el concepto de atributos de Gibson, o sistemas de aprendizaje por imitación basado en la interacción con objetos [Lopes et al., 2007]. Además, es posible encontrar múltiples estudios que implementan modelos de aprendizaje basados en $affordances$ por medio de redes bayesianas, que relacionan las acciones, los efectos y las características de los objetos [Montesano et al., 2007], siendo este mismo método utilizado en otros trabajos diferentes, tales como: [Osório et al., 2010] y [M. Kammer and Nagai, 2011], donde los autores extienden el concepto de $affordances$ al tomar en consideración también el contexto del entorno.

8.1.5. Affordances emocionales

Desde el desarrollo de las $affordances$ perceptuales de Norman dentro de la IHC, se han desarrollado múltiples extensiones del concepto clásico desarrollado por Gibson [Gibson, 1979], debido a las posibilidades que permite este concepto dentro de diferentes campos de investigación. Una de estas extensiones en el contexto de IHR afectivas, es el concepto de $affordances$ emocionales ($emotional\ affordances$), definido como la relación existente entre los distintos elementos afectivos, los efectos y las posibilidades para las reacciones del observador. El propio Norman, en su libro *Emotional Design: Why we love (or hate) everyday things* [Norman, 2004], hace referencia a las posibilidades que presentan las relaciones existentes entre los objetos y las emociones. La única referencia en este sentido que se encuentra en la literatura viene dada por el trabajo de [Morie et al., 2005], donde se presentan estas $affordances$ emocionales en el contexto de la generación de entornos virtuales y cómo estos afectarían emocionalmente al usuario. En el área de la robótica, la primera mención de estas $affordances$ emocionales se realiza en [Cid et al., 2013a], como un término que relaciona qué posibilidades presenta un objeto o una posible expresión facial, para provocar una determinada reacción en el robot o en el usuario, sabiendo el efecto deseado. Un ejemplo es el caso de alguien que da un regalo y a la vez sonríe a un robot, éste último no sólo sabe cómo puede tomar el regalo ($affordances$ perceptuales), sino que también su reacción lógica es un estado emocional $positivo$. De acuerdo con este concepto, las $affordances$ se pueden ver como un $continuum$, desde las $affordances$ perceptuales hasta las emocionales, que ilustre los dominios complementarios y sobrepuestos de la percepción y la emoción [Morie et al., 2005]. (Ver Figura. 8.4).

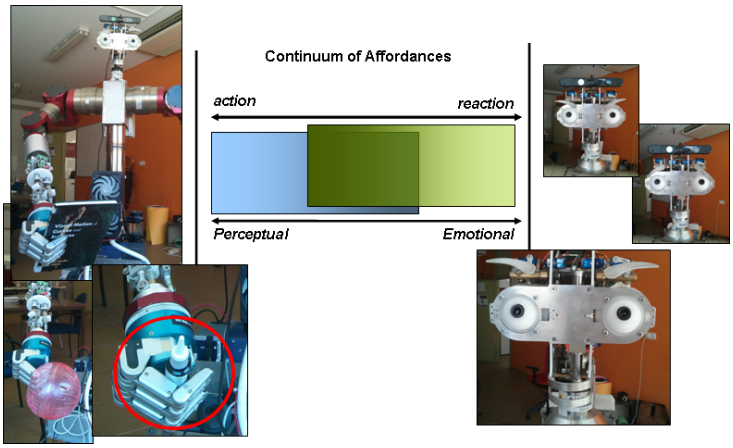


Figura 8.4: Continuum de $affordances$ desde perceptivas a emocional. A la izquierda, diferentes objetos (en tamaño, color y forma) infieren diferentes planes y acciones ($affordances$ perceptivas). A la derecha, estos mismos objetos y el estado emocional de los participantes en la comunicación también infieren una reacción en el robot. (Figura extraída de la publicación [Cid et al., 2013a])

En la Figura 8.5, se observa un representación gráfica de este concepto de $affordances$ emocionales, describiendo cómo los elementos afectivos ($estímulos$) crean relaciones con posibles reacciones y efectos asociados al estado emocional del usuario. En este enfoque, las $affordances$ emocionales pueden ser aplicadas en IHR donde se persiga afectar o manipular

el estado emocional del usuario mediante la predicción del efecto que tendrá sobre el mismo cada reacción emocional del robot.

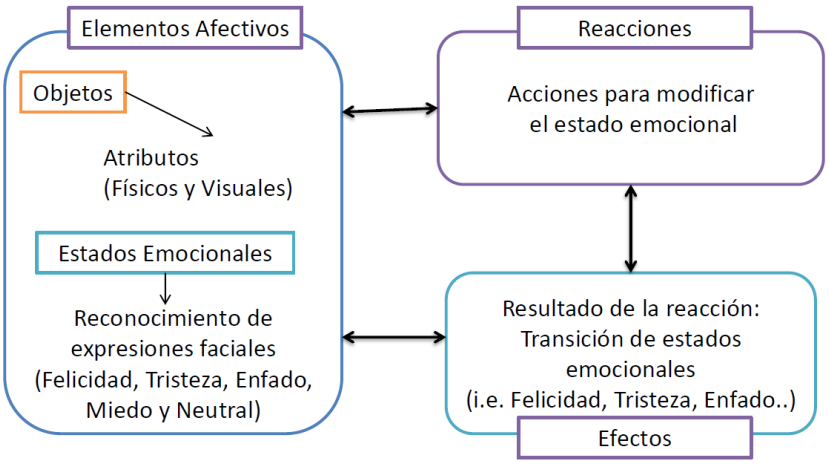


Figura 8.5: Las $affordances$ emocionales describen la relación entre diferentes $estímulos$ (elementos afectivos) con las reacciones afectivas y el efecto esperado. (Figura adquirida de la publicación [Cid and Núñez, 2014])

Capítulo 9

Sistema de aprendizaje de comportamientos afectivos basado en affordances emocionales

9.1. Introducción

En el campo de la Interacción Humano-Robot, el desarrollo de robots autónomos capaces de ayudar en la rutina diaria al usuario, ya sea en entornos controlados o en escenarios reales, se presenta como uno de los desafíos más importantes. Una interacción en el mundo real requiere de avanzadas capacidades sociales que permitan a los robots comunicarse mediante reglas con las personas a través del lenguaje natural. Esto lleva a la necesidad de que cada robot debe disponer de las habilidades para percibir y aprender sobre las diferentes acciones que se pueden llevar a cabo en interacciones con usuarios, siempre considerando las capacidades y limitaciones asociadas al propio diseño de estos robots. Estas acciones asociadas a las capacidades físicas del usuario y su capacidad para percibirla son denominadas $affordances$ perceptuales [Norman, 2004].

Este concepto de $affordances$, como fue descrito en el capítulo anterior, ha sido tomado directamente del campo de la psicología. Las $affordances$ demuestran la importancia de las capacidades del individuo para percibir las posibles acciones del entorno, a través de su propio conocimiento, un conocimiento adquirido a través de los recuerdos, la experiencia previa o por medio de un proceso de aprendizaje. Este último es el aspecto relevante en el caso de los robots, cómo adquirir la información a partir de un proceso previo de aprendizaje, principalmente si consideramos esta idea dentro de la IHR, donde los robots no pueden adquirir conocimiento de forma libre, ni pueden realizar interacciones complejas con el usuario.

Al considerar esta problemática asociada a la robótica y buscando una solución en la teoría de $affordances$, se llega a sistemas como el descrito en [Lopes et al., 2007], donde, considerando las limitadas capacidades de un robot, implementaron una teoría para transferir información directamente entre el usuario y éste mediante una simple imitación. El desarrollo de trabajos como el presentado en [Lopes et al., 2007] permite aplicar estas $affordances$ también en la teoría de las emociones, lo que se ha denominado en esta Tesis Doctoral como $affordances$ emocionales en una IHR afectiva. Este concepto extendido describe las relaciones entre los elementos afectivos o estímulos con respecto a posibles reacciones emocionales y efectos asociados con el estado emocional del usuario. En este enfoque, es posible dotar al robot con la

capacidad de afectar o manipular el estado emocional del usuario mediante la predicción del efecto que tendrá sobre el individuo cada reacción emocional del robot asociada a un elemento.

En este capítulo se describe un sistema de aprendizaje de comportamientos afectivos basados en las $affordances$ emocionales. El objetivo principal del sistema es dotar al robot con la capacidad de orientar el estado emocional del usuario (felicidad, miedo, tristeza, enfado y neutral) hacia una emoción dada, definida por su propio comportamiento. Todo ello durante una IHR afectiva, que incluye expresiones faciales por parte del robot, así como la interacción con los objetos del entorno. Para lograr esto, el robot debe estar provisto de una capacidad de aprendizaje que le permita reconocer y aprender las relaciones entre los diferentes estímulos y cómo afectan al estado emocional final del usuario para una reacción emocional predecida.

En el caso del proceso de aprendizaje, se desarrolla un modelo matemático del concepto de $affordances$ emocionales a través de una red bayesiana que permite analizar y adquirir información de los diferentes elementos afectivos durante la IHR [Cid and Núñez, 2014]. Para la percepción de estos elementos afectivos o estímulos, el robot cuenta con la capacidad de reconocer el estado emocional del usuario y los objetos existentes en el entorno. Este modelo matemático incluye la representación de las relaciones que caracterizan las $affordances$ emocionales durante la interacción, y que se describieron en el Capítulo 8. Por su parte, el proceso encargado del comportamiento del robot para guiar la comunicación hasta un estado emocional final en el usuario está basado en modelos de comportamiento construidos sobre máquinas de estado, similar a los trabajos presentados en [Breazeal et al., 2008] y [Schulte et al., 1999].

9.2. Descripción del sistema

En este capítulo se presenta un sistema de aprendizaje de comportamientos afectivos basado en las $affordances$ emocionales, que tiene como objetivo principal dotar a un robot con la capacidad de orientar el estado emocional del usuario durante una IHR. Para llevar a cabo este propósito, esto es, cambiar el estado emocional del interlocutor (felicidad, tristeza, miedo, enfado o neutral) a un estado específico, el robot ha de usar tanto la información emocional adquirida durante la interacción, como un número limitado de objetos del entorno, comunes en la vida diaria. Con sólo estos elementos afectivos, también denominados estímulos, en la definición del problema, esto es, la información emocional del usuario y los objetos del entorno, se requiere la resolución de dos cuestiones: i) cuáles son los efectos y las relaciones entre los objetos presente en el entorno con el estado emocional del usuario; y ii) cómo se puede orientar un estado emocional específico en el interlocutor con una limitada cantidad de elementos. Para solucionar ambos problemas se desarrollaron dos sistemas que tienen como base las $affordances$ emocionales. El primero es un método de aprendizaje emocional, que permita al robot adquirir información del comportamiento del usuario durante la IHR. El segundo es un sistema relacionado con el propio comportamiento del robot, que le permite desarrollar una serie de condiciones o reglas acerca de cómo afectar el estado emocional del usuario para alcanzar una emoción especifica.

Por un lado, para analizar los efectos de los elementos del entorno en el estado emocional del usuario, ha sido necesaria la implementación de un sistema de aprendizaje que tiene su base teórica en las $affordances$ emocionales, y que permite a un robot aprender las relaciones afectivas entre las respuestas emocionales del usuario (información emocional) y los diferentes objetos usados en la interacción. Para percibir y adquirir la información asociada a estos ele-

mentos afectivos, el sistema de aprendizaje requiere del uso de un sistema de reconocimiento de emociones, para estimar la información emocional del usuario, y de un sistema de reconocimiento de objetos, que permita adquirir información acerca de los atributos físicos y visuales de los mismos.

Por otro lado, para lograr una IHR real, donde se permita al robot condicionar la respuesta emocional del usuario, ha sido necesario la implementación de modelos de comportamientos afectivos basados en maquinas de estado. Estos modelos analizan la interacción y estiman las respuestas emocionales del usuario de acuerdo a las opciones disponibles según su aprendizaje. Así, de acuerdo a cada modelo, es posible llegar a un estado específico pasando por varios estados transitorios, mediante el uso de elementos como los objetos en una interacción.

9.3. Modelado de las $affordances$ emocionales

Dado que la primera parte del sistema propuesto está relacionado con un proceso de aprendizaje por observación, donde el robot, durante una IHR real aprende la correspondencia entre las características de un objeto y las reacciones emocionales del usuario, se ha desarrollado un modelo del concepto de las $affordances$ emocionales. En este modelo, por medio de un enfoque bayesiano, se establece las relaciones existentes entre los efectos o respuestas emocionales del usuario con elementos reales del entorno. Tal y como fue descrito en el capítulo anterior, estas $affordances$ emocionales analizan las correlación entre los distintos elementos afectivos, los efectos y las posibilidades para las reacciones afectivas del observador, esto es, qué posibilidades presenta un objeto o una posible expresión facial del robot, para provocar una determinada reacción emocional en el usuario sabiendo el efecto deseado [Cid et al., 2013a]. Por lo tanto, a través de este modelo, cuando un robot realiza acciones específicas de acuerdo a la información de los elementos afectivos, es posible predecir el efecto o resultado emocional esperado sobre un usuario humano. El número y tipo de acciones posibles que el robot puede realizar durante la interacción está limitado por las propias capacidades básicas del robot para percibir e interactuar con los elementos afectivos.

De acuerdo con esto, el propósito del sistema de aprendizaje descrito en este capítulo es el de adquirir información acerca de los efectos (reacciones emocionales) que provocan sobre el usuario los elementos afectivos durante la IHR. Esta adquisición de información la realiza el robot por observación directa, y es la clave de todo el proceso. Por ello, el sistema de percepción del robot de cada uno de los elementos afectivos durante la interacción determina los principales problemas a resolver por el sistema, puesto que no sólo existen elementos físicos como los objetos, sino también conceptuales como las reacciones emocionales del usuario.

Esta necesidad de percibir los elementos afectivos por parte del robot, requiere de dos sistemas que doten al robot con estas capacidades básicas:

1. La capacidad para reconocer el estado emocional del usuario durante la comunicación (felicidad, tristeza, miedo, enfado y neutral), por medio de algunos de los diferentes canales del lenguaje natural. Dado el tipo concreto de interacción durante el aprendizaje, donde predomina la información facial sobre el resto de las fuentes de información, el sistema propuesto utiliza el método de reconocimiento de emociones basado en el modelo de malla $Candide-3$, descrito en el Capítulo 3.

2. La capacidad para reconocer y distinguir determinados objetos en el entorno, por medio

de sus atributos físicos y visuales. Para emular esta capacidad, común en los seres vivos, se ha implementado un sistema basado en marcas 2D que utiliza la información visual para localizar, reconocer y adquirir la información del objeto, principalmente la relacionada con sus atributos físicos y visuales.

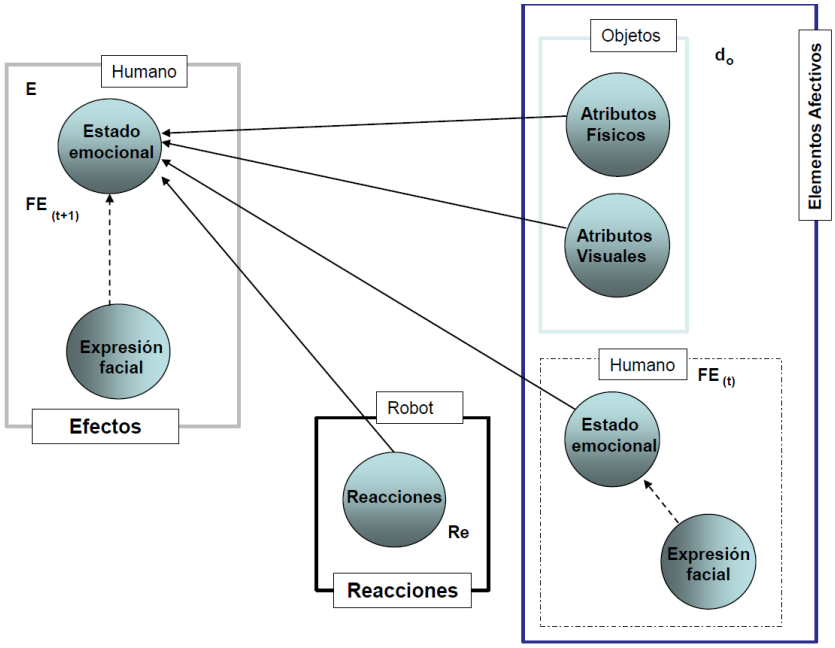


Figura 9.1: Modelo que representa el concepto de $affordances$ emocionales (Figura adaptada de la publicación [Cid et al., 2013a])

Estas capacidades básicas obtienen la información de entrada de este modelo de $affordances$ emocionales, y se consideran igualmente la entrada del sistema de aprendizaje por imitación descrito en este capítulo.

En la descripción del modelo de $affordances$ emocionales, el robot realiza una serie de reacciones en el instante de tiempo t, que son representadas por las variables discretas aleatorias $Re = \{re_t\}$. Por su lado, los elementos afectivos y los efectos en la interacción también son modelados por medio de variables discretas aleatorias. En el caso de los elementos afectivos, los estados emocionales del usuario, estimados por medio del sistemas de reconocimiento basado en expresiones faciales, son representados por la variable FE, en el instante de tiempo t. Los objetos son definidos por una serie de atributos (Visuales y físicos) por medio de $d_o = \{d_o(1), ..., d_o(n_o)\}$., siendo n_o el número total de objetos. Finalmente, $E = \{E(1), ..., E(n_e)\}$ representa los efectos o cambios detectados por el robot en el estado emocional del usuario después de realizar una reacción. El conjunto de nodos G está formado por las variables (Re, FE, d_o, E). Este grafo representa el modelo de red bayesiana implementado en esta Tesis Doctoral. En la Figura 9.1 se ilustra una representación del modelo de $affordances$ emocionales descrito.

Las siguientes secciones describen los diferentes aspectos necesarios durante una interacción basada en $affordances$ emocionales, tanto los elementos afectivos (objetos y los información emocional del usuario), como las respuestas o reacciones emocionales de los participantes en la IHR, o las propias plataformas robóticas utilizadas en la evaluación.

9.3.1. Elementos afectivos

Dentro de las $affordances$ emocionales, los elementos afectivos están relacionados con los diferentes objetos del entorno y con los estados emocionales del usuario. En ambos casos es necesario definir qué atributos son compatibles con el sistema de percepción del robot, para así determinar qué acciones son posibles dentro del rango de acciones deseadas por éste. En el caso de los objetos, estos atributos están determinados por sus propiedades físicas y visuales ($affordances$ perceptuales). Por su parte, en el caso de la información emocional, estos atributos se corresponden con información visual, en este caso, por las características faciales del usuario. Las acciones que el robot puede realizar sobre estos elementos afectivos son limitadas, por un lado su manipulación (objetos) y por otro, la imitación (información emocional). La Figura 9.2 resume la relación entre los objetos y un robot, tomando como base el concepto de $affordances$.

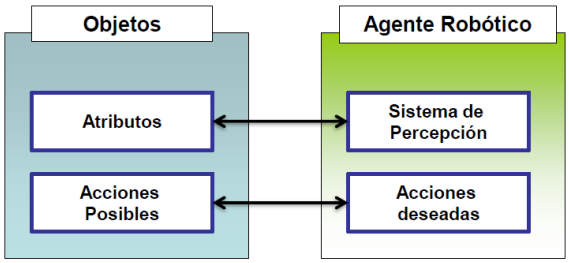


Figura 9.2: Representación de la relación entre objetos y agentes robóticos

9.3.2. Reconocimiento de objetos

El reconocimiento de objetos es uno de los campos más activos en la robótica actual, debido a su enorme complejidad y las múltiples áreas de aplicación, que van desde $grasping$ hasta la IHR. Uno de los principales problemas de estos sistemas es el manejo de la información visual obtenida por el robot, que normalmente requiere de un elevado coste computacional y complejos procesos matemáticos. Puesto que el objetivo de esta Tesis Doctoral no es reconocer los objetos, sino identificarlos en el entorno para la interacción, en este trabajo se ha elegido un enfoque mucho más simple para solventar el problema. Gracias al sistema propuesto, con una gran precisión y rapidez se obtiene información relacionada con las propiedades físicas y visuales de los objetos, lo que se ha denominado como atributos en este trabajo.

Por todo lo anterior, se ha considerado un sistema de reconocimiento de objetos basado en marcas 2D. Cabe destacar que este tipo de sistema permite el uso de múltiples y diferentes objetos, lo cual no sería posible con un sistema de reconocimiento basado en imágenes RGB, incluso en sistemas que utilizan atributos similares como el descrito en [Hermans et al., 2011]

En lo referente a los objetos, este sistema propone su implementación en escenarios afectivos controlados, en los cuales se observan elementos comunes de una casa, por ejemplo, tazas, teléfonos, juguetes o mascotas. El rango de objetos presentes en este escenario está asociado a distintas categorías. La pertenencia o no de un objeto a una categoría depende de los atributos que éste posea, siendo el valor de estos atributos elegidos en base a estudios relacionados con las $affordances$ [Montesano et al., 2007]. El conjunto total de objetos y marcas utilizadas en

este trabajo se presenta en el Apéndice B. En la Figura 9.3 se ilustra un ejemplo de tres objetos, con sus marcas asociadas.

Figura 9.3: Objetos utilizados en la interacción con sus respectivas marcas

9.3.2.1. Sistema de reconocimiento de objetos basado en marcas

Cada uno de los objetos presentes durante la interacción está asociado a una lista de atributos que permiten al robot diferenciar internamente uno de otro. El proceso de reconocimiento llevado a cabo por el sistema hace uso de marcas bidimensionales, colocadas sobre los objetos, de forma que el robot, por observación directa, extrae la información asociada a estos atributos. Siguiendo este enfoque, se decidió utilizar dos librerías diferentes, $ARToolKit$ y $AprilTags$, muy conocidas y usadas comúnmente en robótica. Cada librería tiene asociadas diferentes clases de marcas, como puede observarse en la Figura 9.4, donde se ilustra el objeto real (la taza de la Figura 9.4a), junto a las marcas de $ARToolKit$ y $AprilTags$ utilizadas (Figuras 9.4b y 9.4c, respectivamente).

A continuación, se describirán las características de cada librería, así como otros sistemas alternativos testeados durante este trabajo junto con los diferentes atributos que describen a cada objeto.

- **ARToolKit**

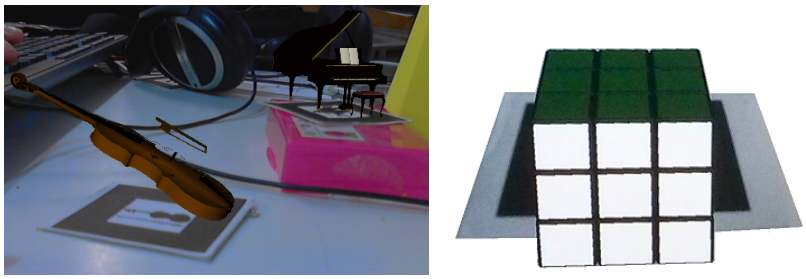
  La librería $ARToolKit$ [Human Interface Technology Laboratory, 2014] descrita en el Apéndice B.2.1, se presenta como una solución para el problema del reconocimiento de objetos basada en el seguimiento de marcas [Kato and Billinghurts, 1999]. Esta librería fue creada principalmente para su trabajo en sistemas de realidad aumentada. El uso de $ARToolKit$ dentro de este sistema se debe a su capacidad de reconocimiento y representación virtual de objetos mediante marcas, lo que fue utilizado durante la etapa de desarrollo, junto con el $toolkit\ OSGART$ [Looser et al., 2006], [HITLabNZ, 2014] para representar los modelos de los objetos reales, permitiendo reducir la magnitud de los experimentos iniciales. En la Figura 9.4b se muestra una marca de $ARToolKit$ asociada a una taza real. Mientras, en la Figura 9.5 se muestra un ejemplo del uso de $ARToolKit$ y $OSGART$ para representar un modelo virtual.

(a) (b) (c)

Figura 9.4: Representación de objetos basados en marcas; en este caso una *taza*: a) taza real en las pruebas; b) marca $ARToolKit$; y c) marca de $AprilTags$.

- **AprilTags**

  La librería $AprilTags$ [Olson, 2011] descrita en el Apéndice B.2.2, representa al segundo tipo de marcas 2D utilizadas en esta Tesis. La elección de esta librería está basada en sus características de detección e identificación de las marcas con bajas condiciones de luz, o con una visión parcial de una parte de la marca, debido a que contiene sólo unos cuantos bits de información. Además, $AprilTags$ que entrega la posición relativa de cada marca con respecto al agente robótico como centro del sistema, también está basada en marcas predeterminadas, de forma que existen conjuntos de marcas que forman familias predefinidas (en este trabajo se utilizó la familia $tag36h11$). En la Figura 9.4.c se ilustra una marca de $AprilTags$ asociada a una taza real.



(a) (b)

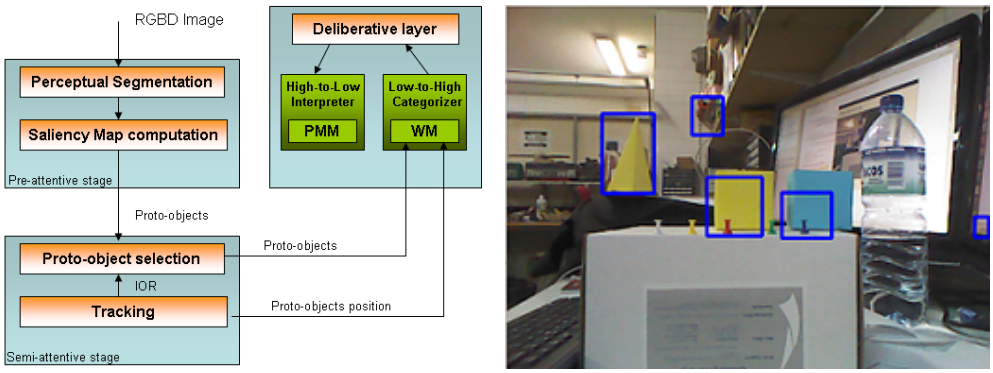
Figura 9.5: Funcionamiento de la librería $ARToolKit$ con $OSGArt$

Finalmente, las pruebas realizadas para comprobar el rendimiento de estos dos tipos de marcas demostraron serias diferencias en la precisión de la localización y en el reconocimiento con diferentes condiciones de luz. Por un lado, la librería $ARToolKit$ demostró una respuesta rápida y eficiente en la detección de las marcas en condiciones óptimas, pero presento serios problemas en los casos donde existían bajas condiciones de luz (ej, interferencia por luz natural directa) o cuando una parte de la marca no es visible. Mientras, que la librería $AprilTags$ no presentó ningún problema con bajas condiciones de luz, y entregó los mejores resultados en la localización de la marca incluso si una parte no es visible o se encuentra inclinada. Por este motivo, y por la entrega de información precisa de la posición y orientación 3D de la marca con

respecto a la cámara del agente robótico, se decidió utilizar las marcas de $AprilTags$ en los experimentos finales del Capítulo 10.

9.3.2.2. Sistemas de reconocimiento alternativos

Durante el desarrollo de la detección y reconocimiento de objetos por el agente robótico Muecas, se tomaron en consideración múltiples sistemas o métodos que pudieran reconocer los elementos del entorno por medio de la pistas visuales adquiridas desde sus sensores. En trabajos anteriores como [Cid et al., 2013a], se consideró otro sistema en la detección de los objetos del entorno. Este sistema de reconocimiento de objetos es descrito en [Palomino et al., 2011], el cual propone un método de percepción básico basado en $proto\text{-}objetos$ [Orabona et al., 2007], una unidad de atención basada en el principio de un modelo de atención visual. En la Figura 9.6 se ilustra una visión general de la estructura de este sistema y el reconocimiento de múltiples objetos del entorno. Mientras, la Figura 9.7 muestra las diferentes etapas del sistema para un caso concreto. Este algoritmo de reconocimiento fue utilizado en el desarrollo inicial del concepto de $affordances$ emocionales, descrito en [Cid et al., 2013a], no obstante, los resultados finales de las pruebas de rendimiento realizadas a este sistema demostraron elevados tiempos de retardo que no permitían un uso adecuado con múltiples y cambiantes objetos en tiempo real, lo cual lo convierte en una opción inviable para su implementación en la cabeza robótica Muecas.

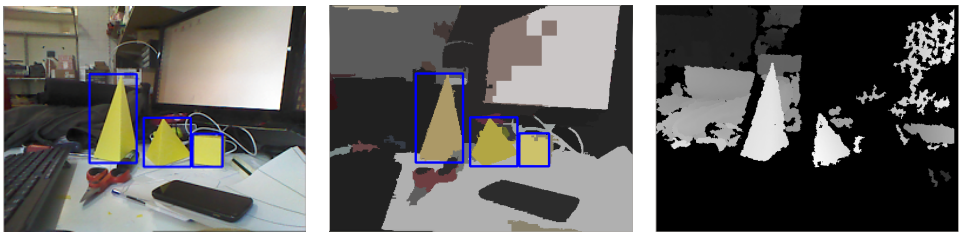


(a) (b)

Figura 9.6: Sistema de reconocimiento de objetos basado en información RGB-D; a) Descripción general del modelo de atención basado en objetos (Figura obtenida de la publicación [Cid et al., 2013a]); y b) Imagen del reconocimiento de objetos.

9.3.3. Atributos

Los objetos utilizados en una interacción basada en $affordances$, requieren de múltiples atributos que permitan al robot distinguir y agrupar cada uno de estos objetos, así como determinar el rango de posibles acciones. Por este motivo, la adquisición de los atributos de cada objeto del entorno mediante marcas 2D se presentó como la mejor solución para proveer información de los elementos del entorno al robot, de forma rápida y precisa. Debido a que el uso de este sistema de reconocimiento basado en marcas, en comparación a otros sistemas de reconocimiento basados en pistas visuales e información de sensores RGB-D [Sun et al., 2009] [Koppula et al., 2013], no limita el número de objetos, ni el tipo de atributos a través de sus



(a) (b) (c)

Figura 9.7: Reconocimiento de objetos basados en un sistema alternativo, a) Identificación de los objetos de color amarillo; b) Segmentación perceptual de las objetos (Extracción de $proto-objetos$); y c) Imagen con contraste de color.

propiedades visuales. Esto permite un gran número de objetos y diferentes tipos de atributos visuales y también físicos. Con respecto a la elección de los atributos se tomó en consideración varios estudios relacionados con la manipulación de objetos basados en $affordances$ [Hermans et al., 2011], [Lopes et al., 2007], que determinaron atributos relacionados principalmente a las capacidades de percepción de los agentes robóticos.

Finalmente, los atributos de los diferentes objetos utilizados en este trabajo se dividen en 6 grupos de variables v:

- **AS** - *Forma*: agrupa los objetos por su forma en 12 posibles valores, tales como: esferas, cilindros, triángulos, conos, octaedros, toroides, cubos, laminas, 4 caras, 6 caras, 8 caras, entre otros.

- **AC** - *Color*: agrupa los objetos por una determinada característica como el color, por medio de 19 posibles valores, tales como: blanco, rojo, negro, azul, verde, celeste, naranjo, gris, entre otros.

- **AM** - *Material*: agrupa los objetos por medio de una determinada propiedad como el material que lo compone en 9 posibles valores, tales como: metal, plástico, cerámica, vidrio, goma, material textil, orgánico-comida, orgánico-plantas y orgánico-animal.

- **AW** - *Peso*: agrupa los objetos por su peso en 6 posibles valores que definen determinados grupos, tales como: muy ligero, ligero, medio, medio pesado, pesado y muy pesado (escala realizada en gramos).

- **ASi** - *Tamaño*: agrupa los objetos por el tamaño en 6 posibles valores que definen determinados grupos, tales como: muy pequeños, pequeños, mediano, medio grande, grande y muy grandes (escala realizada en cm).

- **A1** - *Acción*: agrupa los objetos por las posibles acciones que se pueden aplicar dependiendo de su forma, material y peso, por medio de 7 posibles valores, tales como: empujar, enrollar, arrastrar, levantar, apretar, entre otros.

La lista de los objetos con sus respectivos atributos, que fueron utilizados en los experimentos del Capítulo 10, es descrita en el Apéndice B.1.

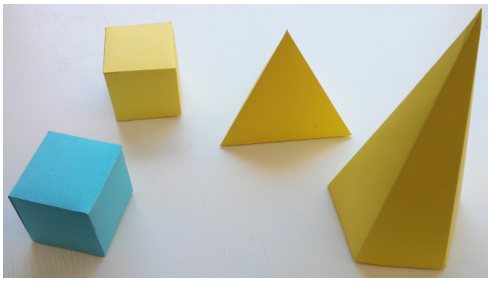


Figura 9.8: Objetos geométricos

9.3.4. Acciones y respuestas emocionales

En el concepto de $affordances$ emocionales , las acciones y respuestas emocionales, ya sean del usuario o del agente robótico, están asociadas a sus capacidades y limitaciones físicas. Debido a que estas capacidades son las que permiten a un agente robótico interactuar con el usuario, a través de múltiples elementos afectivos o acciones que afecten el estado emocional del usuario. Por este motivo, a continuación se describen las diferentes capacidades que se tomaron en consideración, durante el desarrollo y evaluación del sistema de aprendizaje propuesto.

1. Usuario:

   En el caso del usuario, se implementaron múltiples sistemas de reconocimiento de emociones descritos en esta Tesis Doctoral, los cuales toman en consideración 3 enfoques basados en la capacidad humana de transferir información emocional por medio del lenguaje natural, lo que se ilustra en la Figura 9.9a. Por un lado, dos de estos enfoques están basados en sistemas que hacen uso de la información visual para reconocer emociones por medio de las características faciales extraídas de las expresiones faciales y las características corporales del lenguaje corporal, siendo el uso de las expresiones faciales el enfoque más fiable y rápido dentro de lo mencionado en la literatura. Por otro lado, el tercer enfoque utiliza las características acústicas asociadas a la prosodia de la voz, para reconocer el estado emocional del usuario. Finalmente, el último sistema se basa en un enfoque multimodal que hace uso de características faciales y acústicas de forma simultanea, para estimar el estado emocional del usuario.

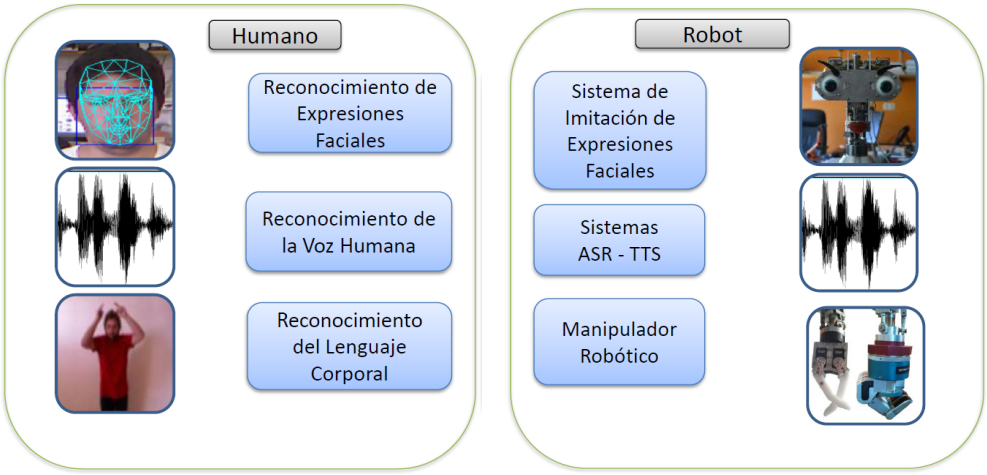
   En conclusión, estos enfoques estiman los mismos estados emocionales: felicidad, enfado, tristeza, miedo y neutral. Por tanto, se decidió elegir solo el sistema de reconocimiento de emociones basado en expresiones faciales para que reconozca la información emocional desde el usuario dentro de este sistema de aprendizaje, debido a que presenta los mejores resultados en todo tipo de entornos no controlados y posee una relación directa con la capacidad de imitar las expresiones faciales del agente robótico Muecas.

2. Robot:

   El robot presenta unas capacidades limitadas desde el momento de su diseño, lo cual causa que sea capaz de realizar acciones asociadas a una sola aplicación. No obstante, para este sistema se esperaba que el agente robótico cumpliera como mínimo con las capacidades descritas en la Figura 9.9b, que le permitiera realizar las acciones: comunicar información objetiva por medio de la voz, transferir información emocional por medio de expresiones faciales y, finalmente, interactuar directamente con el usuario por medio de

los elementos del entorno. Sin embargo, debido a limitaciones con el hardware no se tomaron en consideración la manipulación directa de los objetos en los experimentos del Capítulo 10 de esta Tesis. Esto causó que las respuestas y selección de los objetos en la interacción fuese reemplazada por respuestas verbales sintéticas de un sistema TTS, y con información virtual.

En la Figura 9.9 se ilustran las respuestas asociadas a las capacidades del usuario y el agente robótico. Cabe destacar que no se incluyen los objetos directamente dentro de las capacidades, debido a que estos objetos son realmente elementos que afectan la comunicación, y no están limitados a las capacidades de manipulación física del robot.



(a) (b)

Figura 9.9: Respuestas basadas en las capacidades del usuario y el agente robótico; a) usuario; y b) robot.

9.3.5. Agentes

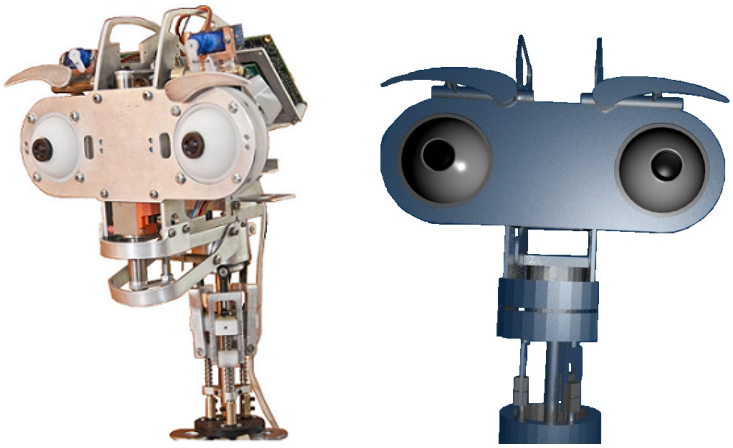
Dentro del concepto de $affordances$ emocionales, los agentes robóticos son los encargados de interactuar con los usuarios mediante el uso de información emocional y de sus propias acciones físicas. Por lo tanto, estos robots deben ser capaces de interpretar y generar elementos del lenguaje natural, tales como expresiones faciales, movimientos corporales y el propio mensaje verbal. Además, debe poseer las capacidades para detectar y seleccionar un objeto del entorno, así como también intercambiar información emocional por medio de los distintos elementos afectivos. Al tomar en consideración la necesidad de estas capacidades, se eligió la cabeza robótica Muecas, como el único agente que cumple con la mayor parte de las capacidades requeridas entre los robots disponibles. Esto es gracias, a que presenta una gran capacidad para percibir el entorno e intercambiar información, ya sea por medio de expresiones faciales o parte del lenguaje natural asociado a la cabeza del usuario. Por este motivo, se eligió este agente para su uso dentro de los experimentos del Capítulo 10, a pesar de que no presenta una capacidad de manipulación básica que le permita la elección e interacción directa con los objetos del entorno.

El agente considerado en esta Tesis Doctoral, es descrito a continuación.

9.3.5.1. Cabeza robótica Muecas

La cabeza robótica Muecas [Cid et al., 2014], es una plataforma multi-sensorial usada en IHR afectiva. En su diseño se decidió por dotarle con apariencia antropomórfica y caricaturizada, como se ilustra en la Figura 9.10. En términos técnicos, Muecas posee 12 Grados de libertad (DoF) distribuidos como sigue: 4 DoF en el cuello, 1 DoF en la boca, 3 DoF en los ojos y 4 DoF en las cejas. Además, el robot Muecas dispone de múltiples sensores de audio (micrófonos y altavoces), inerciales (compás, giróscopo y acelerómetro), vídeo y de profundidad (cámaras estéreos y sensores RGB-D). Gracias a todos estos sensores y al conjunto total de elementos móviles, el robot Muecas es capaz de interactuar con los usuarios expresando y reconocimiento emociones.

En el sistema de aprendizaje descrito en este capítulo, el robot localiza y reconoce las diferentes marcas del entorno (objetos del escenario), así como el estado emocional del usuario. Por un lado, Muecas posee un sensor RGB en el globo ocular, así como un movimiento del $Pitch$ y el Yaw para localizar las marcas asociadas a cada objeto dentro del entorno. Además, Muecas puede reconocer e imitar diferentes expresiones faciales y corporales (asociadas al cuello) que le permiten intercambiar información emocional, como se describe en el Capítulo 7. El robot también integra la implementación del algoritmo de sincronización del movimiento de la boca con el sistema TTS.



(a) (b)

Figura 9.10: a) Cabeza Robótica Muecas; b) Modelo virtual de la cabeza robótica Muecas. (Figura obtenida de la publicación [Cid et al., 2014])

9.4. Sistema de aprendizaje por imitación

El sistema de aprendizaje por imitación tiene como propósito crear relaciones entre los estados emocionales del usuario y los objetos disponibles dentro del entorno, siguiendo el concepto de $affordances$ emocionales. Al seguir este concepto, el proceso de aprendizaje se realiza por medio de una interacción donde un agente robótico entrega un determinado objeto (d_o) del entorno, a un usuario con un estado emocional definido FE, con el objetivo de cuantificar y aprender que efecto causa en la respuesta emocional del usuario E (FE_{t+1}). Este proceso saca partido del concepto de $affordances$ emocionales, para adquirir información que permita

desarrollar relaciones entre la información emocional del usuario y los elementos del entorno dentro de una interacción. En la Figura 9.11 se ilustra un experimento basado en este proceso de aprendizaje, donde se enseñan individualmente una serie de objetos reconocidos por medio de un sistema de reconocimiento de objetos basado en marcas 2D descrito en la sección 9.3.2.1. Mientras, un usuario expresa la expresión facial asociada al efecto emocional que le causa la interacción con el objeto, donde se estima el estado emocional del usuario por medio del sistema de reconocimiento de emociones basado en expresiones faciales descrito en el Capítulo 3. De esta forma, al utilizar la información obtenida de los atributos de cada objeto reconocido por medio de marcas y la información emocional de las expresiones faciales, como datos de entrada para el proceso final de aprendizaje por medio de una red bayesiana, es posible crear una relación entre un objeto y un estado emocional.

En la Figura 9.12 se ilustra una visión general de este sistema, que demuestra cómo la información relacionada a los objetos (atributos) y al usuario (información emocional) es utilizada como datos de entrada dentro de una interacción afectiva, haciendo posible realizar el proceso de aprendizaje por medio de una red bayesiana (RB), que entrena cada objeto en relación a un estado emocional específico del usuario. Tomando en consideración lo anterior, para realizar una verdadera interacción es necesario un proceso de entrenamiento con múltiples y variados objetos del entorno, que le otorguen al robot un conocimiento sobre las relaciones emocionales que crea un humano, y le permita inferir a qué estado emocional está relacionado un objeto desconocido del entorno.

Para entrenar al agente robótico se realiza una interacción controlada, que permita adquirir la información emocional del usuario y los atributos del entorno necesarios para el uso de la red bayesiana. El procedimiento se describe a continuación:

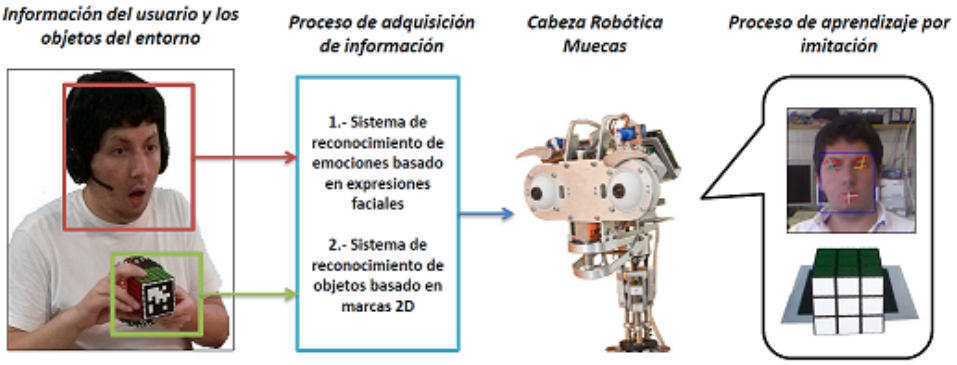


Figura 9.11: Representación del sistema de aprendizaje

1. Primero, un usuario selecciona un objeto y lo enseña al robot, mientras genera una expresión facial reconocible por el robot.

2. El robot estima el estado emocional del usuario por medio del sistema de reconocimiento de expresiones faciales, descrito en el Capítulo 3. Mientras, reconoce el objeto por medio del sistema de reconocimiento basado en marcas 2D, lo cual le proporciona la información relacionada a los atributos del objeto.

3. Al obtener la información relacionada a los objetos y el estado emocional del usuario. se realiza el proceso de aprendizaje por medio de un clasificador bayesiano.

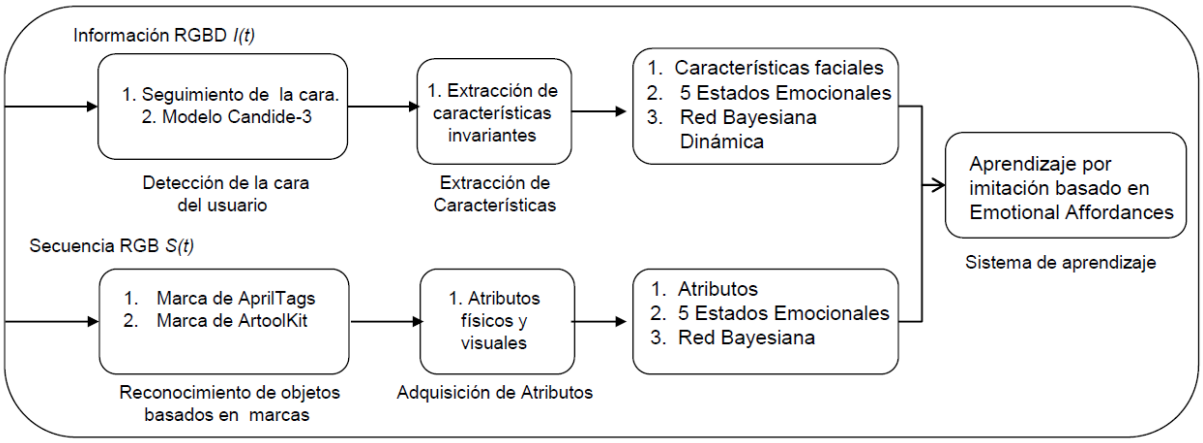


Figura 9.12: Visión general del sistema de aprendizaje basado en $affordances$ emocionales (Figura adquirida desde la publicación [Cid and Núñez, 2014])

9.4.1. Red bayesiana

A continuación, el proceso de aprendizaje utiliza la red bayesiana para crear una relación entre la información asociada al estado emocional del usuario ($E_{[Felicidad]}$, $E_{[Tristeza]}$, $E_{[Enfado]}$, $E_{[Miedo]}$, $E_{[Neutral]}$) y aquella otra relacionada con los atributos de cada uno de los objetos (d_o=AS, AC, AM, AW, ASi, $A1$), por medio del proceso de entrenamiento antes mencionado para esta red bayesiana de dos niveles, como se ilustra en la Figura 9.13. En el primer nivel se encuentra el nodo padre, que representa la reacción emocional con la que se asocia al objeto ($ES_{[Miedo]}$, $ES_{[Tristeza]}$, $ES_{[Enfado]}$, $ES_{[Felicidad]}$, $ES_{[Neutral]}$). Mientras, el segundo nivel contiene los atributos, d_o, para cada uno de estos objetos, independientes entre sí. Estos atributos constituyen las variables de entrada dentro de la red bayesiana, tal y como se describe en el Cuadro 9.1.

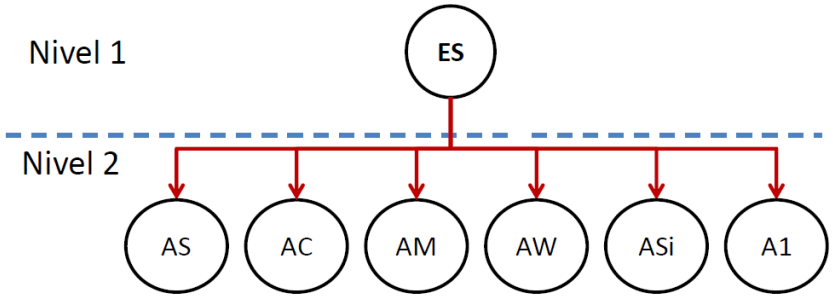


Figura 9.13: Red bayesiana.

Las variables del segundo nivel de la red bayesiana son utilizadas para calcular la distribución conjunta asociada a este clasificador, como se ilustra en la Ecuación 9.1.

Variable v	Atributos	N valores posibles
AS	Forma	12
AC	Color	19
AM	Material	9
AW	Peso	6
ASi	Tamaño	6
$A1$	Acción	7

Cuadro 9.1: Descripción de las variables de entrada de la red bayesiana, cuyos valores se relacionan con los atributos de los objetos.

$$\begin{aligned} & P(ES, AS, AC, AM, AW, ASi, A1) \\ & = P(AS, AC, AM, AW, ASi, A1 \mid ES) \cdot P(ES) \\ & = P(AS \mid ES) \cdot P(AC \mid ES) \cdot P(LE \mid ES) \cdot P(LC \mid ES) \\ & \cdot P(CB \mid ES) \cdot P(MF \mid ES) \cdot P(MA \mid ES) \cdot P(ES) \end{aligned} \tag{9.1}$$

9.5. Modelos de comportamientos afectivos

En esta sección se presentan los modelos de comportamientos afectivos desarrollados durante la Tesis Doctoral. Estos modelos hacen uso de maquinas de estado, de forma que analizan y predicen las respuestas emocionales del usuario en la interacción de acuerdo a un conjunto de opciones disponibles que el robot ha aprendido previamente. De esta forma, a través de diferentes comportamientos del robot, es posible orientar el estado emocional del usuario hacia un estado específico, pasando por medio de varios estados transitorios. Para ello, el robot hace uso del conocimiento adquirido durante el aprendizaje de las distintas relaciones existentes entre los objetos en el escenario, los estados afectivos y las reacciones esperadas en un usuario, tal y como se describen en las $affordances$ emocionales. Con todo lo anterior, los modelos de comportamientos afectivos desarrollados en esta Tesis tienen como objetivo afectar el estado emocional del usuario hasta obtener un estado con un tipo específico de valencia, por medio del modelo circumplejo de Russell [Russell, 1980] (Figura 9.14).

En relación al diseño de estos modelos de comportamientos afectivos, se tomó en consideración varios estudios que presentan sistemas similares en la literatura [Prado et al., 2011] y [Breazeal et al., 2008], los cuales describen una relación entre el comportamiento del robot y las emociones transmitidas en una interacción. Por otro lado, con respecto al uso de sistemas de aprendizaje basados en redes bayesianas dinámicas relacionados a sistemas de modelado de comportamiento, no es algo nuevo, ya que existen trabajos en la literatura como el descrito en [Huang and Mutlu, 2014], que demuestran la importancia del uso de modelos de comportamiento para alcanzar un objetivo específico basado en una IHR. Al considerar cómo muchos de estos trabajos utilizan sistemas de aprendizaje basados en redes bayesianas y presentan resultados interesantes, se decidió seguir esta estructura y realizar la implementación de varios modelos de comportamientos afectivos como maquinas de estado, dentro de una interacción real entre un humano y un robot según el concepto de $affordances$ emocionales, tal y como se

ilustra en la Figura 9.15. Durante la interacción, el objetivo del robot es utilizar su conocimiento respecto a los diferentes elementos afectivos y su relación con los estados emocionales, con el fin de obtener el efecto esperado en el usuario. El efecto esperado queda definido por el propio modelo de comportamiento afectivo.

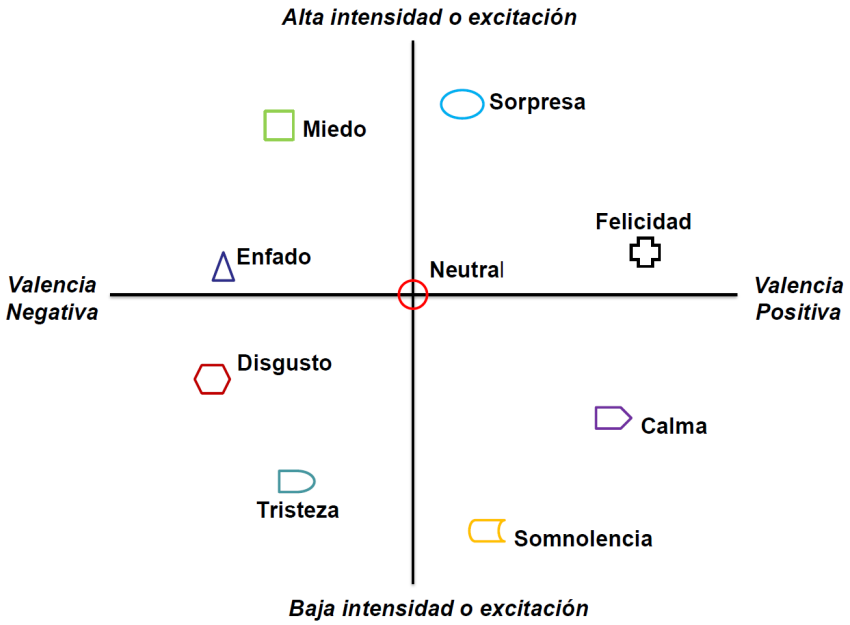


Figura 9.14: Modelo Circumplejo del afecto de Russell [Russell, 1980].

Los modelos de comportamientos afectivos implementados en este trabajo se pueden dividir en tres enfoques diferentes, los cuales hacen uso de la valencia de una emoción para cambiar gradualmente el estado inicial del usuario a un estado con valencia Positiva, Negativa o Neutral. A continuación se describen estos modelos:

1. **Empático** o **Positivo**: este modelo afecta el estado emocional del usuario por medio de los objetos del entorno y las expresiones faciales del robot, cambiando el estado emocional gradualmente en términos de intensidad y posteriormente valencia, hasta orientar al usuario a un estado emocional asociado a la Felicidad.

2. **Antipático** o **Negativo**: este modelo trata igualmente de modificar el estado emocional del usuario, cambiando en este caso el estado emocional gradualmente en términos de intensidad y luego valencia hacia un estado asociado al Miedo.

3. **Neutral**: este modelo afecta al estado afectivo del interlocutor, gradualmente en términos de intensidad y posteriormente valencia, hasta orientarlo a un estado Neutral.

La Figura 9.15 presenta un esquema de una posible IHR basada en los modelos de comportamiento anteriormente descritos. En la misma, el modelo de comportamiento seleccionado para el robot (positivo, negativo o neutral), MC, busca un cambio en el estado emocional del humano (efecto), que debe ser el observado por el robot tras operar con los objetos del entorno y mostrar una reacción emocional. En la Figura 9.16 se muestran los tres modelos de comportamiento asociados al robot Muecas. La Figura 9.16a representa los cambios graduales en los

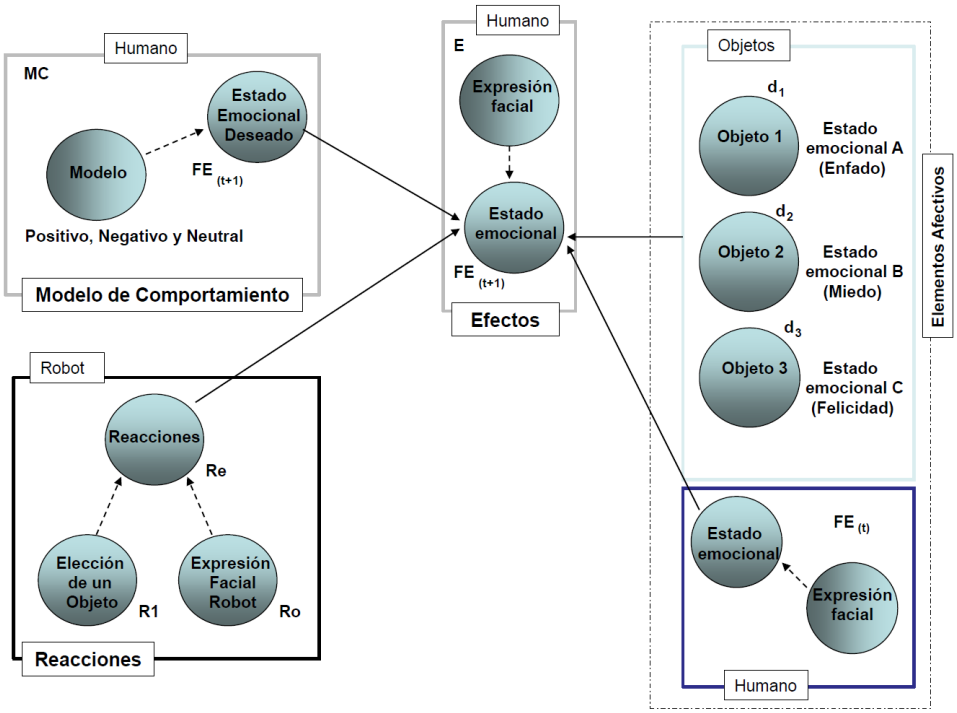


Figura 9.15: Representación de una interacción basada en modelos de comportamientos afectivos.

estados emocionales para llegar al estado final de felicidad, según un modelo de comportamiento positivo. De forma similar, las Figuras 9.16b-c muestran el funcionamiento de los modelos para los comportamientos negativo y neutral, respectivamente.

9.5.1. Maquinas de estado

La implementación de los modelos de comportamientos afectivos se realizó por medio de maquinas de estado, tal y como se muestra en la Figura 9.17. Dentro de la literatura, el uso de maquinas de estado se presenta como una herramienta con gran potencial para controlar las respuestas emocionales de un robot durante interacciones reales con humanos, donde exista un número limitado de estados emocionales [Schulte et al., 1999].

Por este motivo, en esta Tesis Doctoral, las reacciones emocionales del robot y las interacciones con los objetos del entorno que persiguen modificar de forma gradual el estado emocional del humano en una IHR se formalizan según un modelo de máquinas de estado. La transición entre un estado y otro se realiza en función de los objetos del escenario, así como los propios estados emocionales del robot, todo ello condicionado por el aprendizaje previo de las $affordances$ emocionales. Las máquinas de estado aquí presentadas disponen de una serie de condiciones sobre qué objeto seleccionar del escenario afectivo, dado que los objetos que el robot dispone para la interacción son aleatorios y no cubren todos los estados emocionales, llegando incluso a repetirse. Las tres condiciones que regulan los criterios de selección sobre los elementos del entorno son descritas a continuación:

- Condición 1: (1) es el objeto asociado al estado emocional que se desea conseguir.

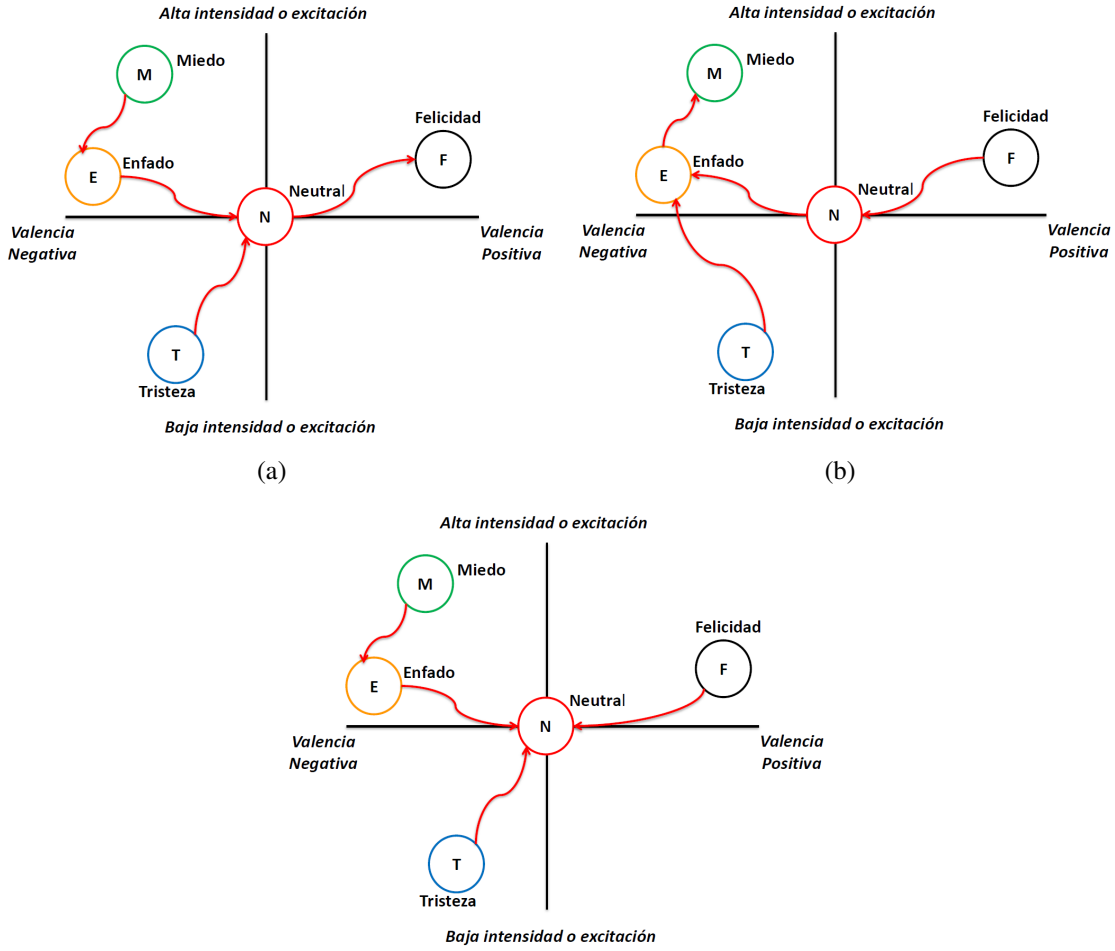


Figura 9.16: Representación visual de los modelos de comportamientos afectivos, basados en el modelo circumplejo de Russell : a) Positivo; b) Negativo; y c) Neutral.

- Condición 2: (2) es el objeto elegido cuando el objeto asociado al estado deseado (1), no se encuentra entre los elementos disponibles.

- Condición 3: (3) es el objeto elegido en base a que las opciones anteriores, ni (1) ni (2) se encontraban entre los objetos. No obstante, el resultado suele ser imprevisible causando que llegue a cualquier tipo de estado emocional. Esta última condición es causada por el limitado número de estados emocionales en el estudio, que no permite hacer un buen uso de la información emocional del usuario dentro de los modelos de comportamientos afectivos.

Finalmente, en la Figura 9.17 se ilustran las maquinas de estado para cada modelo de comportamiento. En la Figura 9.17a se presenta la máquina de estado asociado al comportamiento positivo del robot. Cada una de las condiciones y su efecto sobre la transición de los estados es representada en la figura, entre paréntesis. Al final, tras una serie de interacciones, el robot consigue llevar al usuario a un estado de felicidad. Las Figuras 9.17b-c corresponden a los modelos de comportamiento negativo y neutral, respectivamente. Según se aprecia en la figura, el

robot elige los objetos esperando tener un efecto específico en el estado emocional del usuario, dado que cada objeto está asociado a un estado emocional concreto. En la figura, O_M, O_E, O_F, O_T y O_N representan objetos asociados a los estados emocionales de miedo, enfado, felicidad, tristeza y neutral, respectivamente.

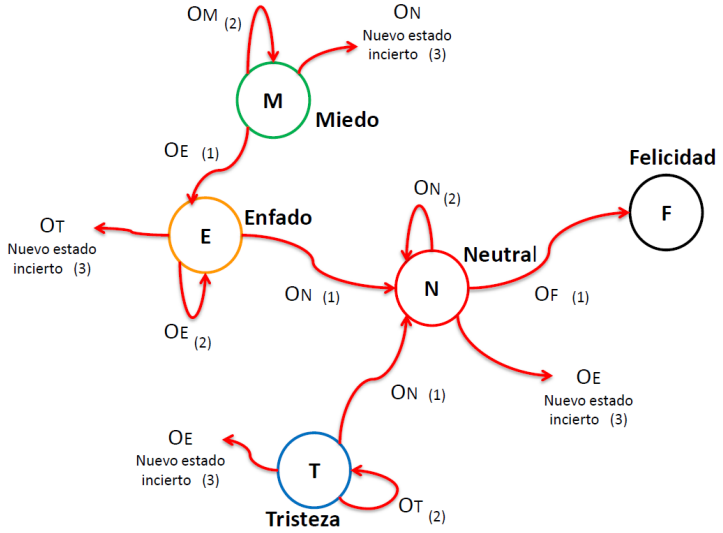
9.6. Escenario afectivo

Una de las necesidades durante el desarrollo y la evaluación del sistema de aprendizaje de comportamientos afectivos es la de disponer de un escenario adaptado a todas las capacidades del robot Muecas. Este entorno para la IHR debe permitir al robot percibir claramente los diferentes elementos afectivos, esto es, los objetos físicos y la información relacionada con el lenguaje natural del usuario. A su vez, el escenario descrito en esta sección debe considerar las limitaciones propias del robot, entre ellas aquellas relacionadas al hecho de que no posee capacidades de navegación ni de manipulación. Por este motivo, se desarrolló un escenario afectivo dentro de $RoboHome$, un laboratorio de pruebas del grupo de Robótica y Visión Artificial de la Universidad de Extremadura. Este laboratorio fue adaptado para las pruebas y experimentos, con el objetivo de cubrir todas las necesidades de Muecas, lo que permitió que el robot pudiera analizar todos los objetos de la habitación e interactuar fácilmente con el usuario. Dado que es un entorno controlado, los cambios y reacciones del usuario pueden ser cuantificados de forma externa por medio de métodos supervisados. Junto a este escenario físico real, se disponía de una reproducción virtual del mismo a partir del simulador $RCInnermodelSimulator$, como se ilustra en la Figura 9.18

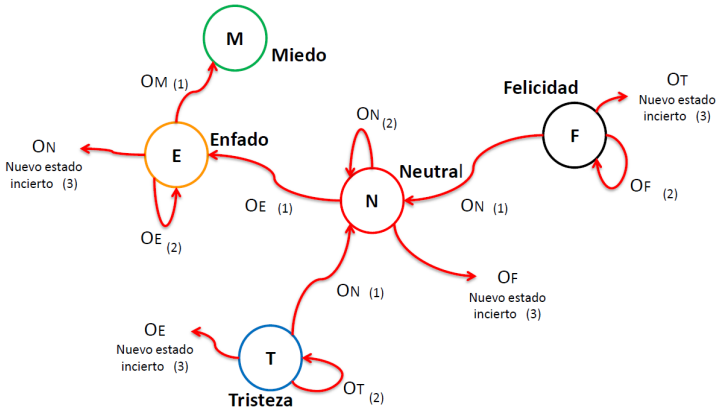
9.7. Conclusiones

En este trabajo se describió un sistema de aprendizaje de comportamientos afectivos basados en las $affordances$ emocionales, con la capacidad de orientar el estado emocional del usuario durante una IHR. Para lograr esto, se presentaron los dos procesos principales de este sistema. El primero es una estructura de aprendizaje basada en $affordances$ emocionales, la cual utiliza un enfoque bayesiano para representar una relación entre los diferentes atributos de un objeto y el estado emocional del usuario, todo ello a través de diferentes sistemas de reconocimiento. Por un lado, para la captura de información emocional del usuario, el sistema de aprendizaje utiliza uno de los sistemas de reconocimiento de expresiones faciales descrito en esta Tesis Doctoral. Por otro lado, para el reconocimiento de objetos se hace uso de un sistema de reconocimiento de objetos basado en marcas 2D, que adquiere la información relacionada a los atributos específicos de cada objeto, evitando limitaciones como la cantidad y el tipo de estos objetos.

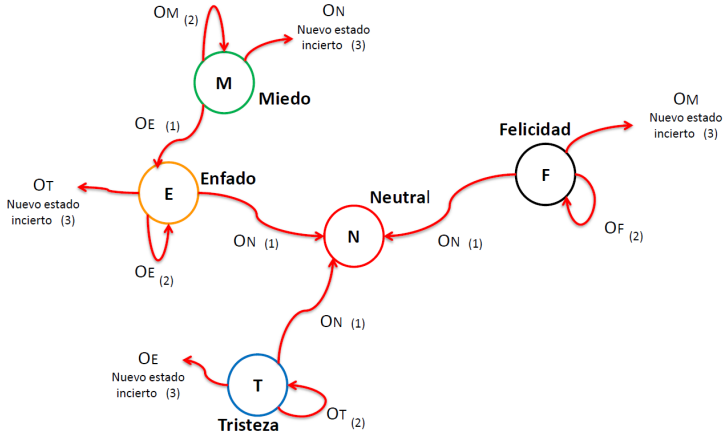
El segundo de los procesos hace referencia a la implementación de diferentes modelos de comportamientos emocionales para un robot durante una IHR afectiva. Estos modelos están implementados en maquinas de estado que analizan y aprenden de las respuestas emocionales del usuario en su interacción con determinados objetos del escenario. El objetivo final de estos modelos es orientar el estado emocional del usuario a un estado especifico mediante el uso de elementos afectivos, como los objetos presentes y las propias reacciones del robot durante la comunicación.



(a)



(b)



(c)

Figura 9.17: Maquinas de estado de los modelos de comportamientos afectivos, según el modelo circumplejo de Russell : a) Positivo; b) Negativo; y c) Neutral.

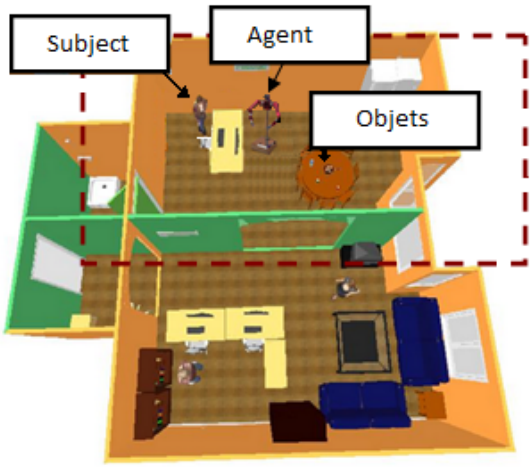


Figura 9.18: Escenario controlado para una interacción afectiva. (Figura adquirida de la publicación [Cid and Núñez, 2014])

Capítulo 10

Resultados experimentales del sistema de aprendizaje

A lo largo del capítulo se describen los experimentos realizados, así como los resultados obtenidos en la evaluación del rendimiento del sistema de aprendizaje basado en $affordances$ emocionales. El desarrollo de las pruebas se plantea como una interacción real entre un humano y un robot en un escenario controlado, a lo largo de la cual se llevan a cabo cada una de las fases descritas en el Capítulo 9. Tal y como fue presentado en el capítulo anterior, el robot está dotado con la capacidad de reconocer el estado emocional del interlocutor, así como también con la capacidad de interactuar con objetos del entorno. El robot, durante la interacción, ha de aprender las relaciones existentes entre algunos objetos del entorno y las propias emociones del usuario, esto es, los estímulos, y las posibles reacciones que provocan. Este aprendizaje emocional durante la interacción, unido al modelo de comportamiento del robot, permite llevar a cabo una IHR afectiva completa, entre el usuario no entrenado y el robot autónomo.

La evaluación del sistema de aprendizaje propuesto sigue cada una de las fases lógicas en el desarrollo del sistema de aprendizaje. En primer lugar se evalúa el sistema de reconocimiento de emociones utilizado, luego el aprendizaje en sí, y finalmente una interacción real entre un robot y un humano siguiendo los modelos de comportamiento del robot. Más concretamente, podemos dividir el desarrollo experimental de este trabajo en las siguientes fases:

1. **Evaluación de sistema reconocimiento de emociones basado en expresiones faciales:** como fue descrito en el Capítulo 9, la adquisición de la información emocional durante la IHR es necesaria durante el aprendizaje de las $affordances$ emocionales. Por ello, es imprescindible una evaluación previa de este sistema de reconocimiento de emociones con idea de comprobar la precisión del mismo en el escenario afectivo utilizado. En la primera parte de esta Tesis Doctoral se describieron varios sistemas de reconocimiento de emociones que trabajan con fuentes de información diferentes (expresiones faciales, análisis del habla o del lenguaje corporal). En el Cuadro 10.1 se resume la precisión de cada uno de estos sistemas, destacando la ventaja del uso de expresiones faciales para reconocer una emoción. En el sistema de aprendizaje evaluado en esta sección, por simplificación, se hace uso del algoritmo basado en la malla $Candide-3$.

2. **Evaluación del sistema de aprendizaje basado en imitación:** el segundo experimento realiza una evaluación del sistema de aprendizaje por imitación presentado en esta Tesis Doctoral. En esta parte se evalúa globalmente el sistema, esto es, se realiza un aprendizaje

completo de las $affordances$ emocionales durante interacciones con un usuario humano en un entorno afectivo controlado.

3. **Evaluación de los modelos de comportamiento en interacciones afectivas:** este último experimento evalúa de forma practica el funcionamiento de los modelos de comportamiento descritos en el Capítulo 9. A lo largo de esta prueba se pretende modificar el estado emocional de un usuario humano según un enfoque determinado por parte del robot. Este experimento requiere de una IHR real, dentro del contexto de las $affordances$ emocionales, y por ello se hace uso de todos los elementos afectivos del escenario, junto con el sistema de aprendizaje por imitación, ya entrenado.

Sistema de reconocimiento	Precisión
Sistema basado en $Candide$-3 [Cid et al., 2014]	**94** %
Sistema basado en $Gabor$ [Cid et al., 2013b]	93 %
Sistema basado en $Voz\ Humana$ [Cid et al., 2014]	77 %
Sistema basado en $Lenguaje\ Corporal$ [Doblado et al., 2013]	74 %

Cuadro 10.1: Cuadro comparativo entre los diferentes sistemas de reconocimiento de emociones descritos en esta Tesis Doctoral.

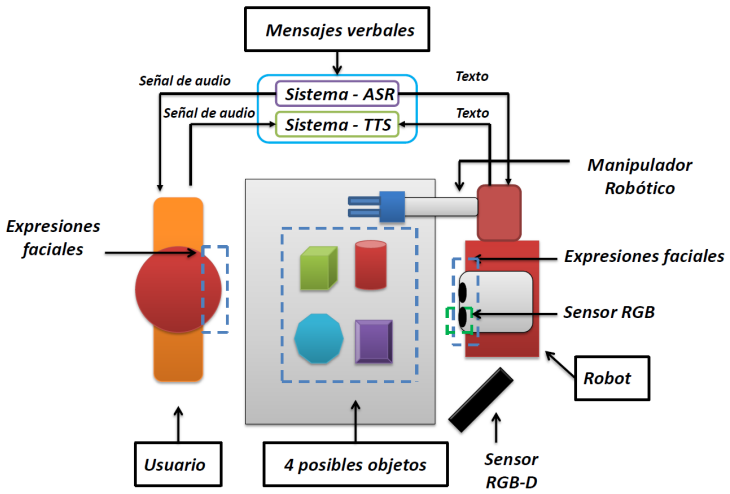
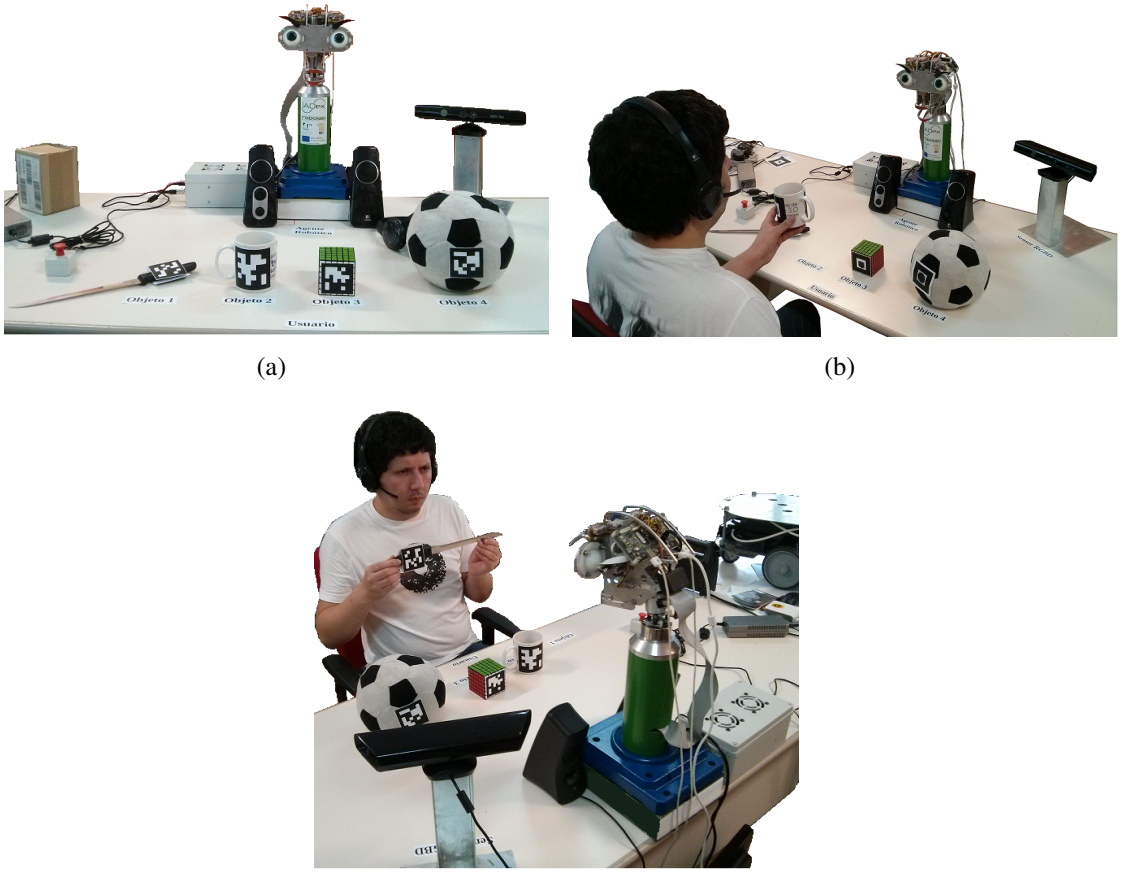


Figura 10.1: Representación de los elementos utilizados en los experimentos basados en $affordances$ emocionales, dentro de un escenario afectivo.

Todos los experimentos en este capítulo se realizaron en un mismo escenario afectivo, representado en la Figura 10.1. Como se observa en la figura, en el entorno utilizado para las pruebas, el robot se encuentra ubicado justo enfrente del interlocutor, separados por una mesa que sirve de espacio de trabajo y a una distancia tal que se asegure un correcto funcionamiento del algoritmo de reconocimiento de emociones. Sobre dicha mesa, el robot dispone de una serie de objetos de los que se quiere, o bien aprender sus $affordances$ emocionales por imitación, o bien afectar el estado emocional del usuario. La Figura 10.2 muestra el escenario real utilizado

en los experimentos. En la Figura 10.2a se observan los cuatro objetos utilizados en las pruebas, así como el sistema de percepción visual del robot. Las Figuras 10.2b-c muestran diferentes momentos durante el aprendizaje.

A continuación se describen los elementos seleccionados para el escenario real y su función dentro de los experimentos:



(a) (b)

(c)

Figura 10.2: Escenario afectivo real dentro de los experimentos

- *Usuario:* el usuario proporciona, por un lado, la información emocional durante el aprendizaje, expresa emociones y manipula los objetos durante el mismo; por otro lado, interactúa con el robot durante la evaluación del sistema completo.

- *Robot:* el agente utilizado durante los experimentos ha sido la cabeza robótica Muecas, capaz de generar reacciones por medio de expresiones faciales, movimientos corporales y mensajes verbales.

- *Objetos:* son 19 los objetos disponibles en el entorno durante los experimentos. El total de estos objetos se recogen en los Apéndice B.1, B.2 y B.3, donde se incluye no sólo la descripción y marcas utilizadas, sino también los modelos 3D empleados en las simulaciones y pruebas iniciales con realidad aumentada ($OSGART$).

- *Marcas:* las marcas utilizadas durante los experimentos de este escenario siguen los patrones de las librerías $ARToolKit$ y $AprilTags$ (Ver Apéndice B.2). El uso de este sistema de marcas permite la localización y el reconocimiento de los objetos con sus respectivos atributos de forma rápida y simple, sin complicar el sistema de reconocimiento de objetos que queda fuera del alcance de esta Tesis Doctoral. En la Figura 10.3 se ilustran dos de los objetos utilizados en las pruebas, una taza y un cubo de Rubik, Figuras 10.3a-b y Figuras 10.3c-d, respectivamente. Las marcas que aparecen sobre el objeto corresponden a la librería $AprilTags$ (Figuras 10.3a y 10.3c) y $ARToolKit$ (Figuras 10.3b y 10.3d).

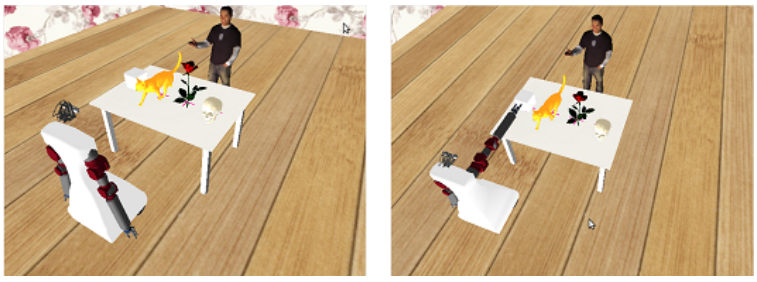
- *Sensores:* para adquirir la información del usuario y del entorno, el agente robótico dispone de dos sensores diferentes. Muecas presenta un primer sensor RGB (cámara $Firewire$) dentro del globo ocular. Mientras, un segundo sensor de tipo RGB-D se encuentra ubicado junto a la plataforma. En concreto, el sensor RGB-D se encuentra a una distancia de 40 cm desplazado a la izquierda, manteniendo la misma altura y distancia hacia el usuario que si estuviera encima del robot. Gracias a este desplazamiento, se dispone de una mayor libertad de movimiento por parte del robot en la comunicación con el usuario.

Es importante describir cómo se realiza la comunicación verbal durante la interacción. Para la transmisión de los mensajes desde el robot al usuario se utilizó la voz sintética generada por el TTS de $Google$, usando el idioma español. De forma similar, para recibir los mensajes de voz del usuario, el robot está equipado con el sistema ASR de $Google$ (ver Figura 10.1).

Finalmente, debido a que el robot Muecas no posee capacidades de manipulación, se ha utilizado un modelo de un robot más complejo que incluye la cabeza robótica Muecas, así como brazos y manos para interactuar con objetos. Este entorno simulado se construye sobre RoboComp a través de la herramienta $RCInnerModelSimulator$ [Manso, 2012]. Un ejemplo del escenario virtual, réplica del real, se ilustra en la Figura 10.4

(a) (b) (c) (d)

Figura 10.3: Representación de dos de los objetos utilizados en los experimentos; a) Taza con la marca de $AprilTags$; y b) taza con la marca de $ARToolKit$; c) cubo de Rubik con la marca de $AprilTags$; y d) cubo de Rubik con la marca de $ARToolKit$.



(a) (b)

Figura 10.4: Imágenes del simulador, con los elementos del escenario real representados virtualmente (usuario, objetos y robot).

10.1. Evaluación del sistema de reconocimiento de emociones basado en expresiones faciales

Como se ha comentado anteriormente, en primer lugar se procede a evaluar la precisión y robustez del sistema de reconocimiento de expresiones faciales descrito en el Capítulo 3 para este escenario controlado. Para ello, se contó con la participación en el experimento de diez usuarios con diferente género, edad y características faciales. Cada uno de ellos realizó diez secuencias aleatorias de expresiones faciales, pasando por todos los estados emocionales.

Durante la interacción con el robot, el usuario no tiene contacto alguno con otros humanos, es decir, no existe supervisión por un experto, y toda la comunicación y evaluación se realiza mediante un procedimiento guiado por el propio robot. Muecas, en cada fase de la evaluación, genera instrucciones que se convierten en voz sintética acompañada de movimientos sincronizados de la boca. En estas instrucciones, el robot solicita al usuario la realización de expresiones faciales de forma aleatoria, y una vez éste la realiza, estima el estado emocional del sujeto. En caso de acierto, el usuario procede a etiquetar el experimento como correcto, y en caso contrario, como error (aportando información también del tipo de error). A continuación, se muestran las instrucciones de las que se compone este experimento. Las instrucciones son mensajes verbales generados por el robot con su sistema TTS, y en cada caso se describen las acciones realizadas por el usuario o el robot por medio de unos paréntesis ().

- Muecas: ”La prueba ha comenzado, por favor exprese una expresión facial.”
- Usuario: (El usuario realiza una determinada expresión facial, tratando de mostrar una emoción concreta)
- Muecas: (El robot, al reconocer la expresión facial del usuario, la imita mediante el movimiento de sus componentes mecánicos)
- Usuario: (Evalúa el acierto/error en el reconocimiento de la expresión facial)
- Muecas: ”Por favor, expresa una nueva expresión facial, diferente de las realizadas anteriormente.”
- Usuario: (El usuario realiza una determinada expresión facial, tratando de mostrar una emoción concreta)

- Muecas: (El robot, al reconocer la expresión facial del usuario, la imita mediante el movimiento de sus componentes mecánicos)
- Usuario: (Evalúa el acierto/error en el reconocimiento de la expresión facial)
- Y así de forma continuada hasta completar el experimento.

La Figura 10.5 describe el proceso llevado a cabo durante el desarrollo del experimento, desde el instante inicial del mismo y distinguiendo entre las acciones llevadas a cabo por el usuario humano y por el robot Muecas. En la parte superior de la imagen se distinguen las acciones llevadas a cabo por el robot tal y como se describieron anteriormente. De igual forma, la parte inferior representa las acciones realizadas por el usuario durante la interacción.

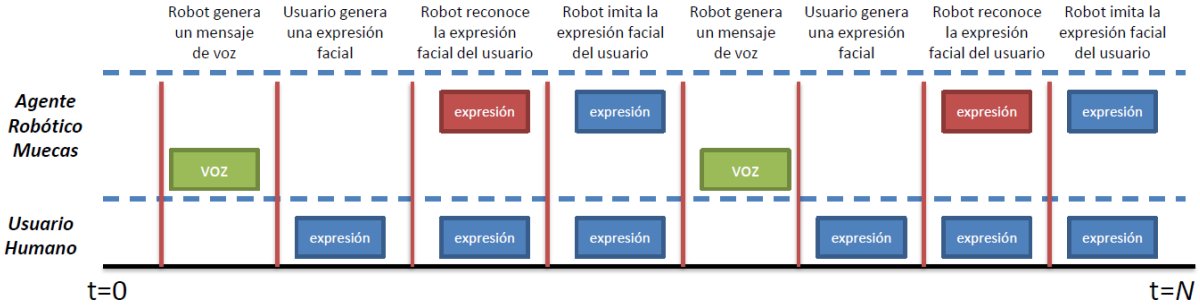


Figura 10.5: Representación de las instrucciones relacionadas al reconocimiento de expresiones faciales.

En los Cuadros 10.3 y 10.2 se resumen las diferencias en cuanto a la precisión y el porcentaje de errores en la estimación del estado emocional del usuario en el experimento actual y el realizado en el Capítulo 3, respectivamente. Del análisis de las tablas, se observa cómo las ventajas de tener un entorno controlado (condiciones de luz y espacio fijas o la supervisión del entorno, entre otras) permite una mejora sustancial de los resultados.

Test	Tristeza	Felicidad	Miedo	Enfado	Neutral
P_{FE} (Cap. 3)	90 %	98 %	95 %	95 %	92 %
P_{FE}	**92** %	**98** %	**96** %	**95** %	**94** %

Cuadro 10.2: Tabla Comparativa de la precisión del sistema de reconocimiento de expresiones faciales. En la tabla se representan los datos extraídos del Capítulo 3, P_{FE}, y aquellos obtenidos en esta sección.

10.2. Evaluación del sistema de aprendizaje basado en imitación

El segundo experimento persigue la evaluación del sistema de aprendizaje descrito en el Capítulo 9. En esta prueba se mantuvo el mismo número de usuarios que en el experimento

Errores	Tristeza	Felicidad	Miedo	Enfado	Neutral
Errores (Cap. 3)	6 %	2 %	2 %	3 %	5 %
Errores	4 %	2 %	1 %	3 %	4 %

Cuadro 10.3: Tabla comparativa de los porcentajes de error del sistema de reconocimiento facial. En la tabla se representan los datos extraídos del Capítulo 3, y aquellos obtenidos en esta sección.

anterior. En este caso, cada participante en la IHR realiza expresiones faciales e interactúa con uno de los elementos del entorno. En los tests se hace uso de los 19 objetos presentados en este capítulo, donde únicamente tres quintas partes de los mismos han sido utilizados en el entrenamiento. Para la localización y el reconocimiento de los objetos, se utiliza la librería $AprilTags$.

El primer paso consiste en el entrenamiento del sistema con los objetos seleccionados para tal fin. El criterio seguido para la selección de un objeto u otro es disponer de una representación balanceada de pares objeto/emoción, de forma que se disponga de suficiente información asociada a cada una de las cinco emociones a partir del total de objetos del entrenamiento. En este paso, el usuario vuelve a no tener contacto con otros humanos durante la interacción y de nuevo todo el procedimiento es guiado por el robot Muecas. El procedimiento seguido se describe a continuación (de nuevo se presentan los mensajes generados por el robot y las acciones llevadas a cabo por el usuario o el robot, representadas por medio de paréntesis ()).

- Muecas: ”El entrenamiento ha comenzado, por favor elige un objeto.”
- Usuario: (El usuario escoge un objeto de los existentes en la mesa, éste lleva asociado una marca $AprilTag$)
- Muecas: (El robot reconoce el objeto).
- Muecas: ”A continuación, por favor, exprese qué emoción le genera ese objeto.”
- Usuario: (El usuario expresa la emoción)
- Muecas: (El robot, al reconocer la expresión facial del usuario, la imita mediante el movimiento de sus componentes mecánicos. Los datos se incluyen en el sistema como válido)
- Se reinicia el procedimiento, que se repite hasta completar el entrenamiento.

La Figura 10.6 muestra el procedimiento descrito según su evolución en el tiempo. La figura muestra los instantes donde se generan mensajes de voz por parte del robot y donde se ejecutan las acciones. En la parte superior se reflejan aquellas acciones relacionadas con el robot Muecas, mientras que en la parte inferior se ilustran las llevadas a cabo por el usuario.

Tras finalizar el entrenamiento del sistema de aprendizaje, estos mismos objetos se mezclan con el resto, y comienza el experimento en sí. Aleatoriamente, del total de objetos se seleccionan cuatro de ellos que se sitúan en la mesa de trabajo, justo entre ambos participantes. A continuación, el usuario decide el estado emocional inicial en el que se encuentra, generando la expresión facial correspondiente. Como se ha comentado al inicio de la sección, el robot Muecas

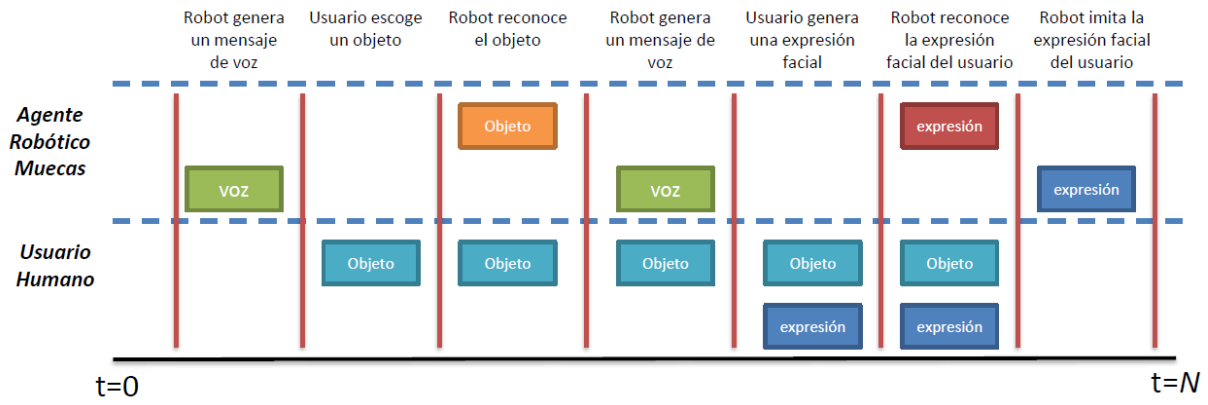


Figura 10.6: Representación de las instrucciones relacionadas al sistema de aprendizaje basado en imitación.

en esta interacción sigue un modelo de comportamiento Neutral. Como tal, su principal objetivo es hacer uso del conocimiento del estado emocional del interlocutor y de los objetos sobre la mesa ($affordances$ emocionales) para conseguir una transición del estado emocional del usuario hasta el estado Neutro. Para ello, el robot toma uno de los objetos y se lo presenta al usuario, junto con una expresión facial. La elección de un objeto u otro depende del aprendizaje de las $affordances$ emocionales de los objetos durante el entrenamiento. Si el aprendizaje se ha realizado correctamente, el robot selecciona el objeto más parecido (presenta unas $affordances$ emocionales similares) a aquel que conseguía llevar al usuario a un estado Neutro durante el entrenamiento. Tras la interacción, el usuario evalúa el éxito o error del entrenamiento, de forma similar al experimento presentado en la sección anterior.

Los resultados del experimento se ilustran en el Cuadro 10.4. En esta tabla se representa la probabilidad de un correcto aprendizaje, P_l, evaluado como el porcentaje de éxito para llevar al usuario desde el estado emocional mostrado, al Neutral. Como se observa en la figura, el aprendizaje de las $affordances$ emocionales no es igual para todas las emociones. En el caso de los estados Tristeza y Felicidad, las tasas de éxito son superiores al resto, dando a entender que la selección de objetos utilizada facilita este tipo de resultados. El estado Neutral, por contra, tiene la probabilidad de aprendizaje correcto más baja, sin lugar a dudas condicionado por la selección de objetos en el aprendizaje y la dificultad que conlleva asociar un objeto a un estado Neutral.

Pruebas	Probabilidad de correcto aprendizaje (P_l)
Tristeza	83 %
Felicidad	78 %
Miedo	71 %
Enfado	67 %
Neutral	59 %

Cuadro 10.4: Probabilidad de correcto aprendizaje basado en imitación

10.3. Evaluación de los modelos de comportamiento dentro de IHR afectivas

Los modelos de comportamiento son métodos, en general basados en maquinas de estados, que tienen como objetivo cambiar el estado emocional del usuario durante una IHR hacia una emoción especifica. Para la evaluación de los modelos de comportamiento se realizaron diferentes pruebas en el escenario afectivo descrito en este capítulo. Las pruebas se llevaron a cabo de forma similar, con el robot ubicado enfrente del usuario, separándolos una mesa sobre la que se sitúan aleatoriamente cuatro objetos. El usuario elige una emoción de inicio y genera para ello una expresión facial. A continuación, dependiendo del modelo de comportamiento seleccionado para el robot (positivo, negativo o neutro, según el Capítulo 9), comienza una interacción entre Muecas y el usuario hasta alcanzar, en caso de éxito, la emoción deseada.

En el experimento presentado en esta sección se utilizaron las siguientes condiciones para las maquina de estados (ver Figura 10.8):

- Condición 1: (1) es el objeto asociado al estado emocional que se desea conseguir.
- Condición 2: (2) es el objeto elegido cuando el objeto que debería estar asociado al estado deseado (1), no se encuentra entre los elementos disponibles.
- Condición 3: (3) Es el objeto elegido cuando, ni (1) ni (2) se encuentran entre los objetos sobre la mesa. No obstante, el resultado suele ser imprevisible causando que llegue a cualquier tipo de estado emocional.

De nuevo la interacción es guiada por el robot, sin supervisión externa. Este usuario es igualmente el encargado de evaluar el resultado final del experimento. Debido a que Muecas no dispone de la capacidad de manipular objetos reales, se muestra en una pantalla, a la vista del usuario, el entorno virtual con la representación interna del escenario afectivo, junto al modelo de robot y usuario. A su vez, Muecas genera un mensaje de voz con la información necesaria. Esto causa menos influencia en el usuario y disminuye el nivel de interacción dentro del experimento, pero permite comprobar cómo sería la interacción real final en un sistema completo, con un robot equipado con brazos y la capacidad de agarrar objetos de una mesa (problema nada trivial). A continuación, se describen los pasos que se realizan en la evaluación de los modelos de comportamiento durante la interacción humano-robot.

- Muecas: ”Por favor, elija un modelo de comportamiento para el robot entre los posibles (positivo, negativo o neutral).”
- Usuario: (El usuario escoge un modelo de comportamiento)
- Muecas: ”A continuación exprese su estado emocional por medio de una expresión facial.”
- Usuario: (El usuario expresa la emoción)
- Muecas: (El robot reconoce el estado emocional del usuario, busca y reconoce los elementos del entorno, y determina por medio de la maquina de estado, qué objeto es elegido de acuerdo al modelo de comportamiento y el estado emocional actual del usuario. Este objeto se muestra al usuario)

- Muecas: ”Elijo este objeto”
- Muecas: (El robot realiza la expresión facial asociada al estado emocional relacionado al objeto, a su vez, actualiza el modelo virtual del escenario afectivo).
- Usuario: (El usuario expresa la nueva emoción)
- Muecas: (El robot reconoce el estado emocional del usuario).
- Se repite el proceso hasta obtener el estado emocional deseado en el usuario, siendo éste el encargado de evaluar el éxito o error del comportamiento del robot.

La Figura 10.7 muestra el procedimiento llevado a cabo durante la evaluación. En la figura se muestra tanto las acciones llevadas a cabo por el robot como por el usuario humano, así como el intercambio de mensajes.

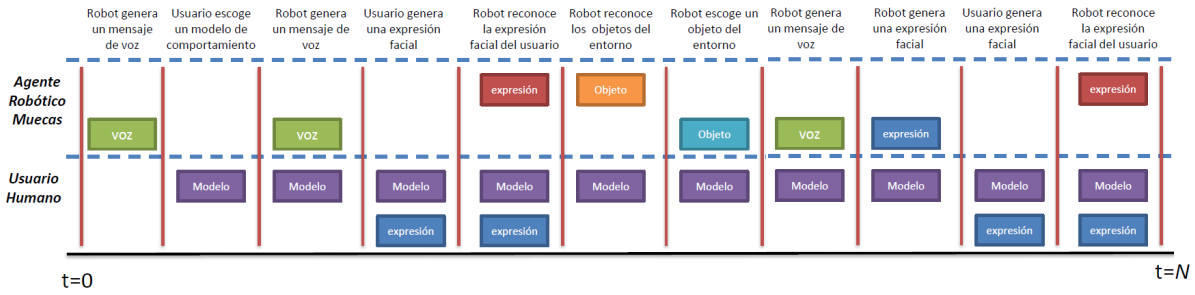


Figura 10.7: Representación de las instrucciones relacionadas a los modelos de comportamiento.

En el Cuadro 10.5 se muestran los resultados del uso de los modelos de comportamiento, dentro de un escenario IHR con $affordances$ emocionales. La tabla representa la probabilidad de un correcto funcionamiento del modelo de comportamiento seleccionado, y para ello mide el resultado final del experimento para cada uno de los modelos (positivo, negativo o neutro). A continuación se explican los posibles resultados recogidos en la tabla:

Pruebas	Positivo	Negativo	Neutral
Caso 1: respuesta correcta	47 %	59 %	75 %
Caso 2: resultado imprevisible	31 %	12 %	9 %
Caso 3: resultado erróneo	22 %	29 %	16 %

Cuadro 10.5: Probabilidad de correcto funcionamiento de los modelos de comportamiento.

1. Caso 1: Al seleccionar el objeto por parte del robot, el estado emocional del usuario cambia al estado esperado dentro de la maquina de estado asociada al modelo de comportamiento elegido. Por ejemplo, si se eligen las opciones de la condición (1) o (2).

2. Caso 2: Al seleccionar la regla (3), al no encontrarse un objeto asociado a un estado emocional necesario da lugar a un estado imprevisible.

3. Caso 3: Al seleccionar el objeto correcto por medio de la maquina de estado de un modelo de comportamiento especifico, termina presentando un estado emocional incorrecto por parte del usuario. Este tipo de resultado erróneo suele ser causado por problemas o errores en el aprendizaje.

Los resultados del Cuadro 10.5 demuestran la dificultad del problema, con unas probabilidades de alcanzar el objetivo planteado, en algunos casos, inferior al cincuenta por ciento. En el caso del modelo de comportamiento positivo, su baja respuesta se debe principalmente a las pocos estados emocionales asociados a este nivel de valencia (positivo), lo cual causa que la transición del estado emocional del usuario, desde otros estados al estado Felicidad siempre necesite el paso por el estado neutral (ver Figura 10.8), complicando el problema al sólo existir cuatro objetos durante la interacción.

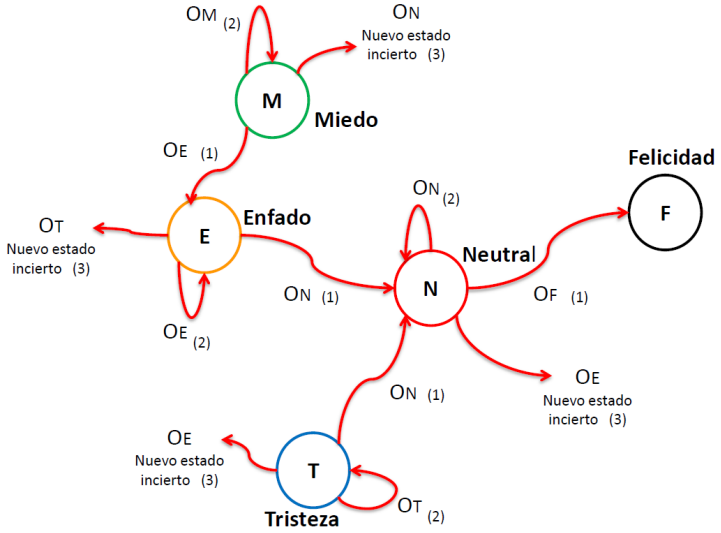
Además, debido a que se intenta recrear condiciones reales en la interacción, en muchos casos existe una tendencia a que la maquina de estados sea dirigida a estados inciertos, principalmente en aquellas ocasiones donde no existe un objeto relacionado a los estados emocionales que se esperan (3). No obstante, es común que existan un valor similar en los resultados erróneos, debido a que los sujetos en el experimento tienen percepciones diferentes acerca de qué objeto está relacionado con cada emoción, incluso esta percepción puede ser influenciada por los demás objetos en el escenario. Para evitar estos problemas con la percepción de las personas en el un proceso previo de selección de los objetos, se intentó mantener una relación emocional común para un grupo heterogéneo de personas. Por su parte, el elevado valor de respuesta positiva en el estado neutral se debe también a lo descrito anteriormente, debido a que cuando un usuario se encuentra en la transición entre estados emocionales de baja intensidad y no puede decidir una emoción valida para la situación, mantiene un estado neutral.

10.4. Conclusiones

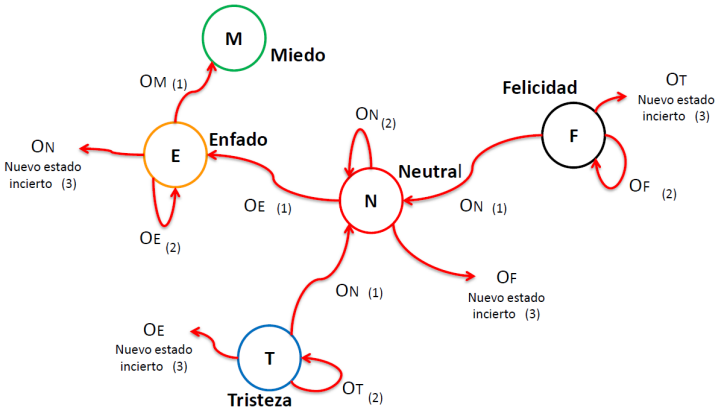
El presente capítulo describe la evaluación llevada a cabo del sistema de aprendizaje basado en $affordances$ emocionales, así como de los modelos de comportamiento para robots sociales. En el experimento se ha utilizado la cabeza robótica Muecas, capaz de generar emociones, y el reconocedor emociones basado en el análisis de las expresiones faciales del usuario, en concreto, el método que usa el modelo de malla $Candide-3$. En este capítulo se ha definido un escenario real para la interacción entre un humano y un robot, donde diferentes objetos son presentados como estímulos para provocar, junto con las expresiones faciales del robot, una reacción sobre el usuario.

En los experimentos aquí presentados se demuestra que el uso de las $affordances$ emocionales permite al robot aprender comportamientos afectivos durante una interacción, actuando como un agente capaz de modificar sus emociones en función de los objetos con los que interactúe o el propio estado anímico del usuario. A su vez, la evaluación de los modelos de comportamiento arroja interesantes resultados, demostrándose cómo el robot puede alterar su comportamiento y la interacción con el entorno para conseguir efectos determinados en la persona, haciendo la IHR más natural y real. Aún así, dentro de la evaluación de estos modelos de comportamiento, se observó cómo la limitada cantidad de estados emocionales causa la mayor

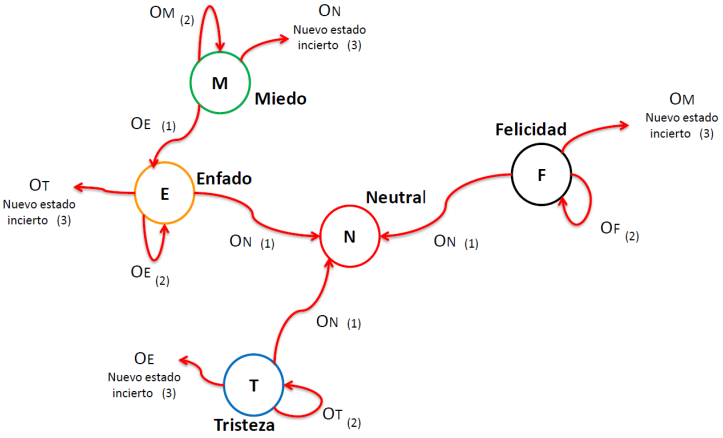
parte de los errores asociados a estados finales del usuario inciertos o imprevisibles. Esto a pesar del uso de un escenario controlado y un entrenamiento inicial limitado a solo cinco estados emocionales que influye en la capacidad de respuesta de los actores, quienes en muchos casos se limitan a expresar emociones más claras, evitando generar emociones secundarias que el robot difícilmente puede reconocer (por ejemplo, el cansancio).



(a)



(b)



(c)

Figura 10.8: Maquinas de estado de los modelos de comportamiento utilizados en los experimentos de la Sección 10.3: a) Positivo; b) Negativo; y c) Neutral.

Parte III

Conclusiones y trabajo futuro

Capítulo 11

Conclusiones

Esta Tesis Doctoral consigue el objetivo principal perseguido, desarrollar un sistema inteligente capaz de aprender e interactuar de forma natural a través de un robot antropomórfico, mediante respuestas emocionales que generen un cambio en el estado emocional del usuario. Esto llevo al desarrollo de múltiples sistemas capaces de estimar la información emocional del usuario por medio de diferentes canales del lenguaje natural. A su vez, para lograr este objetivo, se implementó un sistema de aprendizaje que permitiera a un robot reconocer y crear relaciones entre la información emocional del usuario y los diferentes elementos del entorno, por medio de las $affordances$ emocionales, término acuñado en esta Tesis Doctoral y que representa una extensión de la teoría de las $affordances$.

En la primera parte de este documento se describen y evalúan diferentes sistemas de reconocimiento e imitación de emociones basados en el lenguaje natural, que tienen como objetivo el desarrollo de capacidades de percepción y reacción para los robots, que sean similares a las humanas dentro de una interacción. Por un lado, desde la perspectiva de la percepción era necesario que el robot reconozca cinco estados emocionales del usuario con la mayor precisión posible, utilizando canales de comunicación basados en la información verbal y no verbal durante una IHR. Esto llevó al desarrollo de sistemas que sacaran partido de las capacidades físicas y sensoriales del robot, seleccionando enfoques del lenguaje natural que el mismo agente pudiera realizar, utilizando los principales enfoques en el reconocimiento de emociones que existen en la literatura actualmente, tales como los sistemas basados en expresiones faciales, voz y el lenguaje corporal.

En esta Tesis Doctoral se contribuye con dos sistemas de reconocimiento de emociones que analizan la expresión facial del usuario (Capítulo 3. Ambos siguen un procedimiento similar, basando su funcionamiento en el uso de Unidades de Acción exclusivas, dentro del *Facial Action Code System* (FACS), en un clasificador dinámico bayesiano. El primero de ellos utiliza principalmente un filtro de $Gabor$ previo a la fase de extracción de características, reduciendo considerablemente el ruido de la imagen de entrada RGB. El segundo sistema captura información RGB-D y hace uso del modelo de malla $Candide-3$ para la extracción de los elementos característicos de la imagen. Los resultados obtenidos para estos sistemas, en entornos no controlados (variaciones en las condiciones de luz, número de personas en el escenario durante el experimento, diferentes tipos de participantes, por ejemplo) y con usuarios no entrenados, arrojaron valores relevantes en cuanto a la precisión global del sistema, llegando a mejorar a trabajos similares de la literatura.

También a lo largo de este trabajo se aporta un nuevo sistema de reconocimiento de emo-

ciones basado en el análisis del habla durante una IHR. Este método, descrito en el Capítulo 4, analiza en tiempo real características de la prosodia de la voz humana, que luego determinan la salida de un clasificador dinámico bayesiano. El sistema de reconocimiento basado en la información verbal se presentó como una solución simple y robusta, para su uso en entornos no controlados con un nivel de ruido impredecible.

Junto con los sistemas anteriores, esta Tesis Doctoral hace además nuevas aportaciones tanto en el reconocimiento de emociones basado en el análisis de los movimientos corporales durante la IHR, como en el desarrollo de un sistema multimodal. El primero, analizado en el Capítulo 5, a partir de información RGB-D se extrae un conjunto de características del movimiento humano a través del análisis del esqueleto, las cuales están directamente relacionadas con las emociones. Para la definición de estas características se tuvo en cuenta el Análisis de Movimiento $Laban$, y las categorías que relacionan un movimiento con una determinada emoción. Los resultados obtenidos con este sistema muestran mejoras con respecto a métodos similares del estado del arte. Por su parte, el desarrollo y evaluación de un sistema multimodal basado en las tres modalidades del lenguaje natural descrito anteriormente, permitió mejorar los resultados en el caso de usar uno solo de ellos. El sistema, presentado en el Capítulo 6, está basado en una modalidad predominante, en este caso las expresiones faciales, de forma que sólo cuando existe información emocional del resto de sistemas, se realimenta en el sistema para conseguir un resultado más exacto.

Por otro lado, desde una perspectiva centrada en las capacidades físicas de reacción del robot frente a estímulos emocionales por parte del usuario, era necesario que el robot fuese capaz de interactuar y no sólo reconocer las emociones e intenciones de los usuarios. Esto llevó al desarrollo de un sistema de imitación que recrea tanto las expresiones faciales, como la voz y el lenguaje corporal. Todo esto se concreta finalmente en la cabeza antropomórfica Muecas, un robot diseñado para IHR afectivas. En el sistema de imitación, descrito en el Capítulo 7, se presentan a su vez un sistema de imitación de expresiones faciales y un sistema de generación de voz con componentes emocionales. El primero de ellos basa su funcionamiento también en las Unidades de Acción del FACS, permitiendo su extensión a cualquier otra cabeza robótica capaz de generar emociones. El segundo, aparte de generar la voz sintética de Muecas con las modificaciones oportunas para expresar emociones, desarrolla también un algoritmo que sincroniza los movimientos de la boca con el habla por medio de la cuantificación de la entropía en la señal de audio. Este algoritmo de sincronización también es una contribución en esta Tesis Doctoral.

En la segunda parte se presentó el sistema de aprendizaje emocional para IHR, objetivo principal de esta Tesis (Capítulo 9). Este sistema toma como base la extensión del concepto clásico de las $affordances$, denominado $affordances$ emocionales, y a partir de ellas se generan relaciones afectivas entre la misma información emocional del usuario y los elementos del entorno (elementos afectivos). Debido a que el sistema de aprendizaje asocia determinados objetos con estados emocionales específicos del usuario, se planteó una aplicación basada en las $affordances$ emocionales con el objetivo de utilizar modelos de comportamientos afectivos en el robot. Estos comportamientos están basados en maquinas de estado que tratan de orientar el estado emocional del usuario mediante la interacción con elementos afectivos del entorno que estén dentro de la memoria interna aprendida por el robot. Estos modelos de comportamiento tenían tres enfoques diferenciados: i) llevar al humano al estado emocional Felicidad; ii) dirigir el estado emocional del usuario al estado Neutral, y iii) orientar el estado emocional del usuario al estado Enfado. Dentro de estas interacciones con los modelos de comportamiento se utilizó la

cabeza robótica Muecas en un escenario afectivo controlado, donde parte de los objetos habían sido entrenados y la otra parte podía ser inferida por medio de la información de aprendizaje.

La evaluación de este sistema se realizó utilizando una IHR real, guiada por el robot Muecas, y sin la presencia de ningún observador externo. Las pruebas más significativas consistieron en medir el éxito del aprendizaje propuesto, así como la capacidad del robot de llevar a cabo correctamente el modelo de comportamiento con el que fue programado. Ambos experimentos arrojaron resultados prometedores, demostrando cómo un robot puede llegar a aprender a reaccionar emocionalmente en una interacción con humanos, a la vez que aprende a interactuar con el entorno para un determinado fin definido en su modelo de comportamiento.

Como valoración final de esta Tesis Doctoral, después de evaluar todos los aspectos presentados, queda demostrado que este documento presenta un sistema complejo que reúne y analiza mucha información acerca de diferentes teorías de la psicología y de la conducta humana, aplicándola al campo de la robótica. Se describen cada una de las fases de una IHR real, y se demuestra cómo el lenguaje natural está presente en cada una de ellas. En conclusión, este trabajo cumple con las expectativas y los objetivos planteados inicialmente, a través de una serie de procesos de desarrollo y experimentación con usuarios no entrenados, que proporcionan información y resultados reales acerca del rendimiento y la efectividad de los métodos propuestos. No obstante, también deja un margen para mejorar su contenido en futuros trabajos de investigación, debido a que múltiples aspectos y principios que no se consideraron inicialmente fueron apareciendo durante el desarrollo y evaluación final de los experimentos, principalmente en las teorías relacionadas a las $affordances$ emocionales, permitiendo que se llegue a un final de este trabajo con la esperanza de que el contenido sea apreciado y desarrollado por más personas en un futuro, hasta que sea extendido en su totalidad.

Capítulo 12

Trabajo futuro

Dentro de todos los temas analizados en esta Tesis Doctoral, existen una serie de conceptos que pueden ser trabajados y extendidos en el futuro. A continuación se hace una reflexión acerca de los mismos:

- Uno de los principales trabajos futuros de esta Tesis Doctoral está relacionado con un factor no explorado en la misma, como es la manipulación de objetos por parte del robot y cómo ésta puede favorecer en una IHR afectiva. La inexistencia de un manipulador de objetos real dentro de este trabajo obligó a realizar los experimentos con entornos virtuales, lo que restaba realismo a la interacción.

- Otro de los temas a tratar en el futuro es cómo cuantificar durante una IHR los efectos de la interacción por medio de modelos en la percepción de los usuarios. Esto está directamente relacionado con la necesidad de estimar cómo afecta a la percepción del usuario el uso de modelos de comportamiento que modifican directamente un factor psicológico como son las emociones, por medio de estímulos externos como los objetos y las expresiones faciales basadas en el lenguaje natural.

  Sin embargo, y relacionado con el punto anterior, este proceso de cuantificación y evaluación de la percepción del usuario respecto a las interacciones con un agente antropomórfico, sólo obtendría una relevancia significativa si se implementara la capacidad de manipulación de objetos en el robot. Esto permitirá estudiar aspectos tan importantes como el efecto $Mori$ [Mori, 1970] dentro de las IHR, a través de factores relacionados a la percepción de los humanos (por ejemplo, la atención, la expresividad o la naturalidad, entre otros).

- Otro objetivo para desarrollar en el futuro, es corregir o ampliar las limitaciones asociadas al uso de sólo cinco emociones básicas, como sucede en el caso de los sistemas de reconocimiento e imitación y los modelos de comportamiento emocional descritos en esta Tesis Doctoral. Las causas detrás de estas limitaciones se deben a una elección inicial de los estados emocionales basada en una versión modificada de los estudios de la teoría de las emociones de Ekman. Esta modificación, pensada más para IHR, excluía estados como el de disgusto y sorpresa, debido a su similitud con otros estados como el de enfado o miedo, e introducía un estado necesario para monitorizar las emociones de las personas, que se correspondía con el estado neutral. No obstante, después de analizar los resultados globales de la Tesis, se observa un brecha de mejora con la posibilidad de incluir un nuevo

enfoque basado en la teoría de las emociones de Plutchik, formada por ocho emociones básicas sin incluir el estado neutral. Es posible que con un nuevo enfoque en la teoría de las emociones, la solución presente mejoras sustanciales en los resultados, principalmente en los procesos relacionados con las maquinas de estado descritas en el Capítulo 9.

- Finalmente, otro aspecto a considerar como una posibilidad de investigación futura, es extender aun más el concepto de $affordances$ emocionales, incluyendo el *contexto del entorno* en el que se encuentran los elementos afectivos. Esta propuesta está basada en las denominadas $Situated\ Affordances$ descrita por Kammer en [M. Kammer and Nagai, 2011], las cuales formulan una teoría que toma en consideración no sólo las acciones, los efectos y los objetos, sino también el contexto del entorno dentro del concepto de $affordances$.

Capítulo 13

Publicaciones

2014:

1. ”Muecas: A Multi-Sensor Robotic Head for Affective Human Robot Interaction and Imitation”. **F. Cid**, J. Moreno, P. Bustos and P. Núñez. In *Sensors - Babel* ISSN 1424-8220, vol. 14, no. 5, pp. 7711-7737.
2. ”Learning Emotional Affordances based on Affective Elements in Human-Robot Interaction Scenarios”. **F. Cid** and P. Núñez. In *Proc. of he XV Workshop on Physical Agents - WAF2014*. June 2013, pp. 83 - 92.

2013:

1. ”A Real Time and Robust Facial Expression Recognition and Imitation approach for Affective Human-Robot Interaction Using Gabor filtering”. **F. Cid**, J.A. Prado, P. Bustos and P. Núñez. In *Proc. of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems -(IROS)*. November 2013, pp. 2188 - 2193.
2. ”RGB-D Database for Affective Multimodal Human-Robot Interaction”. C. Doblado, E. Mogena, **F. Cid**, L. V. Calderita and P. Núñez. In *Proc. of he XIV Workshop on Physical Agents - WAF2013*. September 2013, pp. 35 - 40.
3. ”A Novel Real Time Facial Expression Recognition system based on Candide-3 Reconstruction Model”. P. Romero, **F. Cid** and P. Núñez. In *Proc. of he XIV Workshop on Physical Agents - WAF2013*. September 2013, pp. 41 - 46.
4. ”A New Paradigm for Learning Affective Behaviors: Emotional Affordances in Human Robot Interaction”. **F. Cid**, A.J. Palomino and P. Núñez. In *Proc. of he XIV Workshop on Physical Agents - WAF2013*. September 2013, pp. 47 - 52.
5. ”Imitation System for Humanoid Robotics Head”. **F. Cid**, J.A. Prado, P. Manzano, P. Bustos and P. Núñez. In *Journal of Physical Agents* ISSN 1888-0258. Vol. 7, No. 1, pp 22 - 29. January 2013.

2012:

1. ”Development of a Facial Expression Recognition and Imitation Method for Affective HRI”. **F. Cid**, J.A. Prado, P. Bustos and P. Núñez. In *Proc. of Workshop of Physical Agents 2012*. September 2012.
2. ”Engaging human-to-robot attention using conversational gestures and lip-synchronization”. **F. Cid**, R. Cintas, L.J. Manso, L.V. Calderita, A. Sánchez and P Núñez. In *Journal of Physical Agents* ISSN 1888-0258. Vol. 6, No. 1, pp 3-10. March 2012.

2011:

1. ”A real-time synchronization algorithm between Text-To-Speech (TTS) system and Robot Mouth for Social Robotic Applications”. **F. Cid**, R. Cintas, L.J. Manso, L. Calderita, A. Sánchez and P. Núñez. In *Proc. of Workshop of Physical Agents* 2011. September 2011.
2. “Mecanismos de Interacción Hombre-Máquina para el diseño de robots sociales”. *MSc Thesis* **F. Cid**. Cáceres, September 2011.

Apéndices

Apéndice A

Librerías utilizadas en el reconocimiento e imitación de emociones durante una IHR

El objetivo de este apéndice es proporcionar una información detallada de las librerías o programas utilizados dentro de la Parte I de esta Tesis Doctoral, específicamente, aquella relacionada a la adquisición de la información visual o auditiva, para los sistemas de reconocimiento e imitación de emociones.

La estructura de este documento sigue un orden similar al presentado en la Parte I: i) una descripción de los programas que adquieren, analizan y transmiten la información visual de los usuarios a los diferentes sistemas de reconocimiento e imitación; ii) una descripción de las librerías de captura, visualización y reproducción de audio, necesarias en los sistemas basados en la voz humana; y iii) se dan a conocer las bases de datos externas que permiten evaluar los sistemas de reconocimiento de emociones por medio de diferentes fuentes de información en una serie de experimentos, descritos en los capítulos 3 y 4.

A.1. Facial Action Code System

$FACS$ ($Facial\ Action\ Coding\ System$) es un sistema desarrollado por P. Ekman y W.V. Friessen [Ekman et al., 2002], que identifica y categoriza el comportamiento de la actividad muscular facial de los humanos, siendo un elemento importante dentro de sistemas que analizan los movimientos de los músculos faciales y corporales (cuello). Este sistema no proporciona una información detallada de cada musculo facial, sino que categoriza o clasifica cada distorsión facial, causada por la actividad muscular de un pequeño grupo de músculos faciales, en elementos denominados $Unidades\ de\ acción$ ***AUs*** . Estas Unidades de Acción están relacionadas con movimientos específicos, ya sean de la cara, del cuello o de los movimientos necesario para la generación de las expresiones faciales. Por este motivo, múltiples sistemas consideran que cada movimiento o expresión facial está compuesta de diferentes AUs asociadas a distintos elementos de la cara. Lo cual da lugar, en el caso de las expresiones faciales, a que el sistema $FACS$ sea presentado como un estándar utilizado en múltiples estudios que permiten el reconocimiento de las diferentes expresiones faciales del usuario.

Dado que cada AUs está asociada a una actividad especifica de un grupo de músculos faciales, se pueden considerar que estas Unidades de Acción presentan propiedades independientes. Sin embargo, debido a que algunos músculos faciales generan cambios específicos en los ele-

mentos de la cara (por ejemplo, abrir o cerrar la boca), pueden ser considerados también como antagónicos. Estas propiedades se ilustran en la Figura A.1, donde $AU12$ y $AU15$ presentan propiedades opuestas, al ser físicamente imposible mantener la comisura de los labios arriba y abajo en el mismo instante de tiempo.

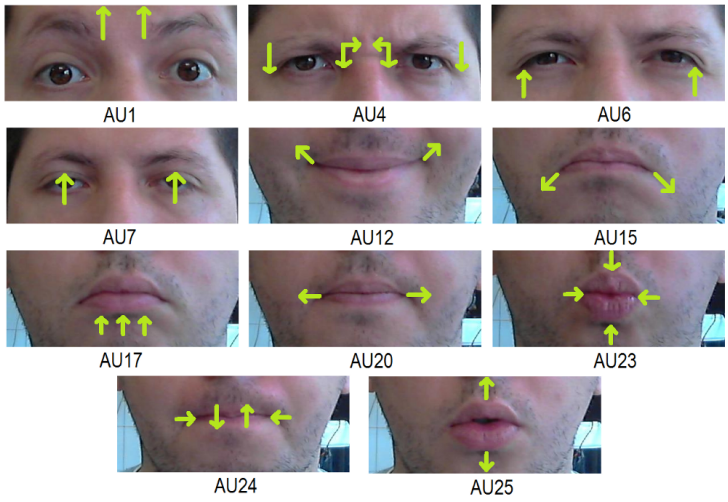


Figura A.1: Ejemplo de algunas Unidades de Acción AUs utilizadas en esta Tesis.

A continuación, se explican dos aspectos importantes a considerar dentro del uso de las Unidades de acción AUs:

1. **Sistema de Intensidad:**

   Las Unidades de Acción están relacionadas directamente a las deformaciones musculares faciales, las cuales influyen en diferentes escalas dentro de las áreas o elementos faciales, de acuerdo a que músculos estén relacionados en cada deformación. Por este motivo, muchas de estas AUs no pueden ser percibidas visualmente por otros humanos, lo que genera un sistema de evaluación de la intensidad de estas deformaciones. Este sistema relaciona y categoriza el nivel de intensidad de cada deformación con la capacidad de ser percibida por otros humanos, por medio de las siguientes letras:

   A: (un rastro de evidencia) El nivel más bajo, asociado a una deformación casi imperceptible por otros humanos. Este nivel permite definir un límite que determina si se debe considerar a una leve actividad de los músculos faciales como una deformación asociada a una AU.

   B: (evidencia leve) este nivel indica un cambio en la apariencia de la cara, a través de una o dos AUs. Este nivel describe que las deformaciones son visibles por múltiples usuarios.

   C: (Marcado o pronunciado) es considerado el nivel que cubre el rango más completo de deformaciones o cambios faciales, principalmente perceptibles a través de los elementos de la cara, como la boca, las cejas u otros.

D: (Extremo) este nivel presenta un gran numero de cambios faciales, normalmente relacionadas a emociones de alta intensidad como el asombro o el miedo.

E: (máxima deformación) El nivel más alto, el cual se asocia a una deformación máxima de los músculos faciales, presentando cambios o expresiones faciales poco naturales o exageradas.

Estos niveles representan diferentes rangos de cambios faciales, debido a que los niveles intermedios C-D están asociados a cambios mas significativos y al uso de más deformaciones faciales. En cambio, niveles como A-B sólo presentan pequeñas deformaciones, las cuales únicamente pueden representar grupos muy definidos de músculos que no puedan ser percibidos rápidamente. Por ultimo, el nivel E representa un pequeño rango de acción en comparación a los niveles intermedios, dado que existe un limitado movimiento de los músculos desde una deformación normal a una que alcance el limite máximo.

Dentro del ámbito de la robótica, el nivel más bajo, A, es apenas perceptible para los humanos, lo cual lo hace virtualmente imperceptible para los robots. Por ello, los sistemas de reconocimiento de emociones basados únicamente en información visual [Cid et al., 2013b], requieren como mínimo un nivel B (evidente a la vista) para identificar correctamente las AUs para cada expresión facial. Finalmente, es importante mencionar la dificultad de identificar de forma visual un cambio en los músculos faciales, debido a que muchos factores pueden afectar esta percepción, tales como los cambios repentinos en el cuerpo, las condiciones de luz o fuerzas externas que puedan dificultar el reconocimiento de las Unidades de Acción.

2. **Sistema de clasificación:**

   Las unidades de acción están basadas en un código numérico que identifica el grupo de elementos que estos músculos faciales afectan. Por ejemplo, las Unidades de Acción de la $AU1$ a la $AU46$, representan los movimientos asociados a elementos de la cara del usuario como la boca, las cejas o lo ojos, entre otros. A continuación, en el cuadro A.1 se describen las AUs más importantes y necesarias dentro del desarrollo de esta Tesis Doctoral.

De esta forma, las diferentes deformaciones asociadas a las múltiples expresiones faciales generadas por un ser humano, pueden ser clasificadas por medio de las unidades de acción del sistema $FACS$. En el caso del formato de las Unidades de Acción, por ejemplo: $AU46-A$ representa a la Unidad de acción 46 asociada al movimiento guiño, con un nivel de intensidad de A.

A.2. Modelo $Candide-3$

El modelo de malla $Candide$-3 [Ahlberg, 2001], [Jiang et al., 2012] es un modelo 3D estandarizado de una cara, compuesto de 113 vértices y 168 superficies que pueden ser controlados por medio de las Unidades de Acción (AU) (ver Figura A.2). $Candide$-3 fue creada para mantener un seguimiento continuo de los elementos de la cara, los cuales son afectados por las deformaciones de los músculos faciales en cada expresión facial (los ojos, la boca, las cejas o la

AUs	Descripción
AU1-AU46	Unidades de Acción relacionadas a los elementos de la cara: desde el movimiento de la boca, hasta el movimiento de las cejas.
AU51-AU56	Estas Unidades de acción están relacionadas a los movimientos más comunes de la cabeza, como son el $Pitch$, $Roll$ y Yaw.
AU61-AU64	Estas Unidades de acción están relacionadas a los movimientos de los ojos, denominados: $Eye\ Tilt$ y $Eye\ Pan$.
AU70-AU74	Unidades de Acción relacionadas a la visibilidad de la cara y de algunos elementos que la componen, tales como la parte baja de la cara o los ojos.

Cuadro A.1: Listado de las Unidades de Acción AUs mas utilizadas en $FACS$.

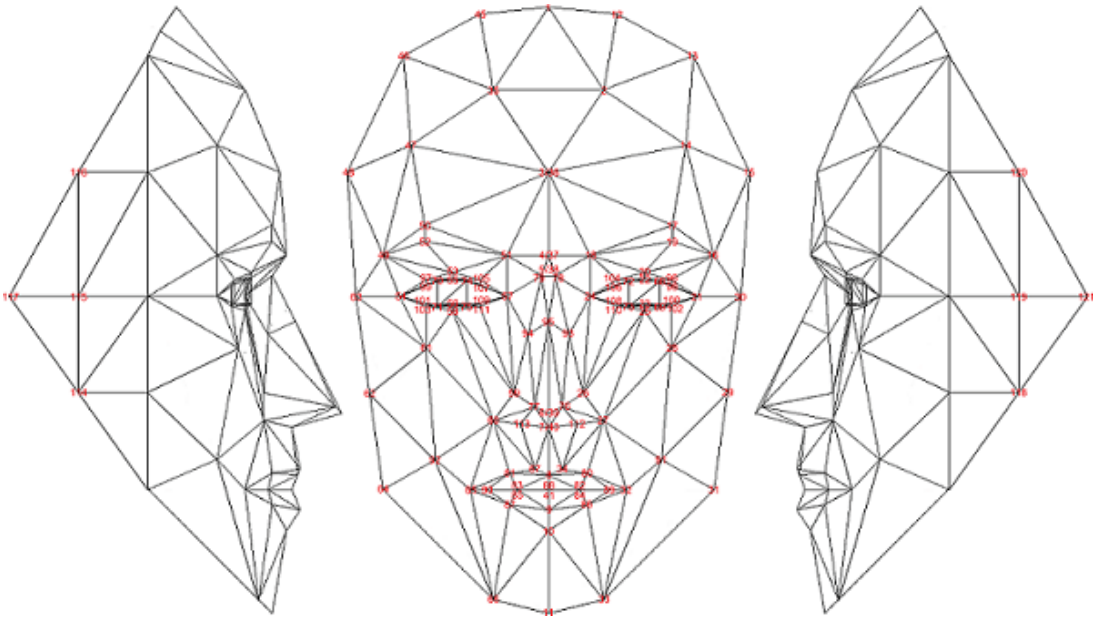
nariz, entre otros). Este seguimiento de los elementos de la cara es necesario para obtener una precisa aproximación de la malla en diferentes orientaciones, desde el usuario al sensor, y es la base de muchos de los sistemas de reconocimiento de emociones por medio de expresiones faciales. La Figura A.3 ilustra los nodos utilizados para calcular las distancias Euclídeas que permiten cuantificar los cambios causados por las deformaciones de los músculos faciales en cada emoción. A su vez, el Cuadro A.2 describe las distancias necesarias en la extracción de características.

El desarrollo de esta malla se encuentra actualmente en su tercera versión, y es utilizada en una gran variedad de programas, librerías y $toolkits$. En esta Tesis Doctoral, la implementación del modelo $Candide$-3 se realiza por medio de la librería $Kinect\ for\ windows\ SDK$ [Microsoft, 2014].

Distancias Euclideas	Nodos	Descripción
d_{eb}	17 – 25	MiddleTopOfLeftEyebrow – UnderMidBottonLeftEyelid
d_{lc}	32 – 65	OutsideLeftCornerMouth – OutsideRightCornerMouth
d_{ma}	8 – 9	MiddleTopDipUpperLip – MiddleBottonDipLowerLip
d_{mf}	9 – 32	MiddleBottonDipLowerLip – OutsideLeftCornerMouth
d_{ch}	21 – 24	OuterCornerOfLeftEye – InnerCornerLeftEye

Cuadro A.2: Descripción de las distancias utilizadas en la extracción de características del Capítulo 3

Por otro lado, la información de los nodos de este modelo es utilizada como dato de aprendizaje. Durante el entrenamiento, previo a los experimentos finales, se capturan los datos de una pequeña muestra de usuarios con diferentes edades, características faciales y género. La información adquirida se relaciona con los posibles valores de las distancias Euclídeas para cada emoción, que serán utilizadas por el clasificador. Estos datos son almacenados en cinco ficheros de texto (*.txt) para uso en cada prueba, cada uno asociado a una emoción. Estos ficheros se encuentran en un directorio llamado $trained$, dentro del componente $AffordancesHumanComp$. Finalmente, en el cuadro A.3, se resume la información para los ficheros por medio de los rangos asociados a las variables de la red en cada estado emocional.



(a)

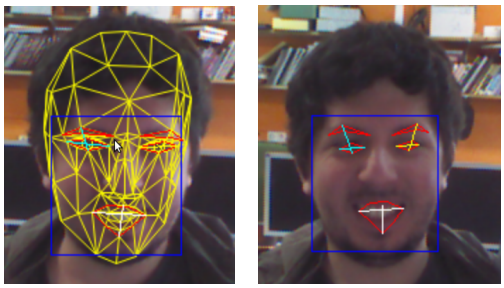
1. TopSkull
2. TopLeftForehead
3. MiddleOfForehead1
4. MidpointBetweenEyebrows1
5. MiddleUpperEdgeOfNoseBone1
6. NoseTip1
7. BottomMiddleEdgeOfNose1
8. MiddleTopDipUpperLip
9. MiddleBottomDipLowerLip
10. AboveChin1
11. BottomOfChin1
12. OneHalfTopLeftOfSkull
13. TopLeftOfSkull
14. LeftBorderBetweenHairAndForehead
15. LeftSideOfSkull
16. LeftOfLeftEyebrow
17. MiddleTopOfLeftEyebrow
18. RightOfLeftEyebrow
19. MiddleBottomOfLeftEyebrow
20. AboveMidUpperLeftEyelid
21. OuterCornerOfLeftEye
22. MiddleTopLeftEyelid
23. MiddleBottomLeftEyelid
24. InnerCornerLeftEye
25. UnderMidBottomLeftEyelid
26. LeftNoseBorder
27. LeftNostrilOuterBorder
28. LeftCheekBone
29. LowerInnerContactPointBetweenLeftEarAndFace
30. UpperInnerContactPointBetweenLeftEarAndFace
31. LeftSideOfChin
32. OutsideLeftCornerMouth
33. LeftOfChin
34. LeftTopDipUpperLip
35. TopRightForehead
36. MiddleOfForehead2
37. MidpointBetweenEyebrows2
38. MiddleUpperEdgeOfNoseBone2
39. NoseTip2
40. BottomMiddleEdgeOfNose2
41. MiddleTopLowerLip
42. Undefined1
43. Undefined2
44. Undefined3
45. OneHalfTopRightOfSkull
46. TopRightOfSkull
47. RightBorderBetweenHairAndForehead
48. RightSideOfSkull
49. RightOfRightEyebrow
50. MiddleTopOfRightEyebrow
51. LeftOfRightEyebrow
52. MiddleBottomOfRightEyebrow
53. AboveMidUpperRightEyelid
54. OuterCornerOfRightEye
55. MiddleTopRightEyelid
56. MiddleBottomRightEyelid
57. InnerCornerRightEye
58. UnderMidBottomRightEyelid
59. RightNoseBorder
60. RightNostrilOuterBorder
61. RightCheekBone
62. LowerInnerContactPointBetweenRightEarAndFace
63. UpperInnerContactPointBetweenRightEarAndFace
64. RightSideOfChin
65. OutsideRightCornerMouth
66. RightOfChin
67. RightTopDipUpperLip
68. OuterTopLeftPupil
69. OuterBottomLeftPupil
70. OuterTopRightPupil
71. OuterBottomRightPupil
72. InnerTopLeftPupil
73. InnerBottomLeftPupil
74. InnerTopRightPupil
75. InnerBottomRightPupil
76. LeftSideOfNoseTip
77. RightSideOfNoseTip
78. LeftUpperEdgeOfNoseBone
79. RightUpperEdgeOfNoseBone
80. LeftTopUpperLip
81. RightTopUpperLip
82. LeftBottomUpperLip
83. RightBottomUpperLip
84. LeftTopLowerLip
85. RightTopLowerLip
86. LeftBottomLowerLip
87. RightBottomLowerLip
88. MiddleBottomUpperLip
89. LeftCornerMouth
90. RightCornerMouth
91. BottomOfLeftCheek
92. BottomOfRightCheek
93. LeftLowerEdgeOfNoseBone
94. RightLowerEdgeOfNoseBone
95. NoseBump
96. AboveThreeFourthLeftEyelid
97. AboveThreeFourthRightEyelid
98. ThreeFourthTopLeftEyelid
99. ThreeFourthTopRightEyelid
100. ThreeFourthBottomLeftEyelid
101. ThreeFourthBottomRightEyelid
102. BelowThreeFourthLeftEyelid
103. BelowThreeFourthRightEyelid
104. AboveOneFourthLeftEyelid
105. AboveOneFourthRightEyelid
106. OneFourthTopLeftEyelid
107. OneFourthTopRightEyelid
108. OneFourthBottomLeftEyelid
109. OneFourthBottomRightEyelid
110. OneFourthBottomLeftOuterEyelid
111. OneFourthBottomRightOuterEyelid
112. LeftNostrilMiddleBorder
113. RightNostrilMiddleBorder
114. LowerOuterContactPointBetweenRightEarAndFace
115. UpperOuterContactPointBetweenRightEarAndFace
116. UpperRightSideOfHead
117. RightSideOfHead
118. LowerOuterContactPointBetweenLeftEarAndFace
119. UpperOuterContactPointBetweenLeftEarAndFace
120. UpperLeftSideOfHead
121. LeftSideOfHead

(b)

Figura A.2: a) Figura del modelo de malla *Candide*-3, que contiene la enumeración de los nodos que la componen; b) Descripción de los nodos con sus correspondientes nombres.

A.3. Librería SoX

La librería *SoX* [C. Bagwell, 2014] es una herramienta de procesamiento de audio basada en líneas de comandos, especializada en grabar, editar y analizar señales por medio de ficheros en el sistema operativo Linux. Esta librería se usa de forma similar en distintos sistemas a lo



(a) (b)

Figura A.3: Extracción de características faciales basado en el modelo $Candide$-3; a) Modelo de malla sobre la cara del usuario; b) Características extraídas del modelo. (Figura obtenida de la publicación [Cid and Núñez, 2014])

Variables	Felicidad.txt	Enfado.txt	Tristeza.txt	Miedo.txt	Neutral.txt
MF	18 – 25	7 – 12	8 – 12	11 – 14	10 – 14
MA	22 – 27	24 – 28	22 – 25	21 – 24	18 – 23
LC	48 – 56	31 – 39	43 – 47	40 – 56	38 – 55
EB	30 – 32	21 – 29	24 – 26	41 – 45	33 – 35
CH	32 – 37	24 – 29	22 – 27	33 – 35	32 – 39

Cuadro A.3: Descripción detallada de los datos de aprendizaje necesarios en el clasificador bayesiano del Capítulo 3

largo de esta Tesis Doctoral:

- **Detección y captura de información acústica**

  Esta primera etapa está asociada a la captura del $stream$ de audio desde los sensores acústicos, por medio de una función de detección de silencios que permite adquirir solamente los períodos de audio que contengan información auditiva asociada a la voz. La naturaleza de esta función de detección permite no sólo analizar cada muestra por medio de umbrales que determinan la existencia de sonidos (no ruido) y de silencios, sino también eliminar períodos específicos de audio antes, durante y después de detectar determinados sonidos. Así, al capturar la información acústica sólo se obtienen las partes asociadas a la voz y las pausas entre las palabras dentro de cada frase (debido a que se eliminan sólo los períodos de silencio con una larga duración). Para ejecutar este proceso se realiza una llamada por medio de la siguiente línea de comando:

  sox -t alsa default ./out.flac silence -l 1 0.01 0.5 % 2 1.0 0.5 %

  Las opciones de esta línea de comando, son descritas en detalle en el cuadro A.4.

- **Pre-procesamiento de audio**

  La segunda etapa está basada en un pre-procesamiento del fichero de salida de la línea de comando anterior ($out.flac$), el cual contiene la señal de audio capturada desde el

Opción de SoX	Descripción
-t	Define el tipo de archivo de audio utilizado en la adquisición de la señal.
alsa	Esta opción hace referencia al controlador del dispositivo de audio (tarjeta de sonido) que es parte del kernel de Linux, el cual presenta una compatibilidad con SoX para capturar los datos desde el sensor acústico.
default	Describe el uso de la tarjeta de la audio configurada por defecto en el equipo.
out.flac	Archivo de audio de salida en formato $flac$ (compresión sin pérdida), que contiene la información acústica.
silence [-l] (above-periods [duration threshold[%]] [below-periods duration threshold[%]])	Esta opción elimina los silencios (al comienzo, mitad y final) durante el proceso de captura de audio, por medio de umbrales asociados al ruido. Sin embargo, en esta llamada sólo se eliminan los silencios al comienzo y al final a través de un $above\text{-}periods$ de 1 segundo (elimina la información hasta que encuentra un sonido) y un $below\text{-}periods$ de 2 segundos (elimina toda la información después de un período de silencio). En el caso de la opción $above\text{-}periods$ que posee un valor distinto de cero, se elimina el audio al comienzo del $streaming$ hasta que encuentre un período que contenga un sonido que supere el valor del umbral ($threshold$), con una duración ($duration$) superior a 0.01 segundos. Por el contrario, la opción $below\text{-}periods$ elimina el audio después de detectar un período de silencio que contenga muestras con valores inferiores al umbral ($threshold$), con una duración ($duration$) superior a 1.0 segundos. El uso de un $below\text{-}periods$ de 2 segundos y una duración ($duration$) superior a 1 segundo evita eliminar los silencios asociados a las pausas entre las palabras de cada frase, gracias a su corta duración. Finalmente, la opción $-l$ asigna los umbrales ($threshold$) como porcentajes del máximo valor de la muestra, siendo 0 % silencio puro digital. Así, estos umbrales determinan los valores de las muestras que deben ser tratados como silencios, siendo diferentes de 0 para eliminar el ruido de fondo y considerar la sensibilidad del sensor.

Cuadro A.4: Opciones de la librería SoX

$streaming$. El objetivo de este proceso, es reducir la cantidad de información acústica, al eliminar las muestras que contengan datos no relacionados a la voz del usuario, tales como el ruido ambiente, la música o los sonidos fuertes, entre otros. Por esta razón, se decidió utilizar una segunda línea de comandos por medio de SoX, que ejecute las siguientes funciones: i) una función específica de detección de voz llamada VAD ($Voice\ Activity\ Detection$); ii) la implementación de un filtro pasa bajo sobre la señal de audio; y iii) un cambio en las propiedades del fichero de audio como la frecuencia de muestreo F_s, el número de bits y canales. De esta forma, por medio de la siguiente línea de comando, se implementarán las funciones descritas anteriormente.

sox out.flac -r 16000 -b 16 -c 1 outvad.flac vad reverse vad reverse lowpass -2 2500

Opción de SoX	Descripción
out.flac	Archivo de audio de entrada en formato $flac$ (compresión sin pérdida).
-r	Esta opción especifica que la frecuencia de muestreo (F_s) debe ser de 16000 (hz) en el archivo de audio de salida.
-b	Esta opción especifica que el número de bits del archivo de salida debe ser igual a 16.
-c	Esta opción determina que el número de canales debe ser igual a 1 en el archivo de salida.
outvad.flac	Archivo de audio de salida en formato $flac$ (compresión sin pérdida).
vad reverse	Función de detección de actividad de voz basado en la cuantificación del $Cepstrum$ de potencia en cada trama de la señal, el cual elimina cualquier silencio, ruido o sonido que no esté relacionado a la voz humana dentro del archivo de entrada. Dado que esta función sólo elimina información en forma frontal, se utiliza la opción $reverse$ para implementar su funcionamiento desde ambos extremos (desde la parte frontal y la parte posterior).
lowpass	Esta opción implementa un filtro paso bajo con una frecuencia de corte (F_c) de 2500 (hz), un Q=0,707 y una respuesta $Butterworth$ sobre la señal del archivo de entrada. La opción -2 determina el uso de un filtro $double\text{-}pole$.

Cuadro A.5: Opciones de la librería SoX para la etapa de pre-procesamiento de audio

A continuación, en el cuadro A.5 se describen las opciones de esta línea de comando y las características del fichero de salida que contiene la información procesada.

A pesar de los buenos resultados demostrados por esta línea de comando, el uso de una segunda etapa parece innecesaria en términos de análisis, debido a que se podría reemplazar la función de detección de silencios por la función VAD en la primera etapa, reduciendo el proceso de adquisición. Sin embargo, las limitaciones de la función VAD están asociadas al hecho que no puede ser implementada para su uso en $streaming$ en tiempo real, ya que la opción $reverse$ no fue desarrollada para un $stream$ de audio. Por ello, sólo se utiliza como un pre-procesamiento para eliminar las tramas de audio que no contengan la voz humana en el archivo de entrada de este proceso, dando como resultado, que el archivo de audio de salida $outvad.flac$ contenga sólo la información verbal y sea utilizado como dato de entrada las etapas siguientes.

Finalmente, en caso de necesitar otro formato para el fichero de salida, por ejemplo un $.wav$ y una frecuencia de muestreo más elevada, como ejemplo, F_s = 44100 hz, se puede utilizar una nueva línea de comandos:

sox outvad.flac outresample.wav rate 44100

En el cuadro A.6 se describen en detalle las opciones de esta línea.

Opción de SoX	Descripción
outvad.flac	Archivo de audio de entrada.
outresample.wav	Archivo de audio de salida, que contiene la señal original con la nueva frecuencia de muestreo (F_s= 44100 hz en este caso).
rate	Esta opción realiza un remuestreo de la señal original en archivo de entrada, y genera un nuevo archivo de salida con una nueva frecuencia de muestreo F_s = 44100 hz.

Cuadro A.6: Opciones de la librería SoX dentro del proceso de re-muestreo

- **Reproducción de audio**

  Esta última etapa del procedimiento, sólo es utilizada para reproducir y comprobar el correcto funcionamiento del archivo de audio de forma eficiente y rápida dentro de los experimentos. Por lo cual, se ejecuta mediante una simple llamada de SoX utilizando la siguiente línea de comando:

  play outvad.flac

  Donde $outvad.flac$ es el fichero de entrada para la reproducción del audio.

A.4. Librería Praat

La librería $Praat$ es un software desarrollado por P. Boesma y V. van Heuven [Boersma and van Heuven, 2001] [P. Boersma and D. Weenink, 2014], para analizar señales de audio que contengan voz desde el punto de vista de fonetistas, lingüistas y fonólogos. Esto se debe a que esta herramienta permite visualizar, editar y analizar desde el dominio del tiempo y la frecuencia ($Spectrum$), elementos característicos de la voz humana, como son el $Pitch$, $Pulses$, $Intensity$ o $Formants$, entre otros. Por esta razón, el principal uso de esta librería en esta Tesis Doctoral está asociada a su capacidad para representar de forma visual las señales de audio y las características de la voz presentes en el capítulo 4. En la Figura A.4 se observa una representación visual de estos elementos desde una frase o sentencia grabada por medio de la librería SoX en un entorno con ruido ambiental.

A.5. Reproductor MPlayer

El reproductor multimedia $MPlayer$ es un software multiplataforma, que permite reproducir ficheros de audio en múltiples formatos de forma rápida y eficiente. El uso de $MPlayer$ dentro de esta Tesis Doctoral, está relacionada con la generación de audio dentro del sistema TTS ($Text\ To\ Speech$) por medio de la reproducción del fichero de audio ($out.wav$). Para ejecutar este proceso, se realiza una llamada por medio de la siguiente línea de comando:

mplayer -really-quiet out.wav

Donde la opción $-really-quiet$, elimina los retardos asociados a la visualización de las propiedades del fichero $out.wav$ durante la ejecución del software.

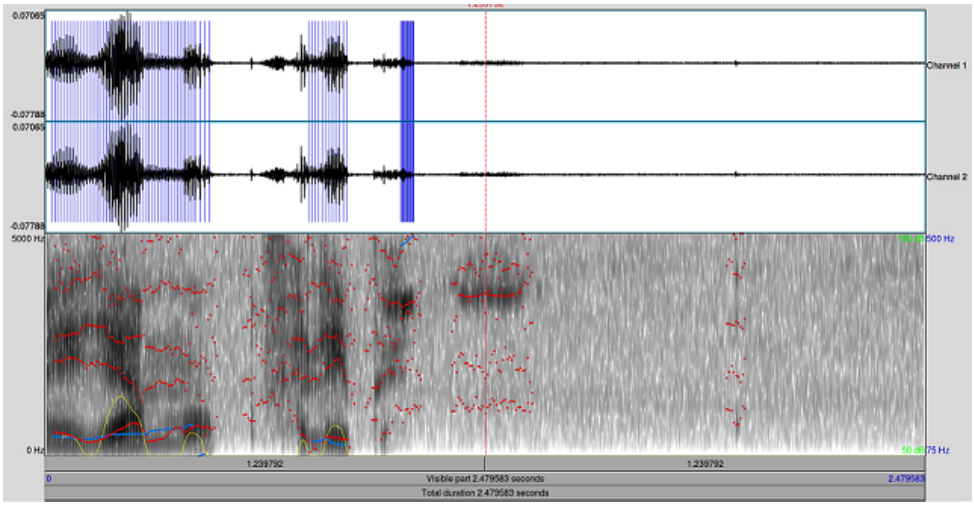


Figura A.4: Representación visual de una señal de audio a través de la librería $Praat$.

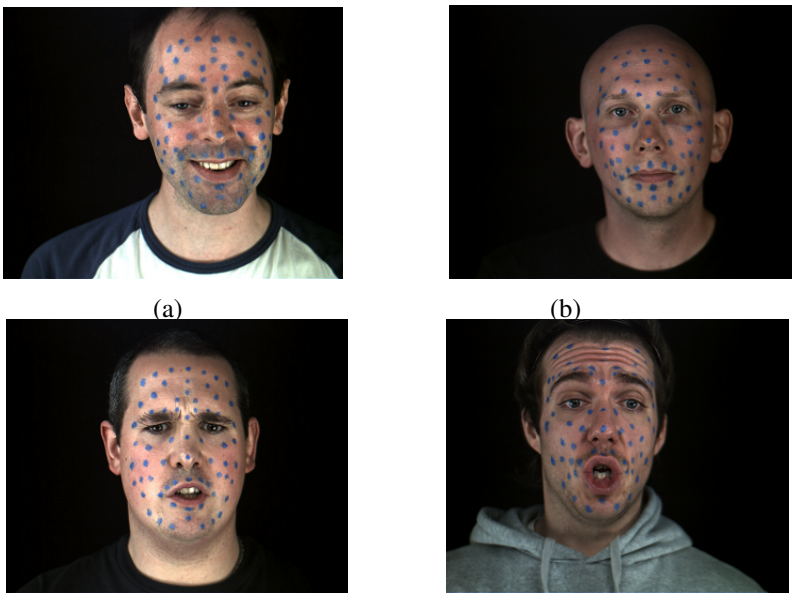
A.6. Base de datos - SAVEE

La base de datos $SAVEE$ fue desarrollada por S. Haq y P. Jackson [Haq and Jackson, 2010], con el objetivo de ser utilizada como una fuente de información que permitiera evaluar sistemas de reconocimiento de emociones basados en información visual y auditiva. Por este motivo, se eligió a $SAVEE$ como una entrada alternativa de datos dentro de los diferentes experimentos del sistema de reconocimiento de emociones basadas en expresiones faciales descrito en esta Tesis Doctoral.

En términos generales, $SAVEE$ está compuesta de cuatro usuarios, de genero masculino, denominados: DC, KL, JE y JK. Cada uno de estos participantes se encuentra dentro de un rango limitado de edad y posee características faciales y acústicas especificas, como se observa en la Figura A.5 y A.6.

Para cada usuario existe una gran cantidad de información y diferentes tipos de ficheros, de audio, imagen o vídeo, los cuales cubren sólo algunos de los siete estados emocionales presentes en la base de datos (felicidad, disgusto, tristeza, enfado, sorpresa, miedo y neutral). En esta Tesis Doctoral, donde sólo se trabajan con cinco emociones, esto puede generar un problema, ya que no todos los usuarios cubren de forma adecuada los cinco estados emocionales necesarios en los experimentos.

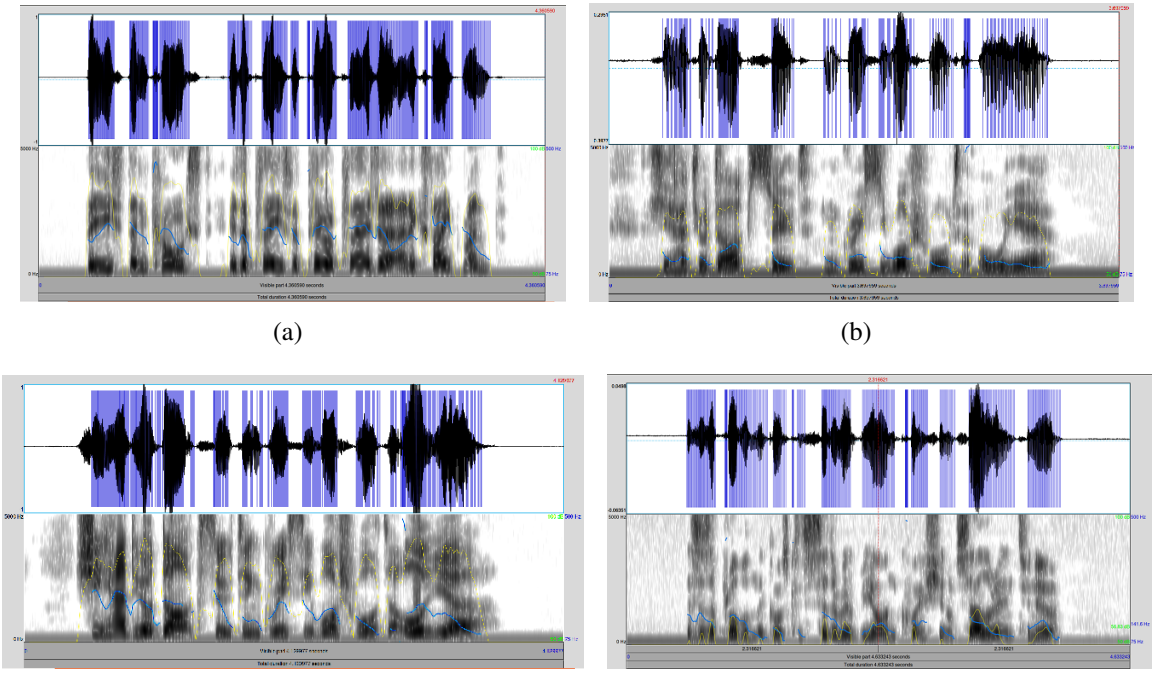
En cada experimento, la implementación y el uso de esta base de datos por parte de los sistemas de reconocimiento de emociones se realizó por medio de ficheros con diferentes formatos para cada tipo de información. Esto se debe a que $SAVEE$ presenta una capacidad para adaptarse a otros sistemas, por medio de múltiples ficheros en formatos de imagen (*.jpeg), vídeo (*.avi) y de audio (*.wav), siendo esto uno de los factores decisivos en la elección de esta base de datos, junto con la buena calidad de las imágenes y el bajo ruido de los ficheros de audio. A pesar de lo anterior, es importante mencionar en términos de comparación y análisis, la existencia de una clara desventaja en el uso de esta fuente de información en los experimentos, debido a que las caras del usuarios poseen marcas o puntos, los cuales provocan una pérdida en la credibilidad de los resultados. Con el fin de corregir este problema se cambiaron de forma mínima los parámetros del filtro de $Gabor$, con el propósito de eliminar las marcas en la imagen, como se puede observar en las Figuras A.7 y A.8.



(a) (b)

(c) (d)

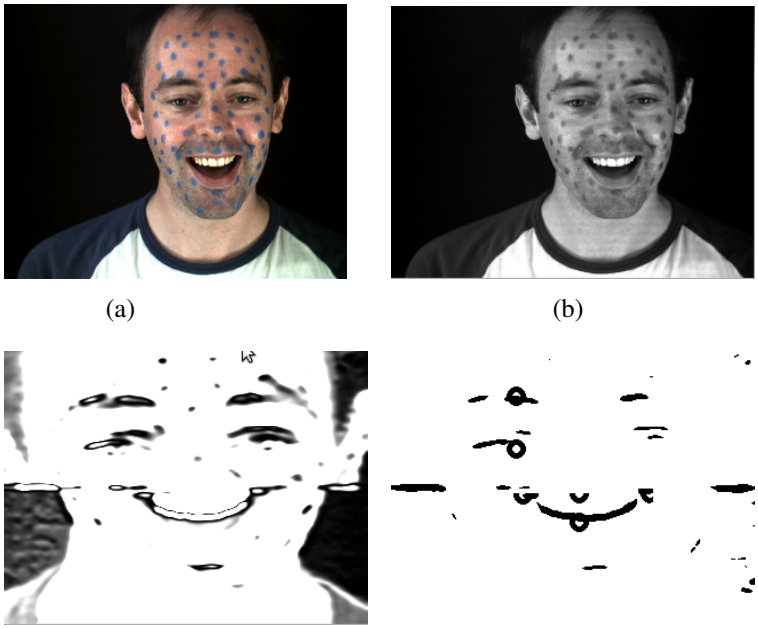
Figura A.5: Imágenes obtenidas de la base de datos audio-visual $SAVEE$; a) Imagen del usuario DC que genera el estado emocional felicidad; b) Imagen del usuario JE que genera el estado emocional neutral; c) Imagen del usuario JK que genera el estado emocional enfado; y d) Imagen del usuario KL que genera el estado emocional miedo;



(a) (b)

(c) (d)

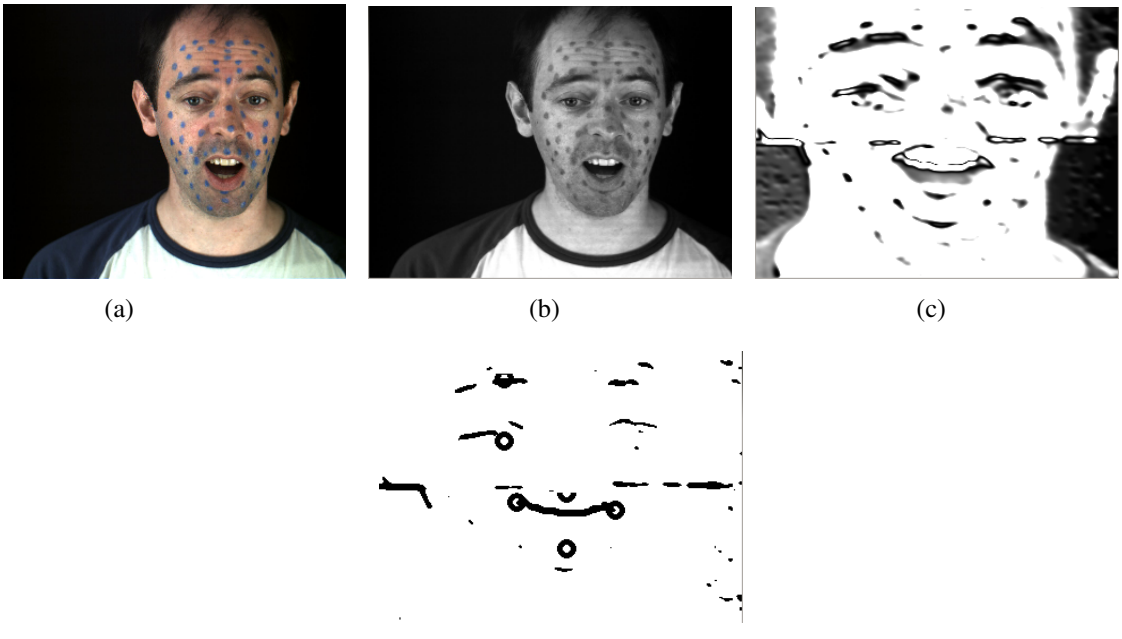
Figura A.6: Señal de audio visualizado por medio del $toolkit\ Praat$ [Haq and Jackson, 2010], obtenida de la base de datos audio-visual $SAVEE$; a) Señal de audio desde el usuario DC que genera el estado emocional felicidad; b) Señal de audio desde el usuario JE que genera el estado emocional neutral; c) Señal de audio desde el usuario JK que genera el estado emocional enfado; y d) Señal de audio desde el usuario KL que genera el estado emocional miedo;



(a) (b)

(c) (d)

Figura A.7: Imágenes obtenidas de la base de datos audio-visual $SAVEE$; a) Imagen original del usuario DC que genera el estado emocional felicidad; b) Imagen en escala de grises del usuario DC; c) Imagen que contiene las sub-regiones ROI_{top} y ROI_{botton}; y d) Imagen procesada por $Gabor$ del usuario DC con el estado emocional felicidad;



(a) (b) (c)

(d)

Figura A.8: Imágenes obtenidas de la base de datos audio-visual $SAVEE$; a) Imagen original del usuario DC que genera el estado emocional miedo; b) Imagen en escala de grises del usuario DC; c) Imagen que contiene las sub-regiones ROI_{top} y ROI_{botton},; y d) Imagen procesada por $Gabor$ del usuario DC.

Apéndice B

Librerías utilizadas en el sistema de aprendizaje emocional

Este apéndice tiene como propósito describir las diferentes librerías y elementos utilizados en la Parte II de esta Tesis Doctoral, aquellas relacionadas principalmente a diferentes métodos de reconocimiento y localización de objetos por medio de marcas visuales para aplicaciones dentro del campo de las $affordances$.

La estructura presentada en este apéndice, comienza con una descripción de los múltiples atributos (visuales y físicos) que permiten diferenciar las clases de objetos utilizados por los sistemas de reconocimiento del Capítulo 9, los cuales buscan relacionar los atributos de un objeto físico con un estado emocional especifico. Por otro lado, se dan a conocer las librerías basadas en marcas de realidad aumentada, utilizadas en el desarrollo y experimentación del sistema de detección de objetos. Finalmente, se ilustran los modelos virtuales utilizados en el simulador de $RoboComp\ RCInnermodelSimulator$ [Manso, 2012], el cual fue utilizado en las pruebas finales del Capítulo 10 para comprobar el correcto funcionamiento del sistema.

B.1. Clases de objetos

Los objetos utilizados dentro de esta Tesis Doctoral, fueron escogidos por ser elementos fáciles de localizar en entornos de la vida diaria, principalmente en escenarios afectivos como un hogar. Han sido agrupados fácilmente en diferentes categorías como juguetes, comida, insectos, animales y objetos cotidianos del hogar. El propósito de estos elementos es cumplir un rol de elemento afectivo dentro del concepto de las $affordances$ emocionales, descrito en el Capítulo 9. Por ello, ha sido necesario la implementación de un método de reconocimiento que sea capaz de analizar, identificar y diferenciar cada uno de los objetos en el escenario, permitiendo al robot, fácilmente, detectar y escoger objetos dentro de los procesos de aprendizaje asociados a las $affordances$. En el cuadro B.1 se ilustran los diferentes objetos con sus respectivos atributos utilizados en los experimentos. Cada atributo representa la siguiente propiedad: forma (AS), color (AC), material (AM), Peso (AW), Tamaño (ASi) y Acciones ($A1$).

Objetos	AS	AC	AM	AW	ASi	$A1$
Manzana	1	9	7	1	1	3
Balón	1	1	2	4	5	1
Taza	2	10	4	3	3	3
Peluche	12	5	6	4	6	6
Gato	6	2	9	6	5	6
Rata	7	10	9	3	3	4
Cucaracha	5	19	9	1	1	4
Pastel	2	5	7	5	5	7
Cubo	7	4	1	2	3	3
Pluma	9	1	9	1	2	7
Teléfono	7	3	1	2	3	3
Rosa	4	9	8	1	4	7
Flores	11	14	8	1	4	7
Cuchillo	11	10	3	3	4	3
Triangulo	3	2	5	1	3	3
Dinosaurio	6	2	1	6	5	6
Escorpión	11	19	9	1	1	4
Araña	11	5	9	1	1	4
Esqueleto	2	1	1	3	3	4

Cuadro B.1: Lista de atributos relacionadas a cada objeto.

B.2. Tipos de marcas

La detección de los objetos requiere de un sistema de reconocimiento que no sólo estime la posición 3D de los objetos, sino que también identifique cada objeto con respecto a otros de apariencia similar. Por este motivo, se decidió utilizar un enfoque diferente a los utilizados por otro trabajos basados en pistas visuales por medio de la información RGB o RGB-D [Koppula et al., 2013], [Katz et al., 2013]. Así, se decidió por el uso de marcas 2D a través de las librerías $ARToolKit$ y $AprilTags$, las cuales permitían obtener la posición y orientación 3D del objeto con respecto al sensor del robot, así como el tipo de objeto.

A continuación, se describen en detalle estas librerías y las marcas utilizadas dentro de esta Tesis Doctoral.

B.2.1. Librería ARToolKit

La librería $ARToolKit$ [Human Interface Technology Laboratory, 2014] es una herramienta basada en marcas 2D para realidad aumentada, que permite estimar la identidad, la posición y la orientación relativa de cada marca con respecto a un sensor RGB en tiempo real. La utilidad de este librería en el ámbito de la robótica está asociada a su capacidad de detección y seguimiento de determinadas marcas asociadas a un objeto especifico dentro de un escenario real, por medio de una secuencia de $frames$ adquiridos por un robot. Esto permite hacer representaciones virtuales (simulaciones) de cada acción que debe realizar

el robot, sobreponiendo modelos virtuales en imágenes reales a través del $toolkit\ OSGArt$ [Looser et al., 2006] [HITLabNZ, 2014]. En esta Tesis, el uso de estas marcas y simulaciones dio lugar a diferentes pruebas en el desarrollo inicial de los sistemas de reconocimiento de objetos, a través de un grupo de 15 marcas relacionadas a objetos geométricos de diferentes colores, como se ilustran en las Figuras B.2 y B.1.

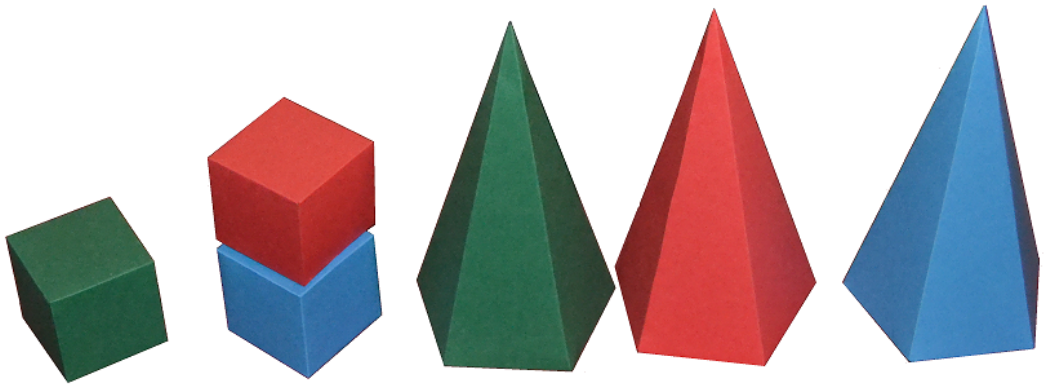


Figura B.1: Ejemplo de algunos de los objetos geométricos utilizados en el desarrollo de las $affordances$ emocionales.

Sin embargo, como se ilustra en la Figura B.3, existe un segundo grupo de marcas asociadas a los 19 diferentes objetos utilizados en los experimentos finales basados en las $affordances$ emocionales.

B.2.2. Librería AprilTags

La librería $AprilTags$ fue desarrollada por E. Olson [Olson, 2011], para la detección e identificación de marcas 2D a través sensores RGB. La implementación de esta librería dentro de esta tesis, tiene como objetivo la identificación (id) y localización (posiciones y orientaciones 3D) de determinados objetos que posean esta marca, por medio de la estimación de la posición relativa del objeto con respecto al sensor RGB localizado en el ojo derecho de la plataforma robótica Muecas.

En la Figura B.4 se ilustra la familia de marcas $tag36h11$ desde el número de identificación $id0$ al $id18$, utilizadas en los experimentos.

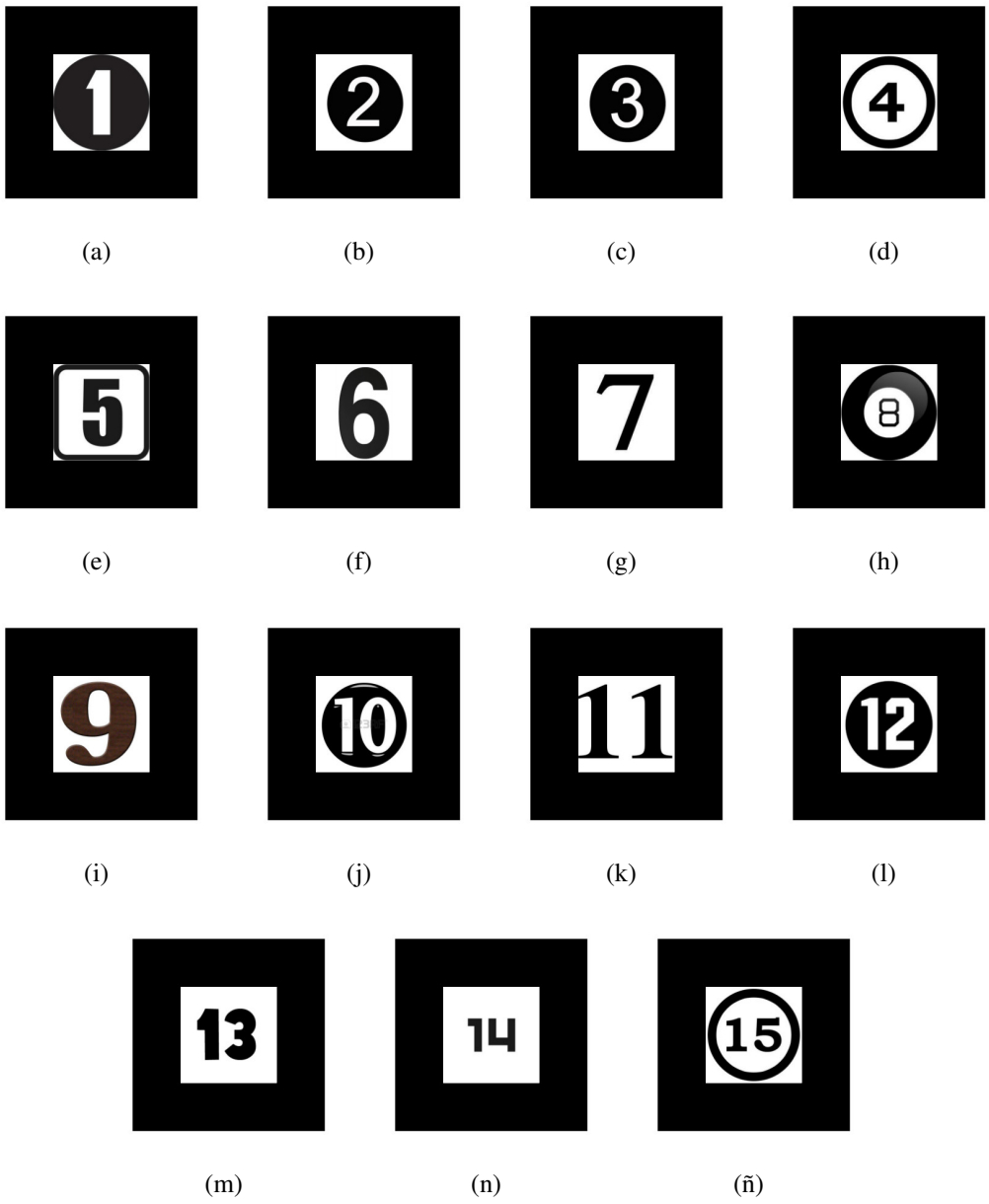
La elección de esta segunda librería se debe al hecho que posee características similares a $ARToolKit$, pero presenta mejores resultados prácticos en términos de detección, identificación y rango de distancia para su uso por robots.

B.3. Modelos 3D

En esta Tesis Doctoral se realizaron una serie de representaciones virtuales mediante modelos 3D de los elementos, agentes y entorno, para comprobar el correcto funcionamiento del sistema de aprendizaje basado en $affordances$ emocionales. Para llevar a cabo esta función, se utilizó el simulador del $framework$ RoboComp denomina-

do $RCInnerModelSimulator$ [Manso, 2012], el cual está basado en el motor gráfico $OpenSceneGraph$ (OSG) [Burns, 2014].

A continuación, en la Figura B.5, se ilustran todos los modelos 3D de los elementos afectivos utilizados para representar los objetos del mundo real.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (ñ)

Figura B.2: Marcas numéricas utilizadas en el desarrollo del algoritmo de reconocimiento de objetos basados en $ARToolKit$, descrito en el Capítulo 9.

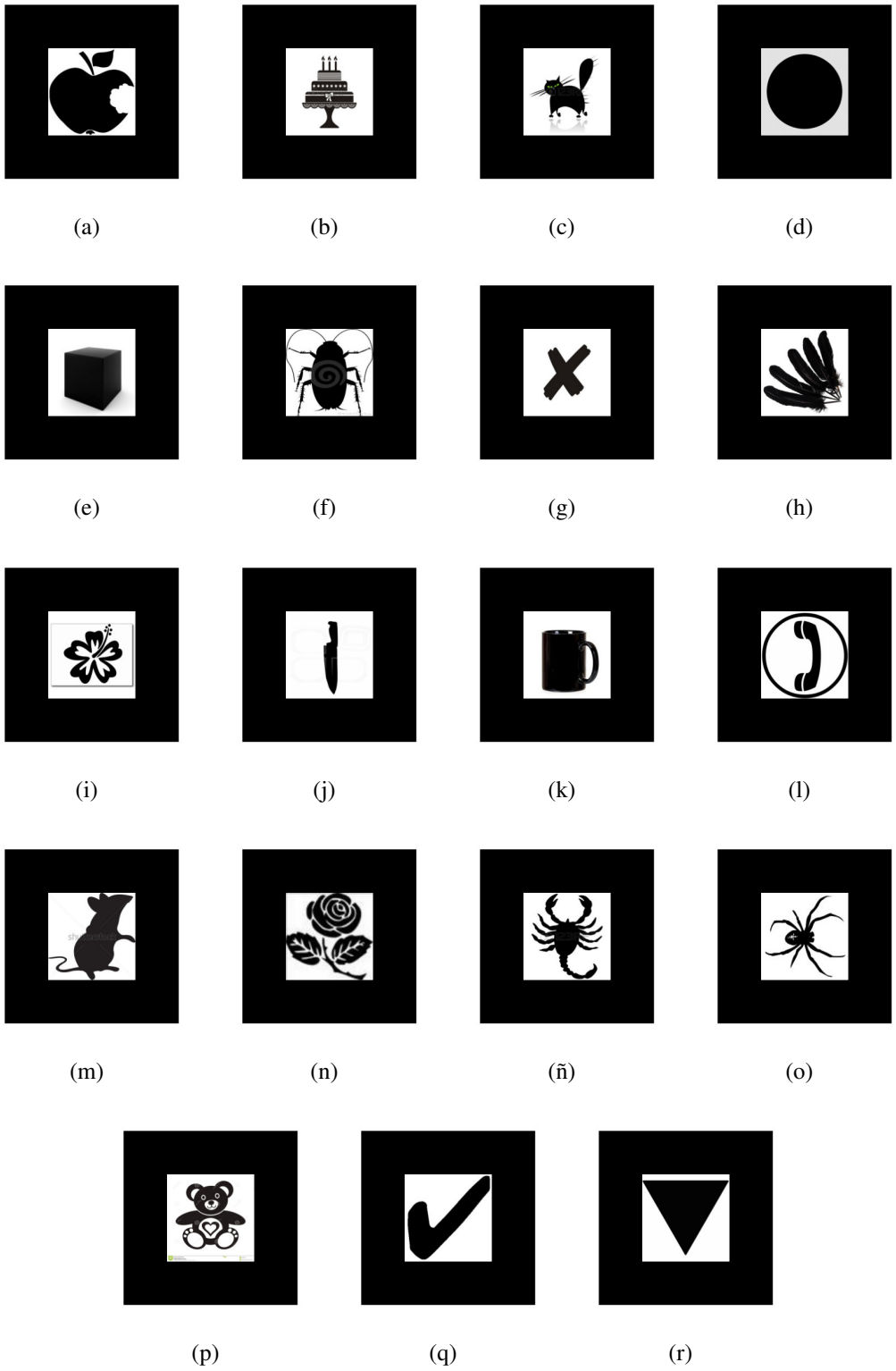


Figura B.3: Marcas utilizadas en el desarrollo del algoritmo de reconocimiento de objetos, descrito en el capítulo 9.

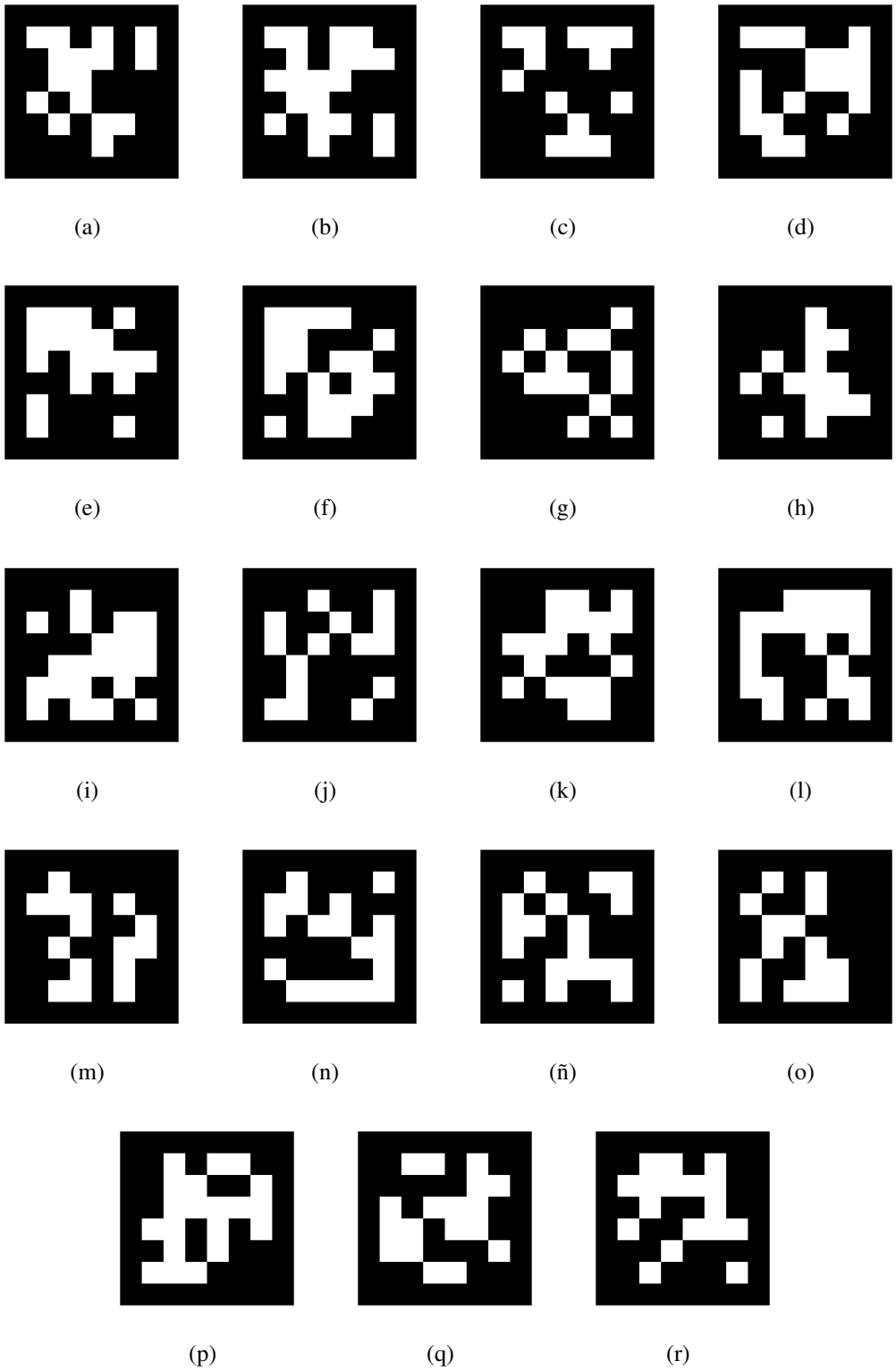


Figura B.4: Marcas utilizadas en el desarrollo del algoritmo de reconocimiento de objetos basado en $AprilTags$ de la familia $tag36h11$ (marcas desde la id:0 a la id:18).

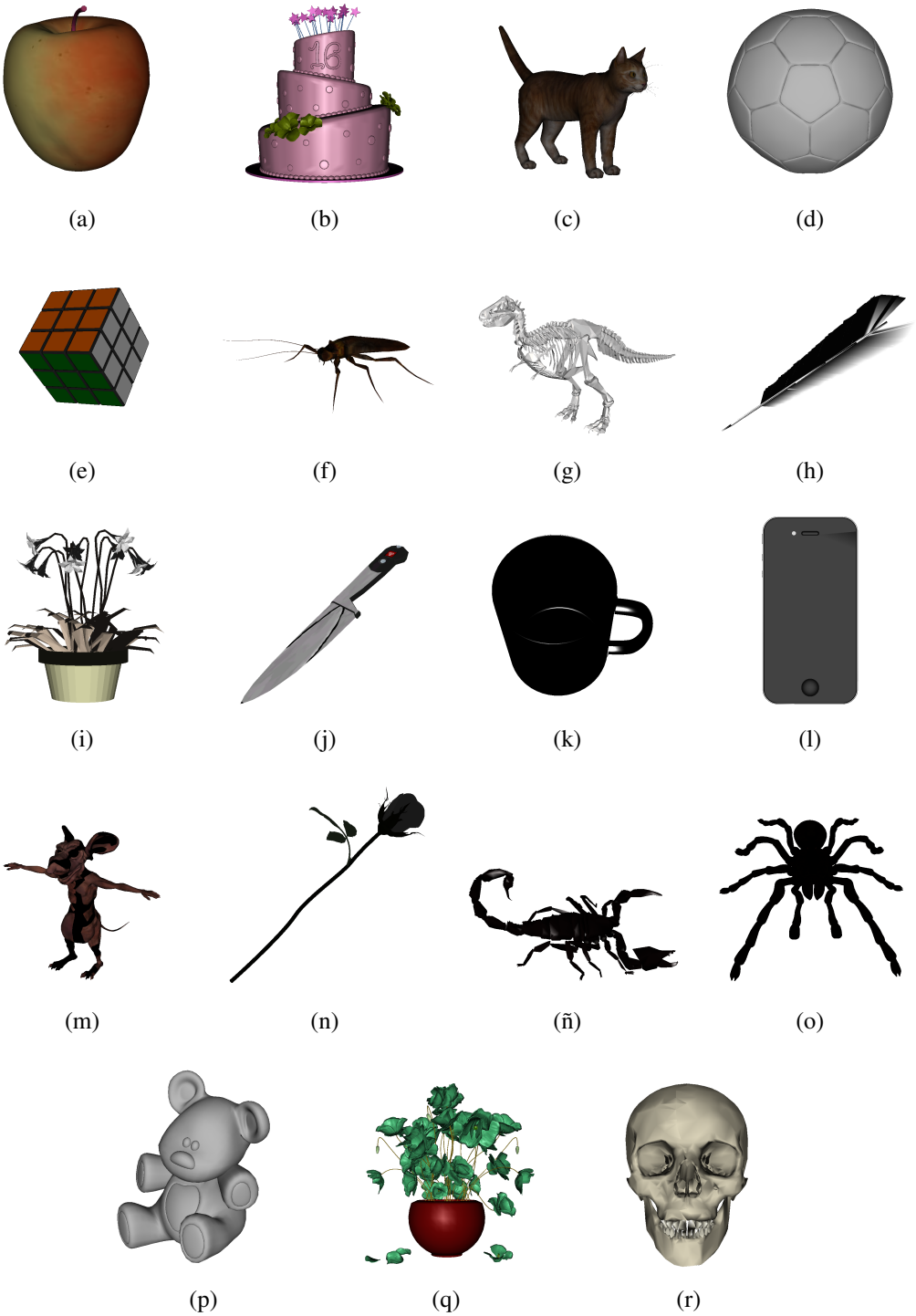


Figura B.5: Modelos en formato .3DS (3D Studio Max), utilizados en el desarrollo del algoritmo de reconocimiento de objetos descrito en el capítulo 9.

Apéndice C

RoboComp

El $framework$ RoboComp [Manso et al., 2010] es el software utilizado para controlar todos los procesos y sistemas descritos en esta Tesis Doctoral. La elección del mismo se debe a que posee una arquitectura preparada para aplicaciones relacionadas con robots autónomos en entornos no controlados, por medio de herramientas que permiten controlar los sensores, motores y hardware, junto a una robusta estructura de comunicación entre componentes basada en el $middleware$ ICE.

Este capítulo describe los componentes esenciales utilizados en este trabajo, su descripción, interfaces y características destacadas. Todo ello está accesible desde la página oficial de RoboComp [1].

C.1. Componentes de RoboComp

El uso de RoboComp permite crear nuevos componentes o utilizar aquellos ya implementados por diferentes desarrolladores. El $framework$ dispone actualmente de una gran cantidad de componentes, asociados principalmente a dispositivos de hardware, como cámaras RGB, sensores RGB-D, sensores inerciales o motores, entre otros.

A continuación, se describirán las interfaces y componentes más importantes y relevantes desarrollados a lo largo del trabajo presentado, desde de los sistemas de reconocimiento e imitación de emociones, hasta los sistemas basados en el uso de las $affordances$ emocionales.

C.2. Interfaces

En primer lugar, se introducen las interfaces que proveen acceso a los componentes relacionados con los elementos hardware utilizados:

C.2.1. JointMotor

La interfaz $JointMotor$ permite controlar los diferentes tipos de motores, por medio de los siguientes componentes:

[1] robocomp.org

- **dynamixelComp**: componente utilizado para controlar los motores "$Dynamixel$", encargados de realizar todos los movimientos del robot social Ursus.

- **faulhaberComp**: componente utilizado para controlar los motores "$Faulhaber$", encargados de los movimientos del cuello y ojos, dentro de la cabeza robótica Muecas.

- **muecasjointComp**: componente utilizado para controlar los servo-motores "$HITEC$", encargados de los movimientos del cejas y la boca, dentro de la cabeza robótica Muecas.

C.2.2. Camera

La interfaz $Camera$ proporciona la información de las cámaras RGB, por medio de los siguientes componentes:

- **camaraComp**: componente utilizado para capturar la información desde cámaras soportadas por Video4Linux en $GNU/Linux$, y cámaras Point Grey por medio del software $Coriander$ para IEEE-1394.

C.2.3. Speech

La interfaz $Speech$ proporciona la información a los sistemas de generación de texto a voz TTS, por medio de los siguientes componentes:

- **speechverbioComp**: componente utilizado para generar un mensaje verbal (voz sintética) a través de la información de un texto, por medio del software $VerbioTTS$.

- **speechGoogleComp**: componente encargado de transmitir la información relacionada al texto y los parámetros como el idioma, al servicio $online$ de $Google - TTS$.

C.2.4. ASR

La interfaz ASR proporciona y transfiere la información a los sistemas de reconocimiento de voz ASR, por medio de los siguientes componentes:

- **speechGoogleComp**: componente encargado de transmitir la información acústica de la voz humana para el proceso de reconocimiento del voz, basado en el servicio $online$ de $Google - ASR$.

C.3. Componentes

En segundo lugar, se describen los componentes relacionados con cada uno de los sistemas presentados en esta Tesis Doctoral.

C.3.1. Reconocimiento e imitación de emociones

Lista de componentes relacionados al $reconocimiento\ e\ imitación\ de\ emociones$ basado en el lenguaje natural, ya sea por medio de expresiones faciales, la voz humana o el lenguaje corporal.

- **muecasemotionComp**: componente encargado del reconocimiento de los estados emocionales del usuario, a través de la extracción y clasificación de las características faciales obtenidas por medio del filtrado de $Gabor$.
- **affordanceshumanComp**: componente encargado del reconocimiento de los estados emocionales del usuario, a través de la extracción y clasificación de las características faciales obtenidas por medio del modelo de malla $Candide$-3.
- **speechrecognitionComp**: componente encargado de reconocer los estados emocionales del usuario, a través de la extracción y clasificación de las características acústicas de la voz humana.
- **bodyrecognitionComp**: componente encargado de reconocer los estados emocionales del usuario, a través de la extracción y clasificación de características especificas, asociadas a las posiciones y movimientos de las articulaciones del cuerpo humano.
- **imitationComp**: componente encargado de analizar la información adquirida del sistema de reconocimiento de expresiones faciales, para su uso en un sistema de imitación del lenguaje corporal mediante el robot Muecas. Este componente permite la imitación en tiempo real, de las expresiones faciales y del movimiento de los elementos del cara del usuario, como la boca, las cejas y el cuello.
- **mouthComp**: componente encargado del algoritmo de sincronización entre un sistema TTS y las bocas robóticas de las plataformas Muecas y Ursus.
- **multimodalrecognitionComp**: componente encargado de reconocer los estados emocionales del usuario a través de un sistema que combina varios enfoques como las expresiones faciales y la voz humana.

C.3.2. Affordances

Lista de componentes relacionados exclusivamente a la implementación del sistema de aprendizaje basado en $affordances$ emocionales, descrito en esta Tesis Doctoral.

- **ObjectAffordancesComp**: componente encargado del reconocimiento de los objetos del entorno por medio de marcas de realidad aumentada, dentro del sistema de aprendizaje basado en $affordances$ emocionales.
- **EmotionalAffordancesComp**: componente principal dentro del sistema de aprendizaje basado en $affordances$ emocionales, el cual se encarga del entrenamiento durante el aprendizaje y la elección de los objetos en el entorno.
- **AgentAffordancesComp**: componente encargado de generar las acciones relacionadas al agente robótico Muecas, ya sean expresiones faciales, mensajes verbales por medio de una voz sintética y la búsqueda de objetos por medio de los sensores RGB de los ojos.

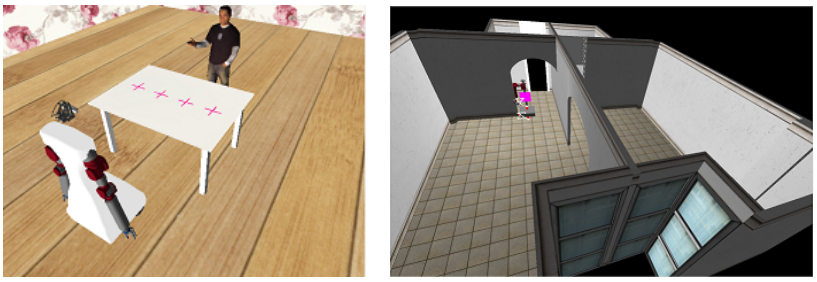
C.4. Otros componentes

Finalmente, se describen los componentes generados por el grupo de Ingeniería de Sistemas Integrados de la Universidad de Málaga y que fueron utilizados en este trabajo:

- **AttentionComp**: componente utilizado para el reconocimiento de $proto\text{-}objetos$ dentro de un escenario. Este componente fue desarrollado por A.J. Palomino [Palomino et al., 2011], y fue utilizado con una de las alternativas iniciales para el reconocimiento de objetos.
- **WinKinectComp**: componente encargado de transmitir la información relacionada al modelo de malla $Candide$-3 desde el sistema operativo $Windows$, a través del $middleware\ ICE$.

C.5. RCInnermodelSimulator

El software $InnerModelSimulator$ [Manso, 2012] es un simulador virtual incorporado dentro del $framework$ RoboComp, que contiene una serie de herramientas para la simulación de escenarios sociales con múltiples objetos y agentes robóticos, con sus respectivos sensores y actuadores. Este simulador está basado en el motor gráfico $OpenSceneGraph$ [Burns, 2014], el cual permite la simulación de múltiples experimentos realizados dentro de esta Tesis, evitando las diferentes limitación físicas del mundo real en el desarrollo de los sistemas anteriormente descritos. En la Figura C.1 se ilustra las representaciones 3D del escenario y del robot social Loki.



(a) (b)

Figura C.1: Imágenes del simulador en diferentes escenarios.

Índice alfabético

Bibliografía

[Ahlberg, 2001] Ahlberg, J. (2001). CANDIDE-3 – an updated parameterized face. Technical report, Report No. LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linköping University, Sweden.

[Aleksic and Katsaggelos, 2006] Aleksic, P. and Katsaggelos, A. (2006). Automatic Facial Expression Recognition using facial animation parameters and multistream HMMs. *IEEE Trans. Information Forensics and Security*, 1:3–11.

[Aly and Tapus, 2011] Aly, A. and Tapus, A. (2011). Speech to head gesture mapping in multimodal human-robot interaction. In *Proc. of the 5th European Conference on Mobile Robots ECMR 2011*, pages 101–108.

[Anderson and Owan, 2003] Anderson, K. and Owan, P. M. (2003). Real-time emotion recognition using biologically inspired models. In *Proceedings of the 4th international conference on Audio- and video-based biometric person authentication*, pages 119–127.

[Anderson and Kewley-Port, 1995] Anderson, S. and Kewley-Port, D. (1995). Evaluation of Speech Recognizers for Speech Training: Applications. *IEEE Transactions on speech and audio processing*, 3(4):229–241.

[Arnold, 1960] Arnold, M. (1960). *Emotions and personality*. New York: Columbia University Press.

[Atassi et al., 2011] Atassi, H., Esposito, A., and Smekal, Z. (2011). Analysis of high-level features for vocal emotion recognition. In *Proceeding of the 34th International Conference on Telecommunication and Signal Processing (TSP)*, pages 361–366.

[Bartlett et al., 2005] Bartlett, S., Littlewort, G., Fasel, I., and Movella, R. (2005). eal Time Face Detection and Facial Expression Recognition: Development and Application to human computer interacation. In *CVPR Workshop on Computer Vision and Pattern Recognition for Human-Computer Interaction*, pages 112–117.

[Bartneck, 2002] Bartneck, C. (2002). *eMuu–an emotional emboid character for the ambient intelligent home*. PhD thesis, University of Eindhoven.

[Bettadapura, 2012] Bettadapura, V. (2012). Face expression recognition and analysis: The state of art. Technical report, Tech Report, College of Computing, Georgia Institute of Technology.

[Boersma and van Heuven, 2001] Boersma, P. and van Heuven, V. (2001). Speak and unSpeak with Praat. *Glot International*, 5(9–10):341–347.

[Breazeal, 2002] Breazeal, C. (2002). *Design Sociable Robots*. MIT Press.

[Breazeal et al., 2008] Breazeal, C., Takanishi, A., and Kobayashi, T. (2008). *Social Robots that Interact with People*, pages 1349–1369. Springer Berlin Heidelberg.

[Burns, 2014] Burns, D. (2014). OSG - OpenSceneGraph.

[C. Bagwell, 2014] C. Bagwell (2014). SoX Sound eXchange.

[Calderita et al., 2011] Calderita, L., Bachiller, P., Bandera, J., Bustos, P., and Núñez, P. (2011). Mimic: A human motion imitation component for robocomp. In *Proc of Int Workshop on Recognition and Action for Scene understanding.*, pages 99–113.

[Caridakis et al., 2010] Caridakis, G., Karpouzis, K., Wallace, M., Kessous, L., and Amir, N. (2010). Multimodal user's affective state analysis in naturalistic interactions. *In Journal on Multimodal User Interfaces*, 3(1–2):49–66.

[Chemerro, 2003] Chemerro, A. (2003). An outline of a theory of affordances. *Ecological Psycology*, 15(2):181–195.

[Chen et al., 2012] Chen, L., Mao, X., and Yue, Y. (2012). Speech emotion recognition: Features and classification models. *Digital signal processing*, 22:1154–1160.

[Chen and Rao, 1998] Chen, T. and Rao, R. (1998). Audio-Visual Integration in multimodal Communication. *Proceedings of the IEEE*, 86(5):837–852.

[Childers et al., 1977] Childers, D., Skinner, D., and Kemerait, R. (1977). The Cepstrum: A Guide to Processing. *Proc. of the IEEE*, 65(10):1428–1443.

[Cid et al., 2011] Cid, F., Cintas, R., Manso, L., Calderita, L., Sánchez, A., and Núñez, P. (2011). A real-time synchronization algorithm between text-to-speech (tts) system and robot mouth for social robotic applications. In *Proceedings of the Workshop of Physical Agents (WAF2011)*, pages 81–86.

[Cid et al., 2012] Cid, F., Manso, L., Calderita, L., Sánchez, A., and Núñez, P. (2012). Engaging human-to-robot attention using conversational gestures and lip-synchronization. *Journal of Physical Agents*, 6(1):3–10.

[Cid and Núñez, 2014] Cid, F. and Núñez, P. (2014). Learning emotional affordances based on affective elements in human-robot interaction scenarios. In *Proceedings of the Workshop of Physical Agents (WAF2014)*, pages 83–92.

[Cid et al., 2013a] Cid, F., Palomino, A., and Núñez, P. (2013a). A new paradigm for learning affective behavior: Emotional affordances in human robot interaction. In *Proceedings of the Workshop of Physical Agents (WAF2013)*, pages 47–52.

[Cid et al., 2013b] Cid, F., Prado, J., Bustos, P., and Núñez, P. (2013b). A Real Time and Robust Facial Expression Recognition and Imitation Approach for Human-Robot interaction using gabor filtering. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2188–2193.

[Cid et al., 2014] Cid, F., Prado, J., Bustos, P., and Núñez, P. (2014). Muecas: A Multi-Sensor Robotics Head for Affective Human Robot Interaction and Imitation. *Sensors*, 14(5):7711–7737.

[Cid et al., 2013c] Cid, F., Prado, J., Manzano, P., Bustos, P., and Núñez, P. (2013c). Imitation System for Humanoid Robotics Heads. *Journal of Physical Agents*, 7(1):22–29.

[Cooley and Tukey, 1965] Cooley, J. and Tukey, J. (1965). On Algorithm for the machine Calculation of Complex Fourier Series. *Mathematics of Computation*, 19(90):297–.301.

[Cos-Aguilera et al., 2003] Cos-Aguilera, I., Canamero, L., and Hayes, G. (2003). Motivation-driven learning of object affordances: First experiments using a simulated khepera robot. In *Proceedings of 9th International Conference in Cognitive Modelling (ICCM)*, pages 57–62.

[Cowie and Cornelius, 2003] Cowie, R. and Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32.

[Cuadra et al., 2001] Cuadra, P. D. L., Master, A., and Sapp, G. (2001). Efficient pitch detection techniques for interactive music. In *In Proceedings of the 2001 International Computer Music Conference*.

[Culjak et al., 2012] Culjak, I., Abram, D., Pribanic, T., Dzapo, H., and Cifrek, M. (2012). A brief introduction to opencv. In *Proceedings of the 35th International Convention MIPRO*, pages 1725 – 1730.

[Damasio, 2003] Damasio, A. (2003). *Looking for Spinoza*. Harcourt Brace & Co.

[Darwin, 1872] Darwin, C. (1872). *The Expression of Emotions in Man and Animals*. London: John Murray.

[Descartes, 1647] Descartes, R. (1647). *The Passions of the Soul*. Cottingham.

[Doblado et al., 2013] Doblado, C., mogena, E., Cid, F., Calderita, L., and Núñez, P. (2013). Rgb-d database for affective multimedia human-robot interaction. In *Proceedings of the Workshop of Physical Agents (WAF2013)*, pages 35–40.

[Ekman, 1999] Ekman, P. (1999). *Basic emotions*. Wiley: New York.

[Ekman and Friesen, 1971] Ekman, P. and Friesen, W. (1971). Constants across cultures in the face and emtions. *Personality Social Psychology*, 17(2):124–129.

[Ekman et al., 2002] Ekman, P., Friesen, W., and Hager, J. (2002). Facial action coding system FACS. Technical report, The manual.

[Ekman et al., 1983] Ekman, P., Levenson, R., and Friesen, W. (1983). Autonomic Nervous System Activity Distinguishes Among Emotions. *Science*, 221(4616):1208–1210.

[Fitzpatrick et al., 2003] Fitzpatrick, P., Metta, G., Natale, L., Rao, S., and Sandini, G. (2003). Learning about objects through action - initial steps towards artificial cognition. In *Proceedings of the 2003 IEEE International Conference on Robotics & Automation*, pages 3140–3145.

[Fong et al., 2003] Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42(3–4):143–166.

[Frijda, 1986] Frijda, N. (1986). *The emotions*. New York: Cambridge University Press.

[Gaver, 1991] Gaver, W. (1991). Technology affordances. In *Proceedings of the Conference Human Factors in Computing Systems, CHI'91*, pages 79–84.

[Ge et al., 2008] Ge, S., Wang, C., and Hang, C. (2008). Facial Expression Imitation in Human Robot Interaction. In *17 IEEE International Symposium on Robot and Human Interactive Communication*, pages 213–218.

[Gibson, 2000] Gibson, E. (2000). Perceptual learning in development: Some basic concepts. *Ecological Psycology*, 12(4):295–302.

[Gibson, 2003] Gibson, E. (2003). The world is so full of a number of thing: On specification and perceptual learning. *Ecological Psycology*, 15(4):283–288.

[Gibson, 1979] Gibson, J. (1979). *The ecological approach to visual perception*. Boston:Houghton Mifin.

[Gonzalez-Sanchez and Puig, 2011] Gonzalez-Sanchez, T. and Puig, D. (2011). Real-time body gesture recognition using depth camera. *In Electronics Letter*, 47:697–698.

[Gray, 1982] Gray, J. (1982). *The neuropsychology of anxiety*. Oxford:Oxford University Press.

[Haq and Jackson, 2010] Haq, S. and Jackson, P. (2010). *Multimodal Emotion Recognition*, chapter 17, pages 398–423. Machine Audition: Principles, Algorithms and Systems. IGI Global Press.

[Hara et al., 1997] Hara, F., Endou, K., and Shirata, S. (1997). Lip-configuration control of a mouth robot for japanese vowels. In *Proc. IEEE International Workshop on Robot and Human Communicatio*, pages 412–418.

[Harris, 1978] Harris, F. (1978). On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. *Proc. of the IEEE*, 6(1):51–83.

[Henning and Spruiell, 2005] Henning, M. and Spruiell, M. (2005). Ice - Internet Communications Engine.

[Hermans et al., 2011] Hermans, T., Rehg, J., and Bobick, A. (2011). Affordance Predictive via Learned Object Attributes. In *Proc. of IEEE International Conference on Robotics and Automation (ICRA): Workshop on Semantic Perception, Mapping, and Exploration*.

[HITLabNZ, 2014] HITLabNZ, H. I. T. L. N. Z. (2014). OSGART - ARToolKit for OpenSceneGraph.

[Hoult, 2004] Hoult, C. (2004). *Emotion in Speech Synthesis*.

[Huang and Mutlu, 2014] Huang, C.-H. and Mutlu, B. (2014). Learning-based modeling of multimodal behaviors for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*, pages 57–64.

[Human Interface Technology Laboratory, 2014] Human Interface Technology Laboratory, U. o. W. (2014). ARToolKit - Library for building Augmented Reality (AR) applications.

[Iliou and Anagnostopoulos, 2010] Iliou, T. and Anagnostopoulos, C.-N. (2010). Svm-mlp-pnn classifiers on speech emotion recognition field- a comparative study. In *Proceeding of the Fifth International Conference on Digital Telecommunications*, pages 1–6.

[Iriondo et al., 2000] Iriondo, I., Guaus, R., Rodriguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., and Longth, L. (2000). Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. In *Proceedings of the International Symposium on Computer Architecture*, pages 161–166.

[Izard, 1971] Izard, C. (1971). *The face of emotion*. Appleton-Century-Crofts.

[Jaimes and Sebe, 2005] Jaimes, A. and Sebe, N. (2005). Multimodal human computer interaction: A survey. In *IEEE International Workshop on Human Computer Interaction in conjuntion whith ICCV 2005*.

[James, 1884] James, W. (1884). What is an emotion? *Mind*, 9(34):188–205.

[J.Cahn, 1990] J.Cahn (1990). Generation Expression in synthesized speech.

[Jiang et al., 2012] Jiang, T., Zhang, L., and Zhang, X. (2012). The Research of the Face's Depth Information Generation Technology Based on the Candide Model. *Advances in Multimedia Information Processing – PCM 2012, Lecture Notes in Computer Science*, 7674:823–831.

[k. Dautenhahn and Nehaniv, 2002] k. Dautenhahn and Nehaniv, C. (2002). *Imitation in Animals and Artifacts*. MIT Press:London.

[Kamarainen, 2012] Kamarainen, J. (2012). Gabor features in image analysis. In *3rd International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 13–14.

[Kato and Billinghurts, 1999] Kato, H. and Billinghurts, M. (1999). Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *Proceedings of the 2nd IEEE and ACM International Workshop of Augmented Reality (IWAR'99)*, pages 85–94.

[Katz et al., 2013] Katz, D., Venkatraman, A., Kazemi, M., Bagnell, J., and Stentz, A. (2013). Perceiving, learning, and exploiting object affordances for autonomous pile manipulation. In *Robotics: Science and Systems Conference*.

[Kessous et al., 2010] Kessous, L., Castellano, G., and Caridakis, G. (2010). Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *In Journal on Multimodal User Interfaces*, 3(1):33–48.

[Koppula et al., 2013] Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *International Journal of Robotics Research*, 32:951–970.

[Kotsi et al., 2008] Kotsi, I., Buciu, I., and Pitas, I. (2008). An analysis of Facial expression recognition under partial facial image occlusions. *Imaging and Computer Vision* , 26(7):1052–1067.

[Krauss et al., 1991] Krauss, R., Morrel-Samuels, P., and Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61(5):743–754.

[Kurtenbach and Hulteen, 1992] Kurtenbach, G. and Hulteen, E. (1992). Gesture in Human-Computer Communication. *The Art of Human-Computer Interface*, pages 309–317.

[Laban, 1980] Laban, R. (1980). *The Mastery of Movement*. London: MacDonald and Evans.

[Laban and Lawrence, 1947] Laban, R. and Lawrence, F. (1947). *Effort*. London: MacDonald and Evans.

[Lang et al., 1990] Lang, P., Bradley, M., and Cuthbert, B. (1990). Emotion attention and the startle reflex. *Psychological review*, 97(3):377–395.

[Lee et al., 2009] Lee, M., Forlizzi, J., Rybski, P., Crabbe, F., Chung, W., Finkle, J., Glaser, E., and Kiesler, S. (2009). The snackbot: Documenting the design of a robot for long-term human-robot interaction. In *In Proc. of HRI 2009*, pages 7–14.

[Looser et al., 2006] Looser, J., Grasset, R., Seichter, H., and Billinghurst, M. (2006). OSGART A Pragmatic Approach to MR. In *Proc. of the IEEE and ACM International Symposium on Mixer and Augmented Reality (ISMAR'06)*.

[Lopes et al., 2007] Lopes, M., Melo, F. S., and Montesano, L. (2007). Affordance-based imitation learning in robot. In *Proceedings of the IEEE/RSJ International Conference on Robots and Systems*, pages 1015–1O21.

[M. Kammer and Nagai, 2011] M. Kammer, M. Tscherepanow, T. S. and Nagai, Y. (2011). From Affordances to Situated Affordances in Robotics - Why Context is Important. *Frontiers in Computational Neuroscience*, (30):1662–5188.

[MacDorman, 2000] MacDorman, K. (2000). Responding to affordances: Learning and projecting a sensorimotor mapping. In *Proceedings of IEEE International Conference in Robotics and Automation (ICRA)*, pages 3253–3259.

[Mancas et al., 2010] Mancas, M., Glowinski, D., Volpe, G., Coletta, P., and Camurri, A. (2010). Gesture Saliency: a Context-aware Analysis. *In Gesture in Embodied Communication and Human-Computer Interaction - Lecture Notes in Computer Science*, 5934:146–157.

[Manso et al., 2010] Manso, L., Bachiller, P., Bustos, P., Núñez, P., Cintas, R., and Calderita, L. (2010). Robocomp: a tool-based robotics framework. In *Simulation, Modeling and Programming for Autonomous Robots, SIMPAR2010*, pages 251–262.

[Manso, 2012] Manso, L. J. (2012). InnerModel*: InnerModel, InnerModelViewer, RCInnerModelEditor and RCInnerModelSimulator. RoboComp's Wiki Tutorials.

[Mayer et al., 2009] Mayer, C., Wimmer, M., Eggers, M., and Radig, B. (2009). Facial expression recognition with 3d deformable models. In *Second International Conferences on Advances in Computer-Human Interactions - ACHI 09*, pages 26–31.

[McDougall, 1926] McDougall, W. (1926). *An introduction to social psychology*. Boston:Luce.

[McGrenere and Ho, 2000] McGrenere, J. and Ho, W. (2000). Affordances: Clarifying and evolving a concept. In *Proceedings of the Conference Graphics interfaces, GI 2000*, pages 179–186.

[McNeill, 1992] McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press.

[Meeren et al., 2005] Meeren, H., van Heijnsbergen, C., and B, d. G. (2005). Rapid perceptual integration of facial expression and emotional body language. *Proc. National Academy of Sciences of the USA*, 102(45):16518–16523.

[Mejías et al., 2013] Mejías, C. S., Echevarría, C., Nuñez, P., Manso, L., Bustos, P., Leal, S., and Parra, C. (2013). Ursus: A Robotic Assistant for Training of Children with Motor Impairments. *Converging Clinical and Engineering Research on Neurorehabilitation Biosystems & Biorobotics*, 1:249–253.

[Microsoft, 2014] Microsoft (2014). Kinect for Windows SDK.

[Moberg, 2007] Moberg, M. (2007). *Contributions to Multilingual Low-footprint TTS System for Hand-held Devices*. Tampere University of Technology.

[Montero et al., 1999] Montero, J., Gutierrez-Arriola, J., Colas, J., Enriquez, E., and Pardo, J. (1999). Analysis and modelling of emotional speech in Spanishn. In *Proceedings of the 14th International Conference on Phonetic*, pages 957–960.

[Montesano et al., 2007] Montesano, S., Lopes, M., Bernardino, A., and Santos-Victor, J. (2007). Modeling Affordances using Bayesian networks. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4102–4107.

[Montesano et al., 2008] Montesano, S., Lopes, M., Bernardino, A., and Santos-Victor, J. (2008). Learning object affordances: From sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26.

[Mori, 1970] Mori, M. (1970). Bukimi no tani – The uncanny valley. *Energy*, 7(4):33–35.

[Morie et al., 2005] Morie, J., Williams, J., Dozois, A., and Luigi, D. (2005). The fidelity of "feel": Emotional affordance in virtual environments. In *Proceedings of the 11th International Conference on Human-Computer Interaction*.

[Mowrer, 1960] Mowrer, O. (1960). *Learning theory and behavior*. New York:Willey.

[Murray and Arnott, 1993] Murray, I. and Arnott, J. (1993). Toward the simulation of emotion in synthetic speech: A review of literature on human vocal emotion. *Journal of Acoustical Society of America*, 93(2):1097–1108.

[Nakata et al., 1998] Nakata, T., Sato, T., and Mori, T. (1998). Expression of emotion and intention by robot movement. In *Proceedings of the 5th International Conference on Autonomous Systems*, pages 352 – 359.

[Nogueiras et al., 2001] Nogueiras, A., Marino, J., Moreno, A., and Bonafonte., A. (2001). Speech emotion recognition using hidden markov models. In *European Conf. on Speech Communication and Technology (Eurospeech 01)*.

[Noll, 1970] Noll, M. (1970). Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum, and a maximum likelihood estimate. In *Proceedings of the Symposium on Computer Processing in Communications, Vol. XIX*, pages 779–797.

[Norman, 2002] Norman, D. (2002). *The Psychology of Everyday Things*. Basic Books, Inc., New York, NY, USA.

[Norman, 2004] Norman, D. (2004). *Emotional design: why we love (or hate) everyday things*. New York: Basic Books.

[Norman, 1999] Norman, D. A. (1999). Affordance, Conventions, and Design. *interactions*, 6(3):38–43.

[Oatley and Johnsin-Laird, 1987] Oatley, K. and Johnsin-Laird, P. (1987). Towards a cognitive theory of emotions. *Cognition & Emotion*, 1(1):29–50.

[Oh et al., 2010] Oh, K., Jung, C., Lee, Y., and Kim, S. (2010). Real time lip synchronization between text to speech(tts) system and robot mouth. In *Proc. of IEEE International Symposium on Robot and Human Interactive Communication*, pages 620–625.

[Olson, 2011] Olson, E. (2011). AprilTag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3400–3407.

[OpenNi Organization, 2014] OpenNi Organization (2014). OpenNi Open Natural Interaction.

[Orabona et al., 2007] Orabona, F., Metta, G., and Sandini, G. (2007). *A Proto-object Based Visual Attention Model*, volume 4840 of *Lecture Notes in Computer Science*, pages 198–215. Springer Berlin Heidelberg.

[Ortony and Turner, 1990] Ortony, A. and Turner, T. (1990). What's basic about basic emotions? *Psychology Review*, 97(2):315–331.

[Osório et al., 2010] Osório, P., Bernardino, A., Martinez-Cantin, R., and Santos-Victor, J. (2010). Gaussian Mixture Models for Affordance Learning using Bayesian Networks. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4432–4437.

[Otsuka and Ohya, 1997] Otsuka, T. and Ohya, J. (1997). Recognizing multiple persons' facial expressions using HMM based on automatic extraction of significant frames from image sequences. In *Proceedings on International Conference on Image Processing*, pages 546–549.

[P. Boersma and D. Weenink, 2014] P. Boersma and D. Weenink (2014). Praat:-doing phonetics by computer. Phonetic Sciences, University of Amsterdam.

[Paiva et al., 2004] Paiva, A., Dias, J., Sobral, D., Aylett, R., Woods, S., and Hall, L. (2004). Caring for agents and agents that care: Building empathic relations with synthetic agents. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagents System*, pages 194–201.

[Palomino et al., 2011] Palomino, A., Marfil, R., Bandera, J., and Bandera, A. (2011). A novel biologically inspired attention mechanism for a social robot. *Journal on Advances in Signal Processing*, 2011:1–10.

[Panksepp, 1982] Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *The Behavioral and Brain Sciences*, 5(3):407–467.

[Parrott, 2001] Parrott, W. (2001). *Emotions in Social Psychology*. Psychology Press.

[Peelen and Downing, 2007] Peelen, M. and Downing, P. (2007). The neural basis of visual body perception. *Nature Reviews Neuroscience*, 8(8):636–648.

[Piaget, 1952] Piaget, J. (1952). *The origins of intelligent in children*. Intl. Univ. Press.

[Picard, 2000] Picard, R. (2000). *Affective Computing*, pages 88–91. MIT Press.

[Plutchik, 2002] Plutchik, R. (2002). *Emotions and Life: Perspectives from Psychology, Biology, and Evolution*. Washington, DC: American Psychological Association.

[Plutchik and Kellerman, 1980] Plutchik, R. and Kellerman, H. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*, 1(2):3–33.

[Prado, 2012] Prado, J. (2012). *A New Probabilistic Methodology to Support an Emotive Dialog between a Human and a Robot*. PhD thesis, University of Coimbra.

[Prado et al., 2011] Prado, J., Simplicio, C., Lori, N., and Diaz, J. (2011). Visuo-auditory Multimodal Emotional Structure to improve Human-Robot-Interaction. *International Journal of Social Robotics*, 4(1):29–51.

[Riaz et al., 2009] Riaz, Z., Mayer, C., Beetz, M., and Radig, B. (2009). Model Based Analysis of Face Images for Facial Feature Extraction. *Computer Analysis of Images and Patterns - Lecture Notes in Computer Science*, 5702:99–106.

[Rockmore, 2000] Rockmore, D. (2000). The FFT - an algorithm the whole family can use. *IEEE Computing in Science & Engineering*, 2(1):60–64.

[Romero et al., 2013] Romero, P., Cid, F., and Núñez, P. (2013). A novel real time facial expression recognition system based on candide-3 reconstruction model. In *Proceedings of the Workshop of Physical Agents (WAF2013)*, pages 41–46.

[Russell, 1980] Russell, J. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

[Rybski et al., 2007] Rybski, P., Yoon, K., Stolarz, J., and Veloso, M. (2007). Interactive robot task training through dialog and demonstration. In *In Proc. of HRI 2007*.

[Sahin et al., 2007] Sahin, E., Çakmak, M., Dogar, M., Ugur, E., and Üçoluk, G. (2007). To Afford or Not to Afford: A New Formalization of Affordances Toward Affordances-Based Robot Control. *Adaptative Behavior*, 15(4):447–472.

[Savva et al., 2011] Savva, N., Scarinzi, A., and Bianchi-Berthouze, N. (2011). Continuous Recognition of Player's Affective Body Expression as Dynamic Quality of Aesthetic Experience. *In IEEE Trans. Computational Intelligence and AI in games*, 4(3):199–212.

[Schlossberg, 1954] Schlossberg, H. (1954). Three dimensions of emotion. *Psychology Review*, 61(2):81–88.

[Schuller et al., 2004] Schuller, B., Rigoll, G., and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Proceeding of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–577–I–580.

[Schulte et al., 1999] Schulte, J., Rosenberg, C., and Thrun, S. (1999). Spontaneous, short-term interaction with mobile robots. In *International Conference on Robotics and Automation*.

[Sebe et al., 2005] Sebe, N., Cohen, I., Gevers, T., and Huang, T. (2005). Multimodal approaches for emotion recognition: a survey. In *Proceeding of the Internet Imaging VI, SPIE IV, Volume 5670*, pages 56–67.

[Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1297–1304.

[Spinoza, 1677] Spinoza, B. (1677). *Ethics*. Hafner Pub.

[Steedman, 2002] Steedman, M. (2002). Plans, affordances, and combinatory grammar. *Linguistics and Philosophy*, 25(5–6):723–753.

[Stoffregen, 2003] Stoffregen, T. A. (2003). Affordances as properties of the animal environment system. *Ecological Psycology*, 15(2):115–134.

[Stoytchev, 2005] Stoytchev, A. (2005). Toward learning the binding affordances of objects: A behavior-grounded approach. In *Proceedings of AAAI Symposium on Developmental Robotics*, pages 17–22.

[Sun et al., 2009] Sun, J., Moore, J., Bobick, A., and Rehg, J. (2009). Learning visual object categories for robot affordance prediction. *International Journal of Robotics Research*, 9:147–197.

[Szokolszky, 2003] Szokolszky, A. (2003). An interview with Eleanor Gibson. *Ecological Psycology*, 15(4):271–281.

[Tan and Triggs, 2010] Tan, X. and Triggs, B. (2010). Enhanced Local Texture Feature Sets for Face Recognition Under Difficult Lighting Conditions. *IEEE Transactions on Image Processing*, 19(6):1635–1650.

[Tomkins, 1984] Tomkins, S. (1984). Affect theory. *Approaches to emotion*, pages 163–195.

[Tsapatsoulis et al., 1999] Tsapatsoulis, N., Fellenz, W., Taylor, J., and Kollias, S. (1999). *Comparing Template-based, Feature-based and Supervised Classification of Facial Expressions from Static Images*. World Scientific and Engineering Society Press.

[Turvey, 1992] Turvey, M. T. (1992). Affordances and prospective control: an outline of the ontology. *Ecological Psycology*, 4(3):173–187.

[Vaughan, 1997] Vaughan, L. (1997). Understanding movement. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 548–549.

[Verbio Technologies, 2014] Verbio Technologies (2014). Text to Speech (TTS) and Speech Recognition (ASR).

[Ververidis and Kotropoulos, 2006] Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: resources, features, and methods. *Speech Communication*, 48:1162–1181.

[Viola and Jones, 2004] Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

[Watson, 1930] Watson, J. (1930). *Behaviorism.* Chicago: University of Chicago Press.

[Weiner and Graham, 1984] Weiner, B. and Graham, S. (1984). An attributional approach to emotional development. *Emotions, cognition, and behavior*, pages 167–191.

[Whissell et al., 1986] Whissell, C., Fournier, M., Pelland, R., Weir, D., and Makarec, K. (1986). A Dictionary of Affect in Language: IV Reliability, Validity, and Application. *Development and Psychopathology*, 62:875–888.

[Zeng et al., 2008] Zeng, M., Pantic, M., Roisman, G., and Huang, T. (2008). A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions. In *Robotics and Mechatronics Conference ROBOMEC2007*, pages 2A1–O10.

[ZeroC, 2014] ZeroC (2014). ZeroC IceStorm.