Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza
1542

Máster en Ingeniería de Sistemas e Informática
Programa Oficial de Posgrado en Ingeniería Informática

Evaluación de un Sistema Multimodal de Reconocimiento de Emociones

Autor:
Sergio Ballano Pablo

Directoras:
Dra. Sandra Baldassarri
Dra. Eva Cerezo

Curso 2010/2011
Zaragoza, Septiembre 2011

GIGA Affective Lab

ita
INSTITUTO TECNOLOGICO DE ARAGON

A Laude y Aniceto, mis padres

Agradecimientos

Quiero agradecer a todas las personas que me han apoyado, de una forma u otra, en la elaboración de este trabajo.

Al Instituto Tecnológico de Aragón que me ha permitido compatibilizar mi desarrollo profesional con la realización del máster.

A mis tutoras Eva y Sandra, que me han orientado en el desarrollo del Trabajo Fin de Máster y a Isabelle, por mostrarme la computación afectiva y sus posibilidades.

A todas las personas que han participado mediante la grabación de datos para este trabajo o la realización de los test de anotación.

También quiero dar las gracias a mi familia y amigos, que me han apoyado, comprendido y aconsejado sin pedir nada a cambio.

Evaluación de un Sistema Multimodal de Reconocimiento de Emociones

Resumen

En el presente Trabajo Fin de Máster se ha llevado a cabo la implementación y evaluación de un algoritmo de fusión multimodal para detección de emociones propuesto anteriormente. Ello ha implicado la realización de las siguientes tareas:

La mejora del módulo de detección facial previo, permitiendo la detección de expresiones faciales de forma automática y posibilitando el procesado de largas secuencias de vídeo.

El desarrollo de un módulo de detección de emociones en texto, basado en el diccionario de Whissell, capaz de proporcionar una salida emocional en el espacio continuo.

El desarrollo de una aplicación, en forma de mensajería instantánea, que permite a dos usuario interactuar emocionalmente por medio de expresiones faciales (vídeo), texto y emoticonos y que recoge todos los datos necesarios para el posterior cálculo de las salidas emocionales de cada uno de los módulos estudiados.

La implementación del algoritmo de fusión y la prueba de su funcionamiento gracias a los módulos faciales, de texto y a los emoticonos.

El diseño de un método de evaluación para sistemas de detección de emociones continuas y el desarrollo de todas las herramientas necesarias para su aplicación. Dicho método ha sido empleado para estudiar el efecto de la fusión multimodal en las tasas de acierto.

El análisis de los resultados obtenidos, identificando las virtudes y los defectos tanto del algoritmo de fusión como del módulo de detección facial.

Publicaciones Relacionadas

TÍTULO: «Scalable multimodal fusion for continuous affect sensing»
AUTORES: Isabelle Hupont, Sergio Ballano, Sandra Baldassarri, Eva Cerezo
TIPO DE PARTICIPACIÓN: Presentación oral en conferencia internacional
CONFERENCIA: Symposium Series on Computational Intelligence
PUBLICACIÓN: IEEE Workshop on Affective Computational Intelligence (WACI)
ISBN: 978-1-61284-083-3
LUGAR: Paris (Francia)
FECHA: 11-15 Abril 2011
ESTADO: Presentado y publicado

TÍTULO: «Continuous Facial Affect Recognition from Videos»
AUTORES: Sergio Ballano, Isabelle Hupont, Eva Cerezo, Sandra Baldassarri
TIPO DE PARTICIPACIÓN: Presentación oral en conferencia internacional
CONFERENCIA: XII Congreso de Interacción Persona-Ordenador (Interacción 2011)
LUGAR: Lisboa (Portugal)
FECHA: 2-5 Septiembre 2011
ESTADO: Presentado

TÍTULO: «Recognizing Emotions from Video in a Continuous 2D Space»
AUTORES: Sergio Ballano, Isabelle Hupont, Eva Cerezo, Sandra Baldassarri
TIPO DE PARTICIPACIÓN: Póster
CONFERENCIA: 13th IFIP TC13 Conference on Human-Computer Interaction (Interact 2011)
LUGAR: Lisboa (Portugal)
FECHA: 5-9 Septiembre 2011
ESTADO: Presentado

TÍTULO: «Emotional Facial Sensing and Multimodal Fusion in a Continuous 2D Affective Space»
AUTORES: Eva Cerezo, Isabelle Hupont, Sandra Baldassarri, Sergio Ballano
TIPO DE PARTICIPACIÓN: Artículo en revista
REVISTA: Journal of Ambient Intelligence and Humanized Computing
ESTADO: Aceptado
FECHA DE ACEPTACIÓN: 29 de Agosto de 2011

TÍTULO: «From A Discrete Perspective Of Emotions To Continuous, Dynamic And Multimodal Affect Sensing»
AUTORES: Isabelle Hupont, Sergio Ballano, Eva Cerezo, Sandra Baldassarri
TIPO DE PARTICIPACIÓN: Capítulo de libro «Advances in Emotion Recognition»
EDITORIAL: Wiley-Blackwell
FECHA DE ENTREGA: 23 de Junio de 2011
ESTADO: En revisión

Índice general

CAPÍTULO 1

Introducción

La computación afectiva busca mejorar la tradicional forma de interacción persona-ordenador permitiendo conocer el estado emocional del usuario. Esto abre posibilidades como ofrecer diferentes respuestas acordes al estado de ánimo del usuario o conocer las reacciones del usuario ante ciertos estímulos o situaciones.

La comunicación de emociones entre personas se realiza de forma natural a través de diversos canales (tono de voz, expresiones faciales, etc). En este sentido, los esfuerzos actuales por mejorar la fiabilidad de los sistemas de detección de emociones se encaminan hacia la multimodalidad, es decir; emplear simultáneamente información procedente de diversas fuentes para obtener una mejor comprensión del estado emocional del usuario.

Actualmente la detección multimodal es un campo abierto de investigación que presenta varios retos siendo el más apreciable la fusión de información procedente de medios muy diversos, como señales fisiológicas (conductividad de la piel, ritmo cardiaco, etc) con imágenes de vídeo (gestos faciales, movimientos de cabeza y expresiones corporales), audio, texto escrito, etc. Otros problemas son la carencia de bases de datos multimodales anotadas y la complejidad de evaluar los resultados obtenidos.

1.1. Contexto

Siguiendo la línea de investigación abierta por Hupont [1] se pretende obtener un sistema fiable de detección de emociones, mejorando las tasas de acierto de sistemas anteriores. Para ello se desarrollará y evaluará un sistema multimodal cuyo esquema se muestra en la figura 1.1.

El sistema parte de la información recogida a través de varios canales (o modalidades), esta información se procesa mediante una serie de módulos emocionales, que pueden haber sido desarrollados de forma independiente. En el caso concreto de este

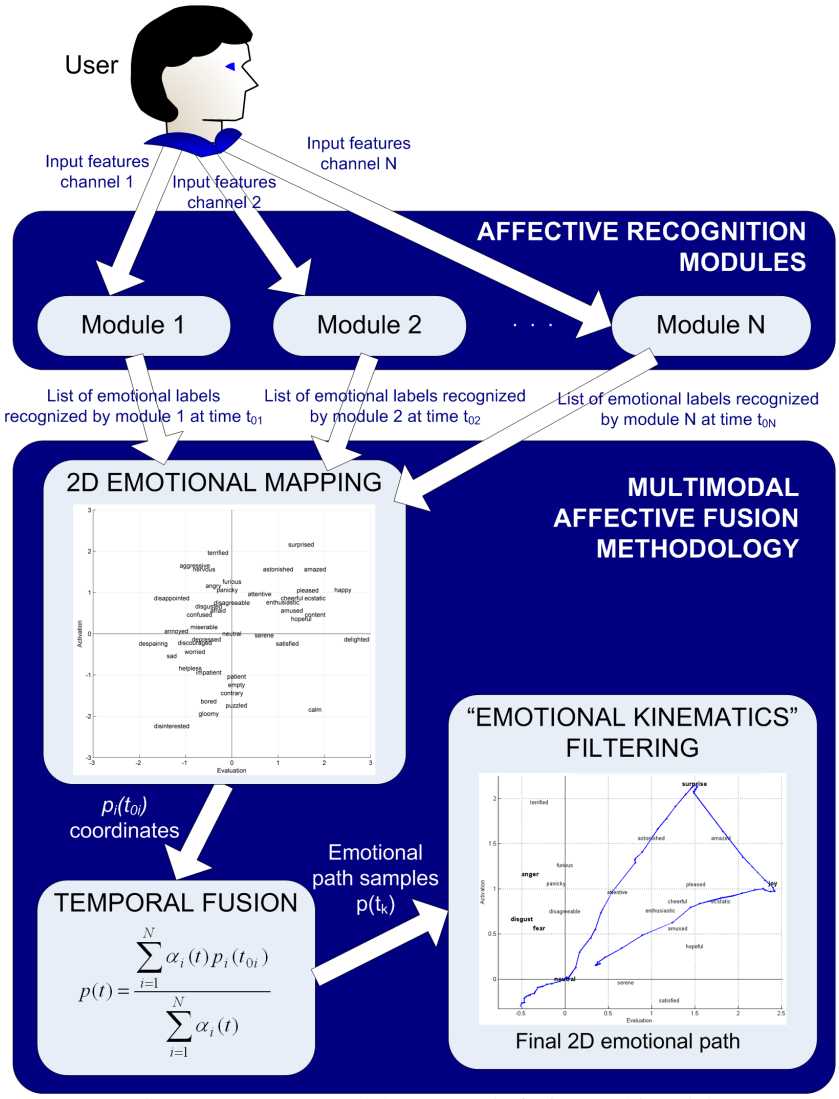

Figura 1.1: Esquema del proceso de fusión multimodal

Trabajo Fin de Máster se trabajará con un módulo facial, uno de texto y un módulo de *emoticonos*, si bien sería posible integrar cualquier número de módulos. Como se verá, los resultados de estos módulos, que inicialmente pueden ser muy heterogéneos, se representan en un espacio continuo 2D, se fusionan y, por último, se filtran obteniendo la salida definitiva del sistema.

1.2. Objetivos

El objetivo principal del Trabajo Fin de Máster consiste en validar el comportamiento del algoritmo de fusión en condiciones reales y valorar las posibles mejoras en las tasa de detección emocional gracias a la fusión multimodal.. Para alcanzar dicho objetivo se realizarán los siguientes pasos:

Se mejorará un «módulo facial», desarrollado previamente, de forma que permita procesar un fichero de vídeo detectando automáticamente los puntos característicos de la cara y proporcione una salida emocional en base a ellos (Sección 3.2.1). Este módulo será entrenado y evaluado individualmente empleando herramientas desarrolladas para tal fin de forma que posteriormente pueda ser comparado con la evaluación global del algoritmo de fusión.

Se desarrollará un «módulo de texto» que proporcione una salida emocional para cada frase de un texto (Sección 3.2.2) y un módulo de emoticonos (Sección 3.2.3) de forma que el usuario pueda reportar directamente su estado emocional si así lo desea.

Se implementará el algoritmo de fusión descrito en el Capítulo 3 que será capaz de aceptar las entradas de los módulos faciales, de texto y de emoticonos, pero también las de cualquier otro futuro módulo que se desee agregar posteriormente.

A fin de recoger los datos necesarios para la evaluación se desarrollará una aplicación de mensajería instantánea (Sección 4.1) que registre la comunicación entre dos usuarios mediante vídeo, texto y emoticonos.

Se propondrá un método de evaluación del algoritmo de fusión (Capítulo 5) y se desarrollarán las herramientas necesarias para llevarlo a cabo (Sección 4.2).

Finalmente, se analizarán los resultados obtenidos (Capítulo 5) y se propondrán mejoras al sistema (Capítulo 6).

CAPÍTULO 2

Estado del Arte

Previamente a la descripción del trabajo realizado es necesario tener en cuenta los antecedentes existentes en el área de la computación afectiva. En la sección 2.1 se presenta la evolución de los sistemas de detección de emociones, de los discretos a los continuos. En 2.2 se muestra cómo se combinan distintos sistemas a fin de mejorar los resultados obtenidos. Por último, en 2.3, se presentan los métodos existentes de evaluación de los métodos de detección de emociones.

2.1. Detección de Emociones

Tradicionalmente los métodos de detección empleados se han limitado a clasificar las emociones según categorías; es especialmente popular la clasificación de Ekman [2] y sus 6 emociones básicas (asco, alegría, ira, miedo, sorpresa y tristeza), por ejemplo Hammal [3] y Littlewort [4], aunque también existen otros que amplían el número de clases con emociones secundarias [5] o estados mentales tales como «concentrado», «interesado» y «pensativo» [6, 7]. Otros sin embargo, reducen la detección simplemente a la detección de dos emociones contrapuestas [8]. No obstante, todas estas aproximaciones suponen una representación discreta, sin ninguna relación entre ellas y que no son capaces de reflejar el amplio rango de emociones complejas que un ser humano es capaz de expresar.

Para superar este problema, investigadores como Whissell [9] y Plutchik [10] prefieren ver los estados afectivos no de forma individual sino relacionados entre sí. Estos autores consideran las emociones como un espacio continuo bidimensional, cuyos ejes representan la evaluación y la activación de cada emoción. El eje de evaluación representa cómo de positiva o negativa es una emoción, mientras que el eje de activación representa lo probable o improbable de que una persona lleve a cabo una acción influida por ese estado de ánimo, desde activa hasta pasiva. En la figura 2.1 se muestra el espacio Whissell así como la localización de varias etiquetas emocionales.

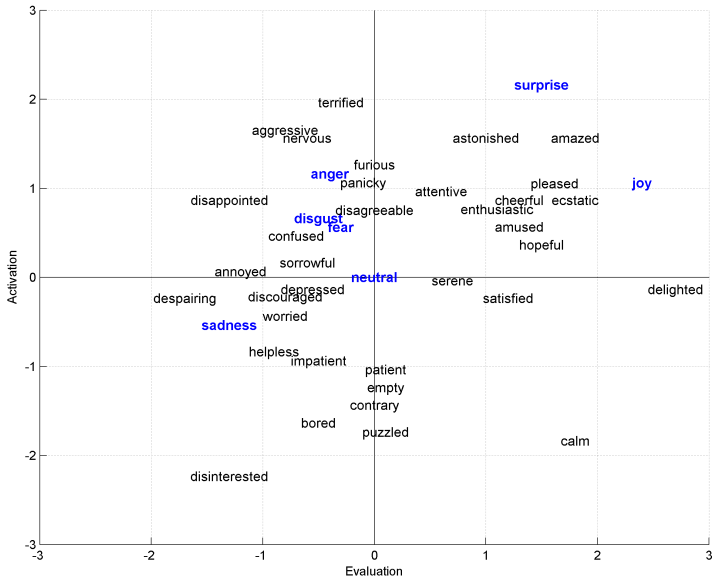

Figura 2.1: Espacio emocional de Whissell

La representación dimensional proporciona una manera de describir un amplio rango de estados emocionales y así como una medida de la intensidad de la emoción. Es capaz de representar variaciones continuas de las emociones no solamente en el espacio, sino a lo largo del tiempo [11]; esto resulta especialmente relevante cuando se estudia un periodo de tiempo en el que se expresa más de una emoción, siendo posible representar la evolución de las mismas. Precisamente, puesto que como se verá en el Capítulo 5 se trabajará con periodos de tiempo relativamente largos y emocionalmente complejos, será este el tipo de salidas emocionales que se manejen en este trabajo: continuas en el espacio y en el tiempo.

2.2. Fusión Multimodal

La fusión multimodal para el reconocimiento de emociones es un campo de estudio muy activo en el que constantemente aparecen nuevos trabajos [12]; sin embargo, debido a los problemas que presenta [13] existen múltiples aproximaciones al problema y todavía no resulta claro qué método resulta más ventajoso. Las tres principales estrategias seguidas en la actualidad son: fusión a nivel de características, fusión a nivel de decisión y fusión híbrida.

La fusión de características combina los datos procedentes de los distintos cana-

les previamente a la clasificación. Varios trabajos han mostrado buenos resultados empleando esta técnica [6, 14, 15] pero presenta el problema de complicarse conforme se incrementa el número de canales, especialmente cuando estos son de diferente naturaleza. Incluir una nueva modalidad representa un gran esfuerzo puesto que requiere re-entrenar el sistema completo de clasificación.

Para superar estas dificultades gran parte de los investigadores trabajan con fusión a nivel de decisión en la que la información procedente de cada uno de los canales es clasificada individualmente y posteriormente estos resultados son combinados empleando distintas técnicas (reglas expertas, mayoría, suma ponderada, etc). Múltiples estudios han demostrado las ventajas de la fusión a nivel de decisión gracias a los errores no correlacionados procedentes de diferentes clasificadores [16] y el hecho de que las dependencias temporales y entre características desaparecen.

Los métodos de fusión híbridos [17, 18] realizan ambos tipos de fusión; a nivel de característica y a nivel de decisión. Mantienen clasificadores separados para cada canal, pero dado que emplean información de todos los canales en cada clasificador presentan los mismos problemas de complejidad que los métodos a nivel de característica.

Kim [19] estudia las diferencias entre las distintas estrategias de fusión aplicadas a la fusión de dos modalidades; audio y señales fisiológicas, obteniendo mejores tasas de acierto mediante la fusión de características, si bien las diferencias resultan únicamente del 3 %.

Si bien existen en la actualidad diversos sistemas de fusión [20, 21, 22], principalmente bimodales, no hay por el momento modelos óptimos de fusión genéricos y que permitan mayor flexibilidad. En este sentido aquí se opta por un modelo de fusión a nivel de decisión [23] en busca de resolver el problema de la escalabilidad, simplificando el proceso de adición de nuevos módulos.

2.3. Métodos de Anotación y Evaluación

El proceso de evaluación de cualquier sistema de detección de emociones comienza por la anotación de los datos. La anotación consiste en definir el resultado que será considerado correcto para un conjunto de datos concreto. Puede ser realizada por uno o varios anotadores, dependiendo de la complejidad de los datos o del tipo de sistema a evaluar. Por otro lado, la evaluación de los resultados ofrecidos por el sistema de

detección consiste en determinar si el resultado se corresponde con la anotación previa o no.

Cuando se pretende evaluar un sistema discreto (basado en categorías) un sólo anotador puede ser suficiente, ya sea un anotador experto o el sujeto que expresa la emoción, puesto que sólo es preciso determinar la categoría a la que pertenece una emoción. No obstante es común emplear varios anotadores y seleccionar la categoría definitiva mediante mayoría de votos. Esto permite aportar mayor información como la tasa de acuerdo entre anotadores y la exactitud de la anotación.

Aunque también se han empleado sistemas de anotación por categorías para analizar sistemas continuos [24, 25] y datos que contienen varias emociones, aparece el problema de determinar el número de etiquetas suficiente para discretizar el espacio emocional. Adicionalmente, dado que este número óptimo depende de los datos estudiados añade una etapa más a la evaluación del sistema y dificulta la comparación de resultados entre datos de distinta naturaleza.

Así, la anotación continua es más adecuada cuando la muestra estudiada incluye más de una emoción o se encuentran emociones mezcladas y para ello se han desarrollado herramientas como Feeltrace [26]. El problema de este planteamiento es que no resulta práctico debido a la gran cantidad de tiempo necesario para la anotación y la falta de correlación entre anotadores [27].

Un método intermedio, que combina la rapidez de anotación con una evaluación adecuada para sistemas continuos, consiste en centrar la anotación a ciertos momentos clave de la misma forma que en [25], habitualmente el inicio (*onset*), máximo (*apex*) y final (*offset*) de cada emoción pero etiquetarlos en el espacio bidimensional del mismo modo que [26]. Estos momentos clave pueden ser definidos por el propio sujeto o por un anotador experto. Posteriormente, el resto de anotadores los etiquetarán en el espacio bidimensional, obteniendo un conjunto más reducido de puntos pero que siguen siendo representativos de la evolución de las emociones estudiadas. Este método se plantea en [28] para un módulo de detección facial pero es fácilmente escalable a un sistema multimodal y será el método empleado en este trabajo.

CAPÍTULO 3

Sistema de Fusión Multimodal

El sistema que se presenta en este capítulo se ha implementado en base al algoritmo de fusión planteado por Hupont [1]. Este algoritmo pretende ser flexible, permitiendo el empleo de múltiples módulos emocionales de distinta naturaleza, proporcionando un resultado coherente y cuya tasa de acierto sea superior a cada uno de los módulos por separado.

La ventaja de este método de fusión es que permite la fusión de módulos tan distintos como los que emplean señales fisiológicas con los que procesan texto introducido por el usuario y que hayan sido desarrollados de forma independiente, sin estar diseñados para trabajar junto con otros módulos. Esto es posible gracias a emplear la fusión a nivel de decisión, es decir, cada módulo procesa la información de los canales que emplea y proporciona una salida emocional; posteriormente las distintas salidas son fusionadas en el espacio continuo proporcionando un único resultado.

3.1. Descripción del Sistema

A continuación se describen las tres partes de las que consta el sistema de fusión multimodal que se pretende evaluar: mapeo emocional, fusión temporal y «cinética emocional».

3.1.1. Mapeo Emocional a un Espacio 2D Continuo

El primer paso en el algoritmo de fusión consiste en tener los resultados de cada uno de los módulos expresados en los mismos términos, en este caso, en el espacio Whissell [9]. En el caso de que un módulo proporcione como salida una única categoría discreta, su paso al espacio continuo es directo, al estar todas las categorías representadas como puntos en el espacio Whissell 2D (ver figura 2.1). Otra posibilidad, como por ejemplo la mostrada en la figura 3.1 y que corresponde con el módulo facial desarrollado por

Hupont et al [29], es que el módulo proporcione no una única categoría, sino una mezcla de ellas, con un peso asociado a cada una. En este caso, conociendo las posiciones 2D de las categorías y los pesos, se puede calcular la suma ponderada de todas las categorías, obteniendo un único punto en el espacio. Por último, en caso de que el módulo proporcione una salida bidimensional, como es el caso concreto del módulo de texto, desarrollado en este trabajo, no es necesario realizar ninguna conversión.

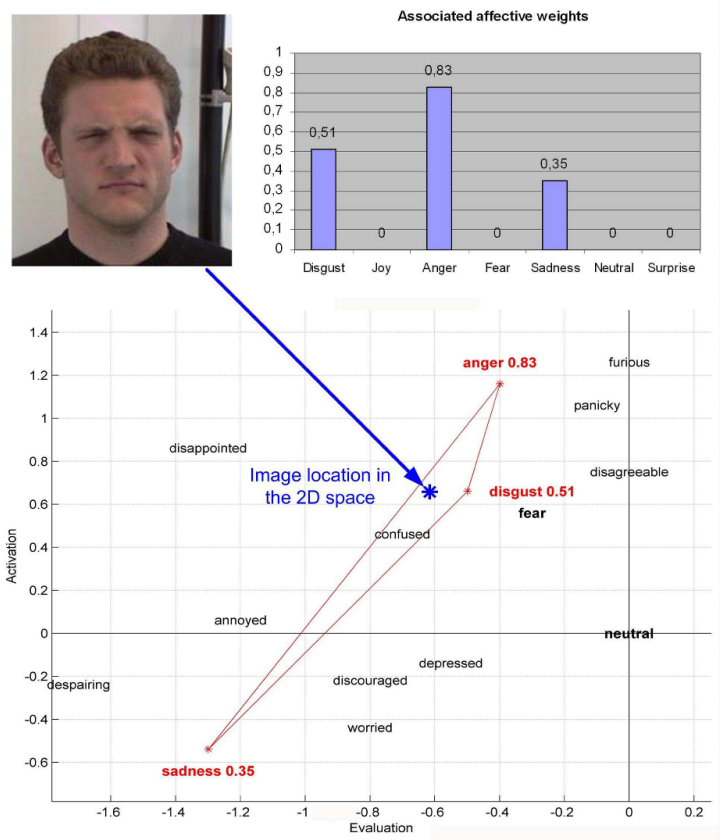

Figura 3.1: Mapeo de la salida del módulo facial al espacio Whissell

3.1.2. Fusión Temporal de Diferentes Módulos

Las emociones humanas se presentan inherentemente siguiendo un patrón temporal continuo [11]. Partiendo de esta base y gracias al empleo del espacio bidimensional evaluación-activación, el progreso emocional de un usuario puede ser representado como un punto (correspondiente a la posición de un estado afectivo concreto en el momento t) moviéndose por este espacio a lo largo del tiempo. El segundo paso de la metodología de fusión busca calcular el recorrido emocional mediante la fusión de

diferentes puntos $p_i(t_{0i})$ obtenidos por los distintos módulos a lo largo del tiempo.

La dificultad principal para conseguir la fusión multimodal se da por el hecho de que el momento exacto en el que se obtendrá una nueva entrada temporal (t_{0i}) para cada módulo puede ser conocido o no y el tiempo entre entradas puede ser muy diferente para cada módulo. Para solucionar este problema se propone la siguiente ecuación que calcula el estado afectivo global $p(t) = [x(t); y(t)]$ en un momento de tiempo arbitrario t:

$$p(t) = \frac{\sum_{i=1}^{N} \alpha_i(t) p_i(t_{0i})}{\sum_{i=1}^{N} \alpha_i(t)} \tag{3.1}$$

donde N es el número de modalidades fusionadas, t_{0i} el momento en el que se produce el último estímulo detectado por el módulo i y $\alpha_i(t)$ es la confianza entre 0 y 1 asignados a cada modalidad i en cada momento t.

De esta forma, la respuesta afectiva completa es la suma de la contribución de cada una de las modalidades $p_i(t_{0i})$ ponderada por el coeficiente $\alpha_i(t)$ a lo largo del tiempo. La definición de $\alpha_i(t)$ es especialmente importante dado que gobierna la respuesta temporal de la fusión. Tal y como propone Picard [30], la respuesta afectiva humana es análoga a los sistemas con respuesta aditiva que decaen con el tiempo, y en ausencia de una entrada, la respuesta regresa a un estado neutro. Siguiendo esta analogía, el valor de $\alpha_i(t)$ se define como:

$$\alpha_i(t) = \begin{cases} b_i c_i(t_{0i}) e^{-d_i(t-t_0)} & if \quad t > \varepsilon \\ 0 & elsewhere \end{cases} \tag{3.2}$$

donde:

- b_i es la confianza global que posee cada módulo i, habitualmente la tasa de acierto del mismo.
- $c_i(t_{0i})$ es la confianza temporal de cada módulo i para un instante concreto debido a factores externos (no debidos a la propia clasificación). Esto permite tener en cuenta, por ejemplo, un error temporal en el sensor en caso de señales fisiológicas

o una pérdida de las características faciales en caso de un módulo de expresiones faciales (debido a oclusiones, mala iluminación, etc).

- d_i es la tasa de decaimiento (en s^{-1}) e indica cómo de rápido desaparece un estímulo con el tiempo para el módulo i.

- ε es el umbral por debajo del cual se desprecia la contribución de un módulo. Esto es necesario dado que las funciones exponenciales tienden a cero en el infinito pero sin llegar a anularse completamente; cualquier $\alpha_i(t)$ por debajo de ε será despreciado.

Estableciendo los parámetros antes mencionados para cada módulo i y aplicando las ecuaciones (3.1) y (3.2) se puede calcular el recorrido emocional que caracteriza el progreso afectivo del usuario $p(t)$ a lo largo del tiempo. En otras palabras, el recorrido emocional se construye añadiendo progresivamente muestras a la trayectoria $p(t_k)$, siendo $t_k = k\Delta t$ (con k entero) y Δt el intervalo temporal entre muestras.

3.1.3. Cinética Emocional

En el cálculo de la trayectoria emocional aparecen dos grandes problemas:

1. Si la contribución de cada módulo fusionado es nula en un instante determinado, es decir, todos los $\alpha_i(t)$ son nulos, el denominador de (3.1) es cero y la trayectoria no puede ser calculada. Ejemplos de situaciones en las que la contribución de un módulo es nula son un fallo de un sensor en el caso de módulos fisiológicos, oclusiones en el caso de sistemas por visión o, sencillamente, cuando no se produce ninguna entrada durante un tiempo y $\alpha_i(t)$ decae por completo.

2. Si aparecen conflictos entre distintos módulos, o en caso de que un módulo produzca gran cantidad de ruido, pueden producirse grandes «saltos emocionales» en el espacio de Whissell.

Ambos problemas pueden ser resueltos aplicando un filtrado de Kalman a la trayectoria emocional resultante de la fusión. Por definición, el filtro de Kalman estima el estado de un sistema combinando una predicción inexacta con una observación inexacta de ese estado, de forma que el término con menor incertidumbre recibe el mayor peso

en cada instante t. De esta forma, por un lado, el filtro de Kalman alisa la trayectoria emocional previniendo los «saltos emocionales» que erróneamente pueden aparecer debido al ruido que presentan ciertos módulos. Por otro lado, se evitan situaciones en las que la suma de todos los $\alpha_i(t)$ resulte nula, empleando la predicción del Kalman en lugar del cálculo de la ecuación (3.1) para esas muestras.

De forma análoga a la mecánica clásica, en la «cinética emocional» se tiene un punto 2D moviéndose por el espacio de Whissell: la posición y velocidad son modelados como el estado del sistema X_k en el filtro de Kalman. Así $X_k = [x, y, v_x, v_y]_k^T$ representa la posición y velocidad XY en el instante t_k. Las sucesivas muestras $p(t_k)$ de la trayectoria emocional son consideradas observaciones del estado del sistema. Las dos ecuaciones implicadas en el filtrado de Kalman son la Ecuación de Proceso y la Ecuación de Medida.

3.1.3.1. Ecuación de Proceso

$$X_{k+1} = F_{k+1;k} X_k + w_k$$

$$\begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_{k+1} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_k + w_k$$

donde $F_{k+1;k}$ es la matriz de transferencia lleva el estado X_k del instante k al instante $k+1$ (de una muestra de la trayectoria emocional a la siguiente). Se asume que el ruido de proceso w_k es aditivo, blanco y gausiano con media cero. Tal y como se sugiere en la literatura [31] su matriz de covarianza Q_k se define como:

$$Q_k = \sigma^2 \begin{bmatrix} \frac{1}{3} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 1 & 0 \\ 0 & \frac{1}{2} & 0 & 1 \end{bmatrix}$$

donde σ^2 es la intensidad del ruido blanco gausiano que modela la aceleración del punto 2D, ya que ésta no ha sido tenida en cuenta en la ecuación de estado del sistema.

3.1.3.2. Ecuación de Medida

$$Y_k = H_k X_k + z_k$$

$$\begin{bmatrix} x_m \\ y_m \end{bmatrix}_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}_k \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix}_k + z_k$$

donde Y_k es la medida en el instante k y H_k es la matriz de medida. El ruido de medida z_k se asume aditivo, blanco, gausiano, con media nula y no correlacionado con el ruido del proceso w_k. Su matriz de covarianza R_k es la matriz identidad:

$$R_k = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

esto es asumiendo que las medidas en x e y contienen errores independientes con varianza $\lambda\ unidades^2$.

Una vez que las ecuaciones de proceso y medida están definidas es posible aplicar la estimación iterativa de Kalman a la trayectoria emocional, de forma que cada iteración corresponda a una nueva muestra.

3.2. Caso de Estudio

A fin de demostrar la eficacia del algoritmo de fusión planteado este será probado gracias a tres módulos: un módulo facial, un módulo de texto y un módulo de emoticonos. Los datos que analizarán dichos módulos se obtendrán mediante una herramienta de mensajería instantánea desarrollada para tal efecto y que será explicada en la Sección 4.1. Los usuarios que empleen esta herramienta podrán comunicarse mediante vídeo, texto escrito y emoticonos; así pues la herramienta es capaz de proporcionar una entrada de vídeo para el módulo facial y entradas de texto y emoticonos para sus respectivos módulos.

Los tres módulos presentan salidas muy diferentes: mientras que el facial proporcionará una respuesta para cada *frame* de vídeo y a intervalos de tiempo conocidos a

priori, el módulo de texto únicamente proporcionará una respuesta por cada frase con contenido emocional que el usuario escriba y el módulo de emoticonos proporcionará un menor número de respuestas pero con una gran confianza en las mismas. Fusionando tres canales tan diferentes se pretende demostrar la flexibilidad del algoritmo de fusión. Si bien los resultados proporcionados por el módulo de texto y emoticonos serán relativamente esporádicos, se pretende que estos ayuden a resolver situaciones en las que el módulo facial falla debido a una baja tasa de detección de ciertas emociones (como la distinción entre triste y neutro) o a fallos en la detección de los puntos característicos.

3.2.1. Módulo Facial

El módulo facial de detección de emociones consta de dos partes; detección de distancias faciales y clasificación de las mismas. Este módulo se basa en el algoritmo desarrollado por Hupont [32], con la diferencia de que en este trabajo fin de máster se realiza la detección de distancias faciales de forma automática en lugar de manual. Para ello se ha empleado la librería faceAPI [33] que proporciona los puntos faciales mostrados en la Figura 3.2. Esto permite el procesamiento de vídeos de larga duración, no limitando las entradas a imágenes estáticas o secuencias cortas.

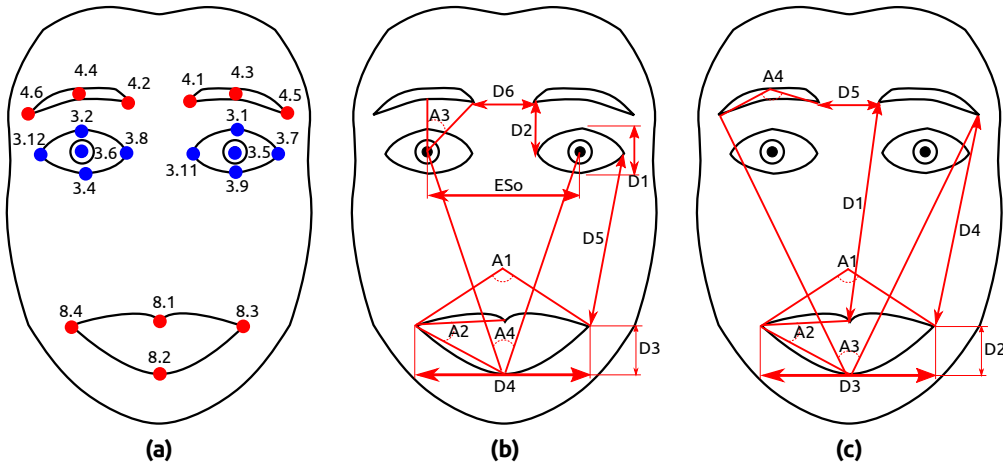

Figura 3.2: Puntos faciales representativos (a), distancias faciales calcuadas por el método manual (b) y distancias faciales calculadas por el método automático (c)

Mediante la posición de estos puntos, se calcula una serie de distancias y ángulos relativos entre puntos y se normalizan con respecto a la cara neutra. Sin embargo, a lo largo del desarrollo del Trabajo Fin de Máster, se observó que los puntos faciales correspondientes a los ojos proporcionados por faceAPI (marcados en azul en la imagen

3.2a) están erróneamente posicionados (faceAPI únicamente ubica estos puntos en base al resto), lo cual imposibilita el uso de los mismos durante la fases posteriores del algoritmo.

A fin de mitigar en lo posible los efectos negativos de la pérdida de la información de estas características faciales, se seleccionaron una serie de distancias faciales alternativas, aprovechando las características disponibles. A pesar de esto, como se comenta en la sección 5.5, la tasa de detección del módulo facial ha disminuido notablemente con respecto a la selección manual de características.

Una vez obtenidas las distancias relativas a la cara neutra estas se clasifican empleando la colección de algoritmos de *machine learning* Weka [34]. En concreto se una combinación de los clasificadores *RIPPER*, *Multilayer Perceptron*, *SVM*, *Naive Bayes* y *C4.5*. El resultado de la combinación de los clasificadores anteriores reduce el riesgo global de fallo, si bien cada clasificador puede tener un mejor funcionamiento para una clase individual [29].

A fin de entrenar los clasificadores mencionados en el párrafo anterior han sido empleadas dos bases de datos de expresiones faciales, la «MMI Facial Expression Database» [35] y la «MUG Facial Expression Database» [36]. Estas bases de datos permiten conseguir una muestra de imágenes anotadas lo suficientemente extensa y universal. De esta forma se facilita que el reconocimiento de expresiones faciales funcione bien en personas de ambos sexos y diferentes rasgos faciales. Un ejemplo de imágenes procedentes de estas bases de datos puede observarse en la Imagen 3.3.

Figura 3.3: Ejemplos de imágenes procedentes de las bases de datos

Con las distancias faciales obtenidas a partir de estas imágenes se entrena el modelo empleado para clasificar las distancias faciales de cada uno de los *frames* de un vídeo. El módulo facial proporciona como salida para cada *frame*, el peso asociado a cada una de las emociones de Ekman más el estado neutro (Imagen 3.4). Esta salida constituye la entrada proporcionada al algoritmo de fusión, junto con una estimación de la precisión de detección de los puntos faciales que será empleada como valor $c_i(t_{0i})$ en el algoritmo de fusión.

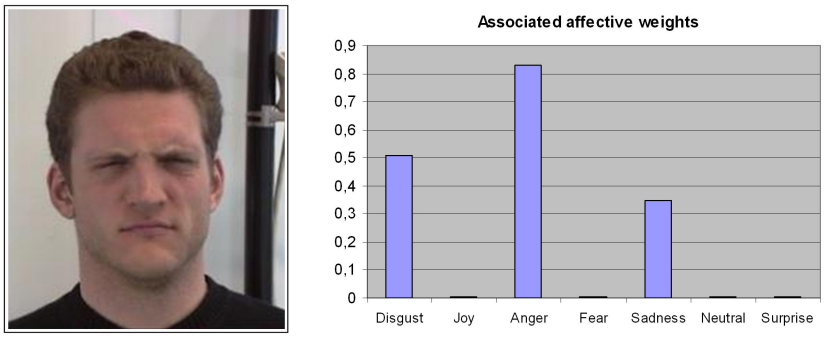

Figura 3.4: Ejemplo de salida emocional proporcionada por el módulo facial

3.2.2. Módulo de Texto

El módulo de texto analiza el texto tecleado por el usuario y sido desarrollado íntegramente en este trabajo fin de máster. El módulo proporciona como resultado un punto en el espacio bidimensional por cada frase con carga emocional. En función de la conversación entre los usuarios el número de este tipo de frases puede ser muy alto o relativamente bajo, pero puede ser de gran ayuda en la detección de emociones que mediante el módulo facial resultan complicadas de distinguir.

Figura 3.5: Esquema de funcionamiento del módulo de texto

El proceso de cálculo del estado emocional a partir de texto se observa en la figura 3.5 y consta de cinco fases:

1. La frase es descompuesta y convertida a su forma más simple; los sustantivos pasan a singular masculino, los verbos a infinitivo, etc. Para esta tarea se emplea la librería de análisis de texto *FreeLing* [37] capaz de trabajar en varios idiomas.

2. Se eliminan las llamadas *«stop-words»* que no aportan información a la frase a fin de simplificarla (por ejemplo «algo», «desde», «esa», «esto», etc).

3. Se detecta la presencia de negadores, que modificarán el valor de las palabras a las que precedan.

4. Se buscan las palabras restantes en el diccionario Whissell, de forma que se les asigne un valor emocional.

5. Se pondera el valor emocional de todas las palabras de la misma frase, proporcionando el resultado final.

El valor emocional de cada frase $f(k)$ queda definido según la ecuación 3.3, donde $p(i)$ es la posición en el espacio Whissell de cada palabra y $\eta(i)$ tomará el valor de -1 en caso de existir un negador o 1 en caso contrario y siendo N el número de palabras con valor emocional.

$$f(k) = \frac{\sum_{i=1}^{N} \eta(i)p(i)}{N} \tag{3.3}$$

3.2.3. Módulo de Emoticonos

Por último, el módulo de emoticonos permite que el usuario reporte directamente su estado emocional durante una conversación. La frecuencia que puede esperarse de este módulo es pequeña, pero de gran importancia por tres motivos; la tasa de acierto de este módulo es del 100 % (a no ser por error o engaño por parte del usuario), permite la entrada de estados emocionales en el algoritmo de fusión que el módulo facial no es capaz de detectar, como el aburrimiento, la calma o la satisfacción y el usuario suele emplearlo en momentos clave de mayor intensidad emocional.

Los emoticonos mostrados en la Figura 3.6 han sido diseñados para ser fácilmente reconocibles para el usuario, resultando representativos de cada una de las emociones y abarcando la mayor parte posible del espacio Whissell.

Figura 3.6: Emoticonos diseñados y su etiqueta emocional

CAPÍTULO 4

Herramientas Desarrolladas

Todas las herramientas presentadas en este capítulo han sido desarrolladas expresamente para permitir la evaluación del algoritmo de fusión y de los módulos presentados en el Capítulo 3, posibilitando la captura de datos y su posterior análisis y anotación.

Se han realizado empleando el entorno de desarrollo *QtCreator* y las librerías *Qt 4.7*, que posibilitan el empleo de las herramientas en la mayoría de sistemas operativos (*Windows*, *Linux*, *Mac OS*).

4.1. Herramienta de Mensajería Instantánea

La aplicación de mensajería instantánea se encarga de registrar el vídeo, texto y emoticonos de la conversación llevada a cabo por dos usuarios para proporcionar las entradas emocionales que se estudiarán mediante los módulos y el algoritmo de fusión explicados en el Capítulo 3.

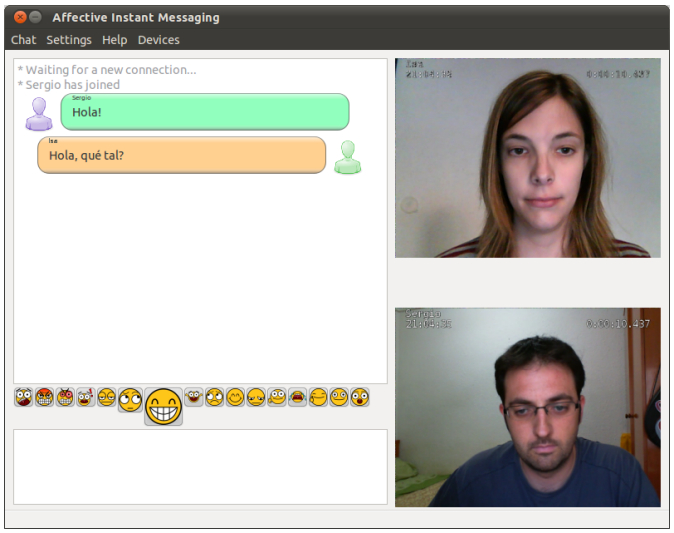

Figura 4.1: Vista del programa de chat

La librería *GStreamer* es empleada para capturar el vídeo mediante una *webcam*

convencional, posibilitando el uso de la aplicación en cualquier computador. La imagen es entonces mostrada en la pantalla local, comprimida, retransmitida hasta el equipo remoto y grabada en disco para su posterior uso. Simultáneamente se registra la escritura en el *chat* y la pulsación de emoticonos a fin de poder sincronizar los datos provenientes de los tres canales en el algoritmo de fusión.

4.2. Herramienta de Selección de Puntos Clave

Esta herramienta ha sido desarrollada para visualizar los vídeos obtenidos mediante la herramienta de mensajería instantánea y facilitar la selección de puntos clave de cara a la evaluación. Como puede observarse en la figura 4.2 en la zona izquierda se muestra la imagen de vídeo mientras que en la derecha se representan los puntos emocionales de ese instante, obtenidos mediante el algoritmo facial y el de fusión, en el espacio Whissell.

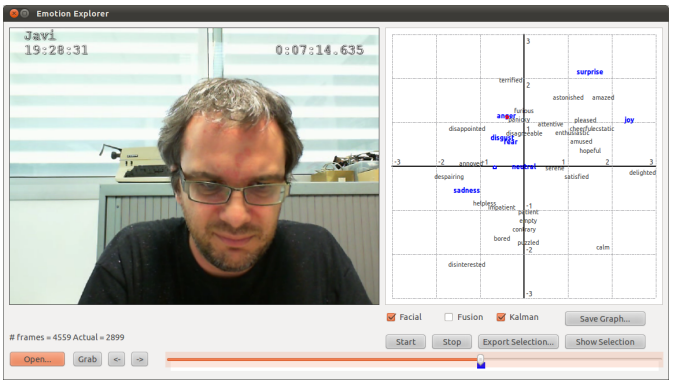

Figura 4.2: Vista del programa de selección de key-frames. Vista de vídeo capturado

Además permite analizar los resultados, visualizando las trayectorias emocionales tanto del módulo facial como del algoritmo de fusión y su evolución. Para ello es posible seleccionar un periodo de tiempo (figura 4.3) y visualizar tanto en su representación en el espacio Whissell como con respecto al tiempo de forma que se visualice de forma sencilla el paso por cada emoción.

De esta forma la herramienta permite a un usuario experto seleccionar los instantes de tiempo de mayor relevancia emocional para que posteriormente un mayor número de usuarios anoten los datos. La herramienta extrae las imágenes y los resultados seleccionados que serán entregados junto con la herramienta de anotación a los anotadores.

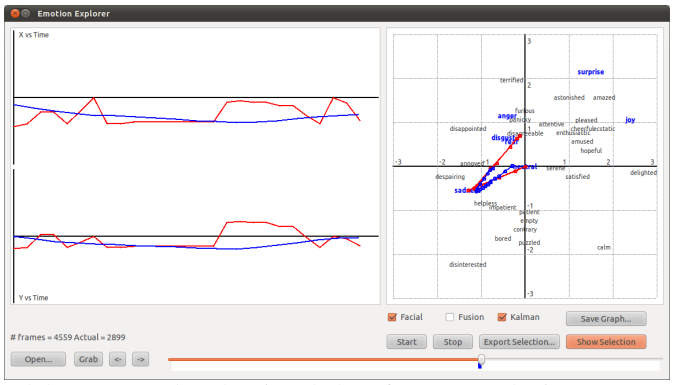

Figura 4.3: Vista del programa de selección de key-frames. Evolución temporal de la emoción.

4.3. Herramienta de Anotación de Puntos Clave

La herramienta de anotación de puntos clave ha sido diseñada para permitir la anotación por parte de usuarios inexpertos de los datos proporcionados por la herramienta de selección. Si bien existen programas similares empleados por otros investigadores como Feeltrace [26] o Anvil [38] estos no resultan lo suficientemente sencillos como para poder ser empleados por anotadores no expertos desde casa, lo que limita los posibles voluntarios. Anvil además no permite una anotación en el espacio bidimensional de forma precisa y rápida, ambos programas requieren de vídeos como datos de entrada con unos *codecs* muy específicos y no admiten imágenes estáticas.

La aplicación desarrollada muestra a los anotadores la imagen clave seleccionada del vídeo y las dos últimas frases ocurridas en el chat a fin de proporcionar el contexto necesario. En la zona de la derecha (como se muestra en la figura 4.4) se muestra el espacio Whissell junto con una serie de etiquetas que faciliten la decisión del punto emocional adecuado.

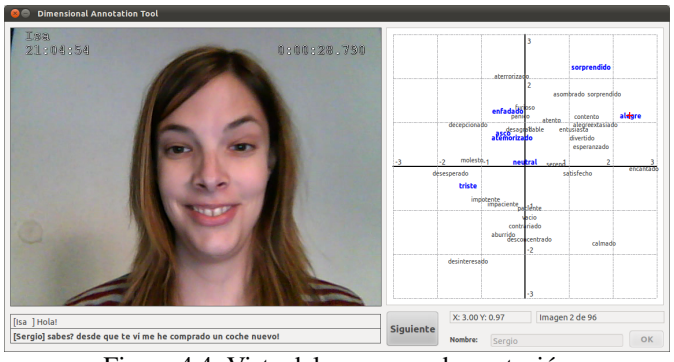

Figura 4.4: Vista del programa de anotación

Tras realizar la evaluación completa de los datos la aplicación permite introducir un comentario o sugerencia para ser tenida en cuenta en posteriores estudios y los registros son enviados automáticamente a un servidor FTP para su análisis. De esta forma la aplicación puede ser descargada por los anotadores desde un servidor web y ser empleada en sus propias casas.

CAPÍTULO 5

Evaluación del Sistema Multimodal

Una vez implementados los tres módulos emocionales, el algoritmo de fusión y todas las herramientas necesarias para el tratamiento de los datos se ha realizado la evaluación del sistema multimodal y su comparación con respecto a la respuesta proporcionada por el módulo facial en solitario. La metodología de evaluación se compone de cuatro pasos; recogida de datos, selección de puntos clave, anotación de puntos clave y análisis de los datos anotados.

5.1. Sesiones de Recogida de Datos

Para la recogida de datos se ha empleado la herramienta de mensajería instantánea mostrada en la Sección 4.1, realizándose sesiones por parejas de usuarios. Las condiciones de recogida de los datos son las que cualquier usuario pueda tener en su hogar, no se ha empleado ningún fondo ni iluminación especial durante los vídeos y el programa de mensajería instantánea puede ser instalado en cualquier equipo actual.

De esta forma se graban los vídeos, el texto y los emoticonos empleados durante la comunicación. De las múltiples sesiones realizadas para la evaluación final se emplearon 4 conversaciones, es decir 8 vídeos, de entre 2 y 3 minutos de duración y en los que se intercambiaron 84 frases. La selección de los vídeos corresponde a los que se aprecia un mayor número de emociones en un periodo de tiempo menor de forma que la anotación posterior resulte menos pesada para los anotadores y más eficiente.

5.2. Selección de Puntos Clave

Para aligerar la carga de los 16 anotadores que participaron en los test, se extrajeron 96 puntos clave de los 8 vídeos, que representan los puntos emocionalmente más representativos de los vídeos, es decir, los puntos de comienzo, máximo y descenso de

las emociones presentadas en ellos. En la imagen 5.1 puede observarse una trayectoria emocional y un ejemplo de puntos clave seleccionados sobre la misma. De estos instantes clave se extraen las imágenes que serán mostradas a los anotadores (empleando la herramienta descrita en la Sección 4.2) junto con las dos últimas frases intercambiadas antes de cada instante de tiempo de forma que los anotadores tengan mayor conocimiento del contexto.

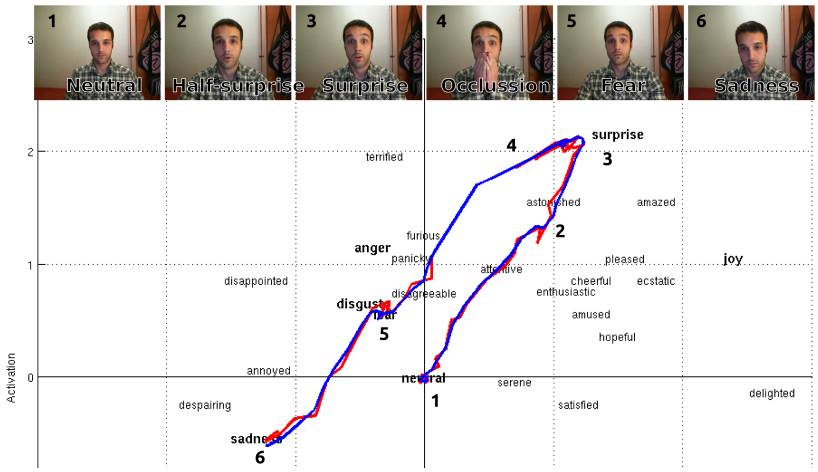

Figura 5.1: Trayectoria emocional y puntos clave seleccionados

5.3. Anotación de Puntos Clave

Los puntos clave fueron anotados en el espacio Whissell gracias a la participación de 18 voluntarios. Los anotadores disponen para cada instante a anotar de la imagen capturada por la cámara, las dos últimas frases escritas por los usuarios del *chat* así como los emoticonos que puedan estar incluidos dentro de estas frases (Imagen 5.2). De esta forma los anotadores disponen de la mayor información disponible para determinar el estado emocional del sujeto.

5.4. Análisis de los Datos Anotados

Los datos de las evaluaciones recogidas han sido empleados para definir una región en el espacio 2D donde el resultado emocional se considera correctamente localizado. El algoritmo que calcula la forma de la región está basado en el *Minimum Volume Ellipsoid* (MVE) empleando el método descrito por Kumar y Yildrim [39]. El algorimo

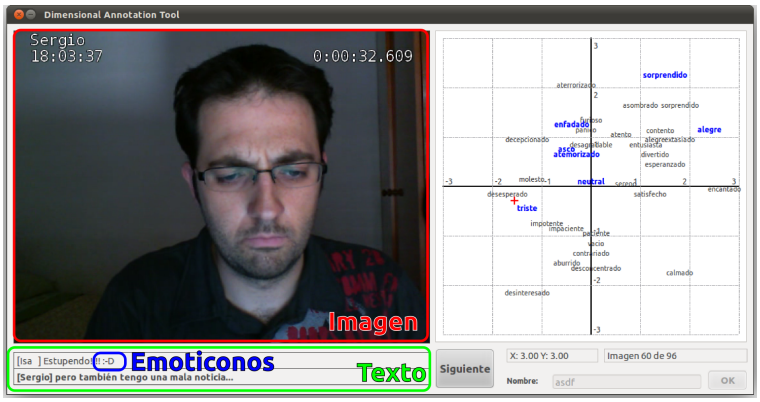

Figura 5.2: Herramienta de anotación e información mostrada al anotador

MVE busca el elipsoide de menor volumen capaz de incluir todo el conjunto de puntos, excluyendo espurios en caso de existir.

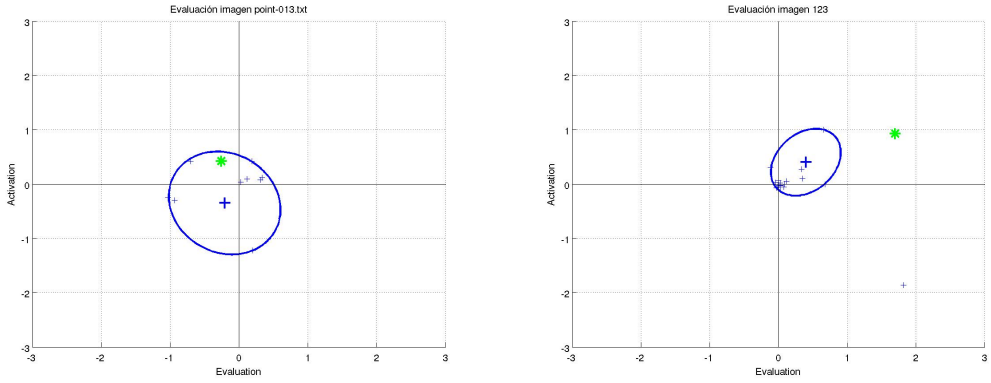

Figura 5.3: Método del Elipsoide de Mínimo Volumen sobre el espacio Whissell

En la imagen 5.3 se observa la representación del Método del Elipsoide del Mínimo Volumen sobre el espacio Whissell. Las marcas representadas mediante «+» son los puntos seleccionados por los anotadores, «**+**» marca del elipsoide del mismo y el símbolo «*» representa la salida emocional del módulo evaluado o del algoritmo de fusión según sea el caso.

Los MVEs calculados son empleados para evaluar los resultados a cuatro niveles diferentes:

1. Criterio de la elipse. Si el punto detectado por el sistema se encuentra dentro de la elipse es considerado acierto (Imagen 5.3a), en caso contrario fracaso (Imagen 5.3b).

2. Criterio del cuadrante. La salida del sistema es considerada correcta en caso de que se encuentre dentro del mismo cuadrante del espacio Whissell que el centro de la elipse calculada. En la Imagen 5.3a el criterio sería fracaso, mientras que en 5.3b sería acierto.

3. Criterio del eje de evaluación. La salida del sistema se considera correcta si está situada en el mismo semi-eje (positivo o negativo) del eje de evaluación que el centro de la elipse. Esta información es especialmente útil para determinar si una emoción es positiva o negativa. En ambos casos de la imagen 5.3 sería acierto.

4. Criterio del eje de activación. Es el criterio anterior aplicado al eje de activación. Esta información es relevante para conocer si un usuario está más dispuesto a realizar alguna acción bajo el estado emocional.En la Imagen 5.3a el criterio sería fracaso, mientras que en 5.3b sería acierto.

Uno de los problemas que surgen a la hora de comparar tasas de acierto entre métodos desarrollados por distintos autores resulta del hecho de que los datos empleados por unos pueden consistir en sujetos que expresan una emoción de forma actuada [35] (con mayor o menor grado de exageración), otros pueden consistir en emociones provocadas deliberadamente mediante un estímulo externo [40] y en un tercer grupo las emociones aparecen de forma natural [5], resultando mucho más difíciles de detectar.

Una posible solución a este problema es el uso del método de las elipses tal como se ha planteado aquí. La ventaja de emplear el método de las elipses consiste en que la distancia máxima permitida para que un punto sea considerado acierto varía en cada caso. Si una expresión resulta muy exagerada los puntos seleccionados por los anotadores se concentran en una región pequeña del espacio, resultando más estricta la aceptación del resultado del sistema como válido. Un ejemplo de esta situación puede observarse en la imagen 5.4.

Sin embargo, en el caso de expresiones no actuadas, o expresiones poco claras y que admiten más de una interpretación nos encontramos con que el área marcada por el conjunto de anotadores resulta mucho más extensa, siendo más flexible a la hora de aceptar un resultado del sistema de detección estudiado (imagen 5.5). De esta forma se obtiene un resultado dependiente del grado de consenso entre anotadores.

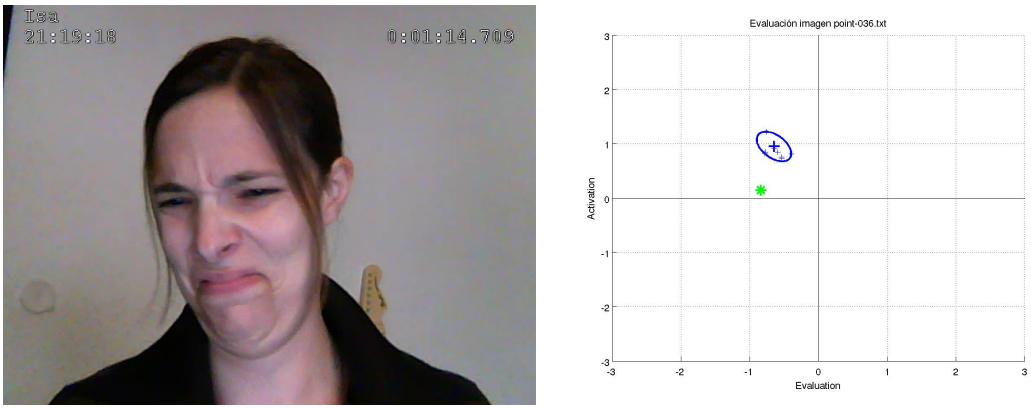

Figura 5.4: Expresión característica en la que todos los anotadores coinciden

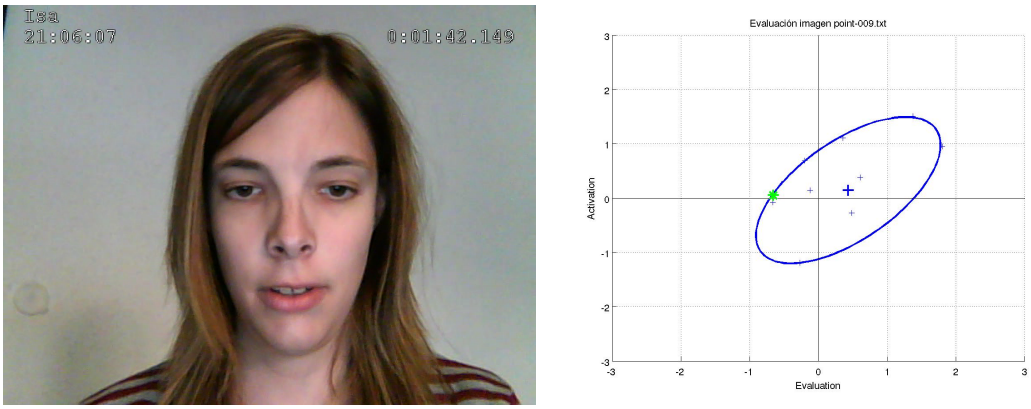

Figura 5.5: Expresión poco clara en la que hay gran dispersión entre anotadores

5.5. Resultados de la Evaluación

Los resultados obtenidos mediante las diferentes estrategias de evaluación se muestran en la Tabla 5.1. Se presentan los resultados de los algoritmos faciales, del de texto y del módulo de fusión multimodal. En este último caso, como puede observarse, la tasa de acierto es del 68,75 % en el caso más restrictivo (el criterio de la elipse) y se sitúa en 94,79 % cuando se considera el criterio de activación. Es importante destacar que, de acuerdo a Bassili [41], un observador entrenado puede clasificar correctamente una media del 87 % de las expresiones faciales.

Tal y como se ha comentado en el apartado anterior, debido a la gran diversidad de métodos de clasificación (discretos y continuos) y de evaluación empleados por los distintos autores, resulta complicado comparar objetivamente los resultados obtenidos por distintos autores. A modo de ejemplo, Busso et al [42] muestra un sistema bimodal (facial + audio), con emociones actuadas y reconocimiento facial basado en marcadores

	Elipse	Cuadrante	Evaluación	Activación
Facial (manual [1])	73,73 %	87,45 %	94,12 %	92,94 %
Facial (automático)	59,38 %	75,00 %	82,29 %	94,79 %
Texto	69,56 %	73,91 %	79,16 %	78,26 %
Fusion	68,75 %	77,08 %	83,33 %	94,79 %

Cuadro 5.1: Tasas de acierto de los algoritmos facial y de fusión

colocados sobre la piel del sujeto a estudiar. La clasificación se realiza únicamente en base a cuatro categorías (enfado, tristeza, alegría y neutra). La tasa de acierto del módulo de audio es del 70,9 %, la tasa del módulo de facial es del 85 % resultando una tasa de acierto combinada del 89,1 %. Por otro lado Kim [19] evalúa un sistema de fusión bimodal (audio + señales fisiológicas), los resultados son proporcionados en el espacio continuo, sin embargo su criterio de evaluación se limita a observar el cuadrante en el que se encuentra el resultado. Se obtienen tasas de acierto de un 51 % para el módulo fisiológico, un 54 % para el módulo de audio y un 55 % para el sistema combinado. No obstante sus resultados parecen muy dependientes del sujeto estudiado, llegando en un caso a obtener una tasa del 92 %.

5.6. Análisis de Resultados

A continuación se presenta el análisis de los resultados obtenidos gracias a las evaluaciones realizadas y a la observación de las trayectorias emocionales del módulo facial y el algoritmo de fusión multimodal.

5.6.1. Resultados del Nuevo Método Facial

Como puede observarse en la tabla 5.1, el módulo facial ha sufrido un descenso importante en la tasa de acierto, esto es debido a dos factores fundamentales:

1. La precisión de detección de características faciales obtenidas mediante faceAPI es mucho menor que las obtenidas mediante una selección manual. No obstante, una característica favorable de faceAPI es que proporciona una medida de la precisión, mediante un factor entre 0 y 1, que puede ser empleado como valor $c_i(t_{0i})$

en la ecuación 3.2. Gracias a ello, este efecto puede quedar mitigado siempre y cuando se disponga de entradas emocionales procedentes de otros módulos.

2. FaceAPI no detecta las características correspondientes al contorno de los ojos lo que provoca que ciertas distancias faciales no puedan ser calculadas. Esto resulta especialmente importante dado que muchas expresiones faciales, como la sorpresa o la ira, modifican el contorno de los ojos y esta información deja de estar disponible para los algoritmos de clasificación.

5.6.2. Resultados del Algoritmo de Fusión

Gracias al empleo de la multimodalidad se aprecian tres grandes mejoras: incremento en la tasa de acierto, ampliación del área emocional detectada y mejora de la trayectoria emocional.

5.6.2.1. Mejora de las Tasas de Acierto

En el caso estudiado se observa un incremento de la tasa de detección emocional del algoritmo de fusión con respecto al módulo facial automático, pasando de 59,38 % a 68,75 %. Este incremento resulta especialmente significativo teniendo en cuenta que las entradas emocionales producidas por los módulos de texto y emoticonos resultan mucho más dispersas que las del módulo facial. Conviene aclarar que, a pesar de que las tasas de acierto de los módulos de texto y emoticonos sean superiores a la tasa global del algoritmo facial, no sería posible construir la trayectoria emocional sin el empleo del módulo facial, puesto que se producirían grandes discontinuidades temporales debido a la falta de datos.

5.6.2.2. Ampliación del Área de Detección Emocional

Uno de los problemas intrínsecos del módulo facial que se solventa mediante la fusión multimodal es la limitación del área de detección del mismo. Dado que el resultado proporcionado por el módulo facial es una media ponderada de los puntos definidos por las emociones de Ekman, cualquier emoción que se encuentre fuera del polígono definido por ellas resulta inalcanzable.

En la imagen 5.6 observamos sombreada el área en la que trabaja el módulo facial

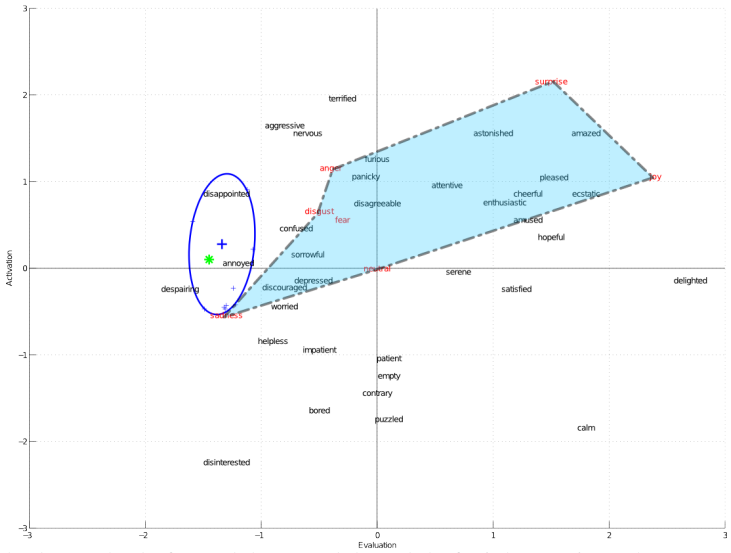

Figura 5.6: Ejemplo de resultado fuera del rango del módulo facial, que sin embargo es alcanzable gracias a la fusión.

y cómo emociones tales como decepción, irritación o desesperación quedan fuera de esta zona. Situaciones como la mostrada siempre resultarán en fallo para el módulo facial puesto que todos los anotadores coinciden en señalar una zona emocional que no intersecta con la alcanzable por el módulo facial. Sin embargo, gracias al empleo de módulos complementarios o con área más extensa es posible alcanzar estas zonas. Esto abre el camino al desarrollo de módulos específicos que mejoren la tasa de aciertos en zonas en las que otros módulos tienen una tasa de aciertos menor.

5.6.2.3. Mejora de la Trayectoria Emocional

El filtrado de Kalman aplicado en la fusión multimodal ha resultado muy eficiente a la hora de suavizar oscilaciones provocadas por el funcionamiento deficitario de la detección facial (figura 5.7), además de permitir obtener una salida emocional coherente incluso en periodos de oclusión facial y ausencia de otras entradas emocionales.

La figura 5.7 muestra la evolución temporal de la respuesta del algoritmo de fusión multimodal sin filtrar (Evaluación vs Tiempo en 5.7a y Activación vs Tiempo en 5.7b) en la que se observa una gran cantidad de ruido, con picos debidos a una detección deficiente de características faciales por parte del módulo facial y que no constituyen una variación emocional real. Estas variaciones desaparecen en la respuesta filtrada mediante Kalman (Evaluación vs Tiempo en 5.7c y Activación vs Tiempo en 5.7d).

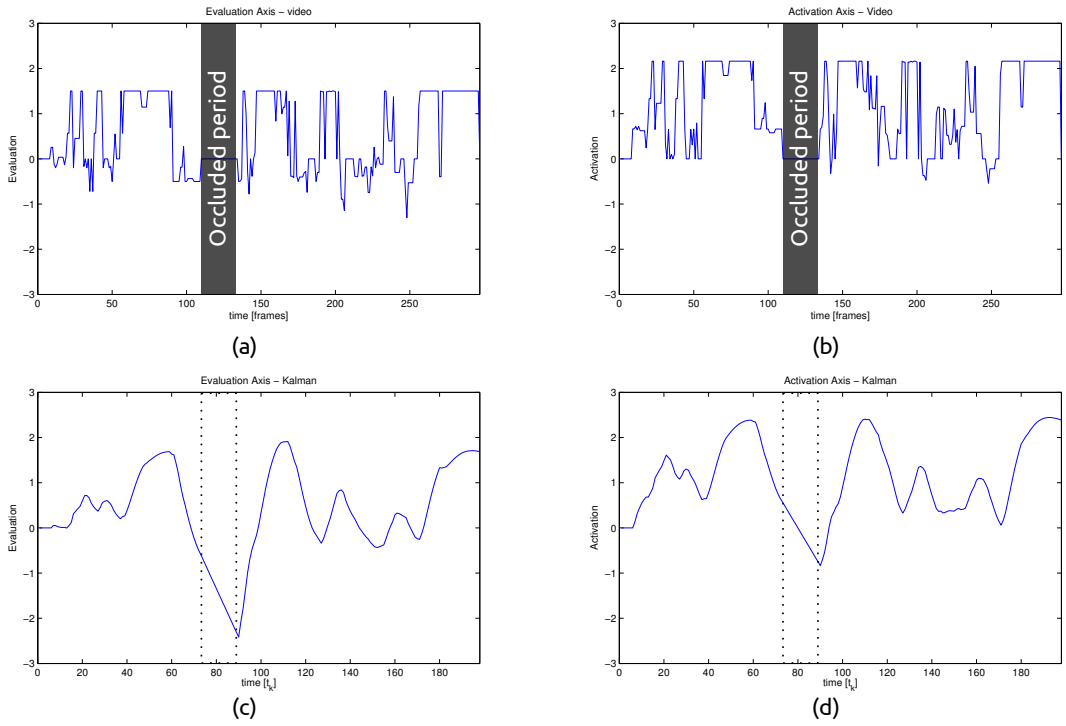

Figura 5.7: Efecto del filtrado de Kalman ante una oclusión

Asimismo se observa que, durante el periodo de oclusión facial dado que no existían otras entradas emocionales procedentes de otros módulos en ese instante, el algoritmo sin filtrar no es capaz de proporcionar una respuesta, mientras que gracias al filtrado de Kalman la respuesta emocional mantiene una trayectoria coherente hasta que se dispone de una nueva entrada emocional.

CAPÍTULO 6

Conclusiones y Trabajo Futuro

En este trabajo fin de máster se ha llevado a cabo la implementación y evaluación de un método de fusión multimodal planteado previamente. Ello ha implicado la realización de las siguientes tareas:

- Se ha mejorado el módulo de detección facial, permitiendo la detección de expresiones faciales de forma automática y posibilitando el procesado de largas secuencias de vídeo. Si bien se ha producido un descenso en la tasa de detección del módulo no ha constituido un problema para el desarrollo del trabajo y la tasa sigue siendo aceptable.
- Se ha desarrollado un módulo de detección de emociones en texto, basado en el diccionario de Whissell, capaz de proporcionar una salida emocional continua.
- Se ha desarrollado una aplicación, en forma de mensajería instantánea, que permite a dos usuario interactuar emocionalmente por medio de expresiones faciales (vídeo), texto y emoticonos y que recoge todos los datos necesarios para el posterior cálculo de las salidas emocionales de cada uno de los módulos estudiados.
- Se ha implementado el algoritmo de fusión y probado su funcionamiento gracias a los módulos faciales, de texto y emoticonos. Dicha implementación se ha realizado de forma flexible, posibilitando el empleo de futuros módulos sin necesidad de modificar el código de la aplicación.
- Se ha propuesto un método de evaluación para sistemas de detección de emociones continuas y se han desarrollado todas las herramientas necesarias para su aplicación. Dicho método ha sido empleado para estudiar el efecto de la fusión multimodal y el empleo de información procedente de distintos canales en la tasa de acierto.

- Se ha realizado un análisis de los resultados obtenidos, identificando las virtudes y los defectos tanto del algoritmo de fusión como del nuevo modelo facial implementado.

En cuanto a las posibles mejoras que pueden llevarse a cabo en futuros trabajos, ha quedado patente durante las pruebas realizadas la existencia de grandes deficiencias en la librería de detección facial faceAPI, resultando en una gran caída de la tasa de acierto individual del módulo facial con respecto a la selección manual de puntos y por consiguiente afectando negativamente al resultado global del algoritmo de fusión. Resultaría por tanto conveniente disponer de un módulo de detección de características faciales alternativo, si bien ninguno de los evaluados hasta la fecha ha proporcionado resultados satisfactorios.

Con respecto al módulo de texto, una posible mejora sería realizar una corrección ortográfica de la frase introducida por el usuario previamente a su reducción a la forma morfológica por FreeLing. Otra mejora pasaría por tener en cuenta los signos de puntuación, especialmente las exclamaciones como «potenciadores del eje de activación», por ejemplo, multiplicando por un factor el valor emocional de activación en una frase. En caso de ser necesario el módulo de texto podría ser empleado en lengua inglesa puesto que se dispone del diccionario Whissell en su lengua original.

En cuanto al sistema multimodal, en el futuro se espera poder evaluarlo empleando módulos de señales fisiológicas (ritmo cardiaco, conductividad de la piel, etc) y módulos de análisis de la voz. Incluso, sería posible combinar un *software* de dictado a fin de emplear el módulo de texto para detección de conversaciones habladas.

Bibliografía

[1] I. Hupont. *Affective Computing: emotional facial sensing and multimodal fusion*. PhD thesis, University of Zaragoza, 2010. 1, 8, 28

[2] P. Ekman, T. Dalgleish, and M. Power. *Handbook of Cognition and Emotion*. Wiley Online Library, 1999. 4

[3] Z. Hammal, A. Caplier, and M. Rombaut. Belief theory applied to facial expressions classification. In *In Int. Conf. on Advances in Pattern Recognition*, 2005. 4

[4] Gwen Littlewort, Marian Stewart Bartlett, Ian Fasel, Joshua Susskind, and Javier Movellan. Dynamics of facial expression extracted automatically from video. In *J. Image and Vision Computing*, pages 615–625, 2004. 4

[5] S. Abrilian, L. Devillers, S. Buisine, and J.C. Martin. Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces. In *Proceedings of HumanComputer Interaction International*, 2005. 4, 26

[6] A. Kapoor, W. Burleson, and R.W. Picard. Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736, 2007. 4, 6

[7] M. Yeasin, B. Bullot, and R. Sharma. Recognition of facial expressions and measurement of levels of interest from video. *Multimedia, IEEE Transactions on*, 8(3):500–508, 2006. 4

[8] P. Gupta and N. Rajput. Two-Stream emotion recognition for call center monitoring. In *INTERSPEECH'07*, pages 2241–2244, Antwerp, Belgium, August 2007. 4

[9] C.M. Whissell. *The Dictionary of Affect in Language, Emotion: Theory, Research and Experience*, volume 4. Academic, 1989. 4, 8

[10] R. Plutchik. *Emotion: a Psychoevolutionary Synthesis*. Harper & Row, 1980. 4

[11] S. Petridis, H. Gunes, S. Kaltwang, and M. Pantic. Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. In *Proceedings*

of the International Conference on Multimodal Interfaces, pages 23–30, 2009. 5, 9

[12] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, January 2010. 5

[13] H. Gunes, M. Piccardi, and M. Pantic. From the lab to the real world: Affect recognition using multiple cues and modalities. *Affective Computing: Focus on Emotion Expression, Synthesis, and Recognition*, pages 185–218, 2008. 5

[14] C. Shan, S. Gong, and P.W. McOwan. Beyond facial expressions: Learning human emotion from body gestures. In *Proceedings of the British Machine Vision Conference*, 2007. 6

[15] T. Pun, T.I. Alecu, G. Chanel, J. Kronegg, and S. Voloshynovskiy. Brain-computer interaction research at the computer vision and multimedia laboratory, University of Geneva. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):210–213, 2006. 6

[16] L.I. Kuncheva. *Combining pattern classifiers: methods and algorithms*. Wiley-Interscience, 2004. 6

[17] M. Wöllmer, M. Al-Hames, F. Eyben, B. Schuller, and G. Rigoll. A multidimensional dynamic time warping algorithm for efficient multimodal fusion of asynchronous data streams. *Neurocomputing*, 73(1-3):366–380, 2009. 6

[18] M. Mansoorizadeh and N.M. Charkari. Hybrid feature and decision level fusion of face and speech information for bimodal emotion recognition. In *2009 14th International CSI Computer Conference*, pages 652–657, Tehran, Iran, October 2009. 6

[19] J. Kim. Bimodal emotion recognition using speech and physiological changes. *Robust Speech Recognition and Understanding*, pages 265–280, 2007. 6, 28

[20] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, and S. Levinson. Audio-visual affect recognition. *IEEE Transactions on Multimedia*, 9(2):424–428, 2007. 6

[21] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4):1334–1345, 2007. 6

[22] P. Pal, A.N. Iyer, and R.E. Yantorno. Emotion detection from infant facial expressions and cries. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 721–724, 2006. 6

[23] I. Hupont, S. Ballano, S. Baldassarri, and E. Cerezo. Scalable multimodal fusion for continuous affect sensing. In *2011 IEEE Workshop on Affective Computational Intelligence (WACI)*, pages 1–8, Paris, France, April 2011. 6

[24] R. Craggs and M. McGee Wood. A categorical annotation scheme for emotion in the linguistic content of dialogue. In Elisabeth André, Laila Dybkjær, Wolfgang Minker, and Paul Heisterkamp, editors, *Affective Dialogue Systems*, volume 3068, pages 89–100. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. 7

[25] L. Devillers, L. Vidrascu, and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005. Emotion and Brain. 7

[26] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schr .ºder. Feeltrace: An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 19–24. Citeseer, 2000. 7, 21

[27] E. Douglas-Cowie, L Devillers, J.C. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox. Multimodal databases of everyday emotion: Facing up to complexity. In *Ninth European Conference on Speech Communication and Technology*, pages 813–816, Lisbon, Portugal, September 2005. 7

[28] S. Ballano, I. Hupont, E. Cerezo, and S. Baldassarri. Continuous facial affect recognition from videos. In *Proceedings of the «Congreso de Interacción Persona-Ordenador»*, 2011. 7

[29] I. Hupont, E. Cerezo, and S. Baldassarri. Sensing facial emotions in a continuous 2d affective space. In *SMC*, pages 2045–2051, 2010. 9, 15

[30] R.W. Picard. *Affective Computing*. The MIT Press, 1997. 10

[31] D.R. Morrell and W.C. Stirling. An extended set-valued kalman filter. In *Proceedings of ISIPTA*, pages 396–407, 2003. 12

[32] I. Hupont, E. Cerezo, and S. Baldassarri. Facial emotional classifier for natural interaction. *ELCVIA*, 7(4), 2008. 14

[33] SeeingMachines. Faceapi technical specifications brochure, 2011. `http://www.seeingmachines.com/pdfs/brochures/faceAPI-techspecs.pdf`. 14

[34] G. Holmes, A. Donkin, and I. H Witten. WEKA: a machine learning workbench. In *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems,1994*, pages 357–361. IEEE, December 1994. 15

[35] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int'l Conf. Multimedia and Expo*, pages 317–321, 2005. 15, 26

[36] N. Aifanti, C. Papachristou, and A. Delopoulos. The MUG facial expression database. In *2010 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4. IEEE, April 2010. 15

[37] TALP Research Center. Universitat Politècnica de Catalunya. Freeling: An open source suite of language analyzers, 2011. `http://nlp.lsi.upc.edu/freeling/`. 17

[38] M. Kipp et al. Spatiotemporal coding in anvil. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco*, 2008. 21

[39] P. Kumar and E. A Yildirim. Minimum-Volume enclosing ellipsoids and core sets. *Journal of Optimization Theory and applications*, 126:1–21, 2005. 24

[40] J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner. Integrating information from speech and physiological signals to achieve emotional sensitivity. In *in Proc. 9th Eur. Conf. Speech Communication and Technology*, pages 809–812, 2005. 26

[41] J.N. Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology*, 37:2049–2058, 1979. 27

[42] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *in Sixth International Conference on Multimodal Interfaces ICMI 2004*, pages 205—211, 2004. 27