UNIVERSIDAD NACIONAL DEL LITORAL

DOCTORADO EN INGENIERÍA

Reconocimiento de Estados Afectivos a partir de Señales Biomédicas

Leandro Ariel Bugnon

FICH
FACULTAD DE INGENIERÍA
Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO
PARA LA INDUSTRIA QUÍMICA

Tesis de Doctorado **2017**

Doctorado en Ingeniería
Mención en Inteligencia Computacional, Señales y Sistemas

Título de la obra:

Reconocimiento de Estados Afectivos a partir de Señales Biomédicas

Autor: Leandro Ariel Bugnon

Lugar: Santa Fe, Argentina

Palabras Claves:

reconocimiento de emociones, interfaces hombre-máquina, procesamiento de señales biomédicas, métodos auto-organizativos, reconocimiento en tiempo real, aprendizaje maquinal.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

RECONOCIMIENTO DE ESTADOS AFECTIVOS A PARTIR DE SEÑALES BIOMÉDICAS

Leandro Ariel Bugnon

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERÍA
Mención en Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2017

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje "El Pozo", S3000, Santa Fe, Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

RECONOCIMIENTO DE ESTADOS AFECTIVOS A PARTIR DE SEÑALES BIOMÉDICAS

Leandro Ariel Bugnon

Lugar de Trabajo:
sinc(i)

Instituto de Señales, Sistemas e Inteligencia Computacional
Facultad de Ingeniería y Ciencias Hídricas
Universidad Nacional del Litoral

Director:
Dr. Diego H. Milone — sinc(i)-CONICET-UNL

Co-director:
Dr. Rafael A. Calvo — The University of Sydney

Jurado Evaluador:
Dra. Silvia Schiaffino — ISISTAN-CONICET-UNICEN
Dr. E. Marcelo Albornoz — sinc(i)-CONICET-UNL
Dr. José Biurrun Manresa — CITER-CONICET-UNER
Dr. Diego Fernandez Slezak — ICC-CONICET

certificacion1

certificacion2

DECLARACIÓN LEGAL DEL AUTOR

Esta Tesis ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería ante la Universidad Nacional del Litoral y ha sido depositada en la Biblioteca de la Facultad de Ingeniería y Ciencias Hídricas para que esté a disposición de sus lectores bajo las condiciones estipuladas por el reglamento de la mencionada Biblioteca.

Citaciones breves de esta Tesis son permitidas sin la necesidad de un permiso especial, en la suposición de que la fuente sea correctamente citada. Solicitudes de permiso para la citación extendida o para la reproducción parcial o total de ese manuscrito serán concebidos por el portador legal del derecho de propiedad intelectual de la obra, por medio escrito.

AGRADECIMIENTOS

Esta tesis no sería posible sin una suerte de circunstancias y personas, que directa o indirectamente han participado de este camino. En primer lugar, agradecer a mis padres, María Luz y Marcelo, que me enseñaron a valorar lo que se tiene y me empujaron a trabajar por una vida mejor. A Dafne, mi hermana, con su punto de vista que me recuerda siempre las cosas importantes. A Leda, mi madrina más por hechos que por formalidad, y a mi abuela, tíos y primos que estuvieron en alguna charla sobre los desafíos actuales y el futuro.

Agradecer de corazón a Luz, que estuvo a mi lado en todo este camino y fue la que más sufrió las horas de no estar. Ella sabe lo que la quiero. También agradecer a su familia, que me han dejado enseñanzas muy valiosas para la vida.

Agradecer a mis amigos, los de siempre y a los nuevos. En particular a Seba y José, con los que hicimos todo el camino del doctorado a la par, pintas mediante. Al resto de los compañeros el *sinc(i)*; puedo decir que tengo un recuerdo o un mensaje de cada uno, y eso es realmente increíble para un lugar de trabajo. También agradecer especialmente a los que han trabajado conmigo en varios proyectos, además de la tesis, y con los cuales espero seguir trabajando.

Un gran gracias a mis directores, Diego Milone y Rafael Calvo, por la paciencia, el trabajo y la predisposición para discutir ideas y revisar manuscritos que fueron y vinieron varias veces. En particular a Diego, que además de ser un director increíble siempre está dispuesto para una charla de lo que sea.

Quiero agradecer a todos los que hicieron posible llegar hasta acá. Recuerdo con algo de nostalgia el paso por la FIUNER, a Javier que me despertó interés por la investigación, algo que veía lejano y sólo para unos pocos.

Es importante agradecer a todos los que hacen posible que este sistema de formación académica pueda funcionar. A las instituciones, a la Facultad de Ingeniería y Ciencias Hídricas, de la Universidad Nacional del Litoral, y a CONICET, por el apoyo en becas, subsidios para compra de equipamiento, lugar de trabajo y servicios.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

RESUMEN

Las emociones constituyen una parte fundamental de los individuos, influyendo en su comunicación diaria, la toma de decisiones y el foco de atención. La incorporación de las emociones en la tecnología ha avanzado en los últimos años, desde estudios exploratorios en la respuesta a los estímulos, a aplicaciones comerciales en interfaces hombre-máquina. Una de las fuentes para identificar estados emocionales es la respuesta fisiológica, registrada mediante señales biomédicas. El uso de estas señales permitiría el desarrollo de dispositivos poco invasivos, como por ejemplo una pulsera, que puedan registrar señales continuamente, en diferentes condiciones, y manteniendo la privacidad de los usuarios. Existen numerosos enfoques para el reconocimiento de afectos, con diferentes señales, técnicas de procesamiento de la señal y métodos de aprendizaje automático. Entre ellos, la combinación de múltiples señales se utilizó ampliamente para mejorar las tasas de reconocimiento, pero resulta inviable en la práctica por su invasividad. Los desafíos actuales requieren clasificadores que puedan funcionar en tiempo real, en aplicaciones interactivas, y con mayor comodidad para el usuario.

En esta tesis doctoral se aborda el desafío del reconocimiento de estados afectivos en varios aspectos. Se revisan las propiedades de cada señal fisiológica en términos de su practicidad y potencial. Se propone un método para adaptar un clasificador a nuevos usuarios, estimando parámetros fisiológicos basales. Luego se presentan dos métodos originales para mejorar las tasas de reconocimiento. El primero es un método supervisado basado en mapas auto-organizativos (sSOM). Este método permite representar los espacios de características fisiológicas y modelos emocionales, para analizar las relaciones en los datos. El otro esta basado en máquinas de aprendizaje extremo (ELM), una novedosa familia de redes neuronales artificiales que tiene gran poder de generalización y puede entrenarse con pocos datos. Los métodos fueron evaluados y comparados con los del estado del arte, en corpus realistas y de acceso libre.

Los resultados obtenidos muestran avances en relación al estado del arte para el problema. El método de adaptación permite, a partir de pocos segundos, mejorar las tasas de reconocimiento en tiempo real, aproximando los resultados del reconocimiento que se podría hacer con posterioridad, sobre los registros completos. Utilizando una única señal de actividad cardiovascular, en particular la variabilidad del ritmo cardíaco (HRV), se lograron avances prometedores, con diferencias significativas en relación a los resultados obtenidos por los métodos del estado del arte. Las ELM obtuvieron excelentes resultados y con bajo costo computacional, por lo que serían útiles para aplicaciones móviles. El sSOM logra resultados similares, con la ventaja de proveer a la vez una herramienta para representar y analizar los espacios complejos de la fisiología y las emociones, en una forma compacta.

ABSTRACT

Emotion is a fundamental part of individuals, it influences daily communication, decision making process and attention. The use of this information in technology has advanced in recent years, from exploratory studies in emotional response to stimuli, to commercial applications of human-computer interfaces. One of the sources to detect emotions is the physiological responses, registered with biomedical signals. These signals have the potential for the development of minimally invasive devices, such as a wristband, that can record signals continuously, in different conditions, and maintaining the privacy of users. There are numerous approaches to the recognition of affects, with different signals, signal processing techniques and machine learning methods. Among them, the combination of multiple signals was widely used to improve recognition rates, but it is unfeasible in practice. The current challenges require classifiers that can work in real time, for interactive applications, and with greater comfort for the user.

In this doctoral thesis the challenge of affect recognition is addressed in different aspects. The properties of each physiological signal are reviewed in terms of the potential and invasiveness. A method is proposed to adapt a classifier to new users and estimate baseline physiological parameters. Then two original methods are presented to improve recognition rates. The first is a supervised method based on self-organizing maps (sSOM). This method allows to represent the spaces of physiological features and emotional models, for further analysis of the relationships in the data. The other is based on extreme learning machines (ELM), a novel family of artificial neural networks that use random projections of features. The methods were evaluated and compared with those of the state-of-the-art, in realistic and freely accessible corpus.

Results show significant progress in relation to the task state-of-the-art methods. The adaptation method makes possible to improve the online recognition rates by using a few seconds of each session, achieving performance rates closer to offline recognition rates. Using the cardiovascular activity as unique signal, specially the heart rate variability (HRV), significant improvements were obtained in emotion recognition. The ELM achieved excelent results, with a low computational cost and good generalization capacity. The sSOM achieves similar results, with the advantage of providing a tool to represent and analyze complex spaces of physiology and emotions in a compact way.

1 Introducción

"Medir todo lo que se pueda medir, y hacer medible lo que no lo es aún"; éste ha sido el principio de Galileo

Wilhelm Dilthey, 1894

El estudio de los estados afectivos en el contexto de la informática ha surgido recientemente como un área de gran interés en la comunidad científica. Es así como actualmente se reconoce una nueva disciplina denominada computación afectiva, que desarrolla herramientas computacionales para estudiar, interpretar y replicar respuestas afectivas [1]. El *afecto* es un término genérico que incluye a las emociones pero también a otros estados mentales, actitudes, estados anímicos, sentimientos o temperamentos.

La información relacionada a los afectos resulta valiosa para toda interfaz hombre-máquina. Un sistema que es capaz de interpretar qué siente una persona puede adaptar su comportamiento o transmitir esta información a otra persona, en una amplia gama de aplicaciones prácticas. Estos sistemas han sido objeto de estudio en áreas como la terapia en desórdenes del espectro autista [2], manejo de estrés [3] y meditación [4]. Se han propuesto tecnologías para el etiquetado automático de contenido multimedia, basándose en la respuesta que incitan en el público [5]. Más recientemente, aparecen aplicaciones en donde se requiere identificar estados emocionales en tiempo real, por ejemplo el control automático de la dificultad en video-juegos [6, 7], la identificación de estados de un alumno durante el aprendizaje [8, 9] o la comunicación de emociones en teleconferencia [10]. El reconocimiento de los afectos también es importante para lograr interacciones más realistas con robots [11] y agentes virtuales [12]. Sin dudas, los afectos juegan un rol importante en la toma de decisiones, comunicación, salud y condicionan las capacidades de desarrollo y trabajo de las personas [13].

En esta tesis, la motivación principal para la investigación en el reconocimiento de los afectos es poder mejorar las capacidades actuales para *hacer medible de forma automática lo que no lo es aún*, partiendo de variables que puedan registrarse fácilmente e incorporarse a la tecnología cotidiana. Los estados afectivos se pueden manifestar de diferentes maneras y pueden inferirse a distintos niveles. En particular la tesis se centra en el estudio de las emociones, que son respuestas sicofisiológicas de corta duración (segundos o pocos minutos), y para las cuales existen numerosos trabajos de reconocimiento basados en la voz [14], expresiones faciales [15], posturas corporales [16] y también a partir de las señales fisiológicas [9]. Si bien todas son en definitiva respuestas fisiológicas, las señales fisiológicas, capturadas usualmente por señales biomédicas, son medidas directas del sistema

nervioso autónomo (SNA) y del sistema nervioso central (SNC), y tienen diversas ventajas. Pueden registrarse continuamente y bajo diferentes condiciones ambientales, lo que actualmente es de gran interés para las aplicaciones prácticas. Las expresiones faciales, la voz y el cuerpo pueden ser espontáneas, pero también están relacionadas con lo que la persona deja ver en un contexto social. Las medidas fisiológicas pueden ser manipuladas (por ejemplo controlando el ritmo respiratorio), pero son más difíciles de enmascarar voluntariamente que las otras fuentes mencionadas. En ciertas patologías, como el caso del trastorno autista, existen comportamientos que afectan al lenguaje y la comunicación, presentando obstáculos para la interpretación regular del lenguaje oral y corporal, siendo las señales fisiológicas una alternativa viable para conocer que sienten las personas que las padecen. Un punto no menor es que el usuario puede tener mayor privacidad en comparación al uso de cámaras o micrófonos, por lo que puede eventualmente sentirse más cómodo [17]. Por último, la transición de la tecnología actual desde las computadoras portátiles a las capacidades de los teléfonos inteligentes y el uso de tecnología corporal hace que estas nuevas fuentes de información, que hace pocos años podrían resultar impracticables, sean una nueva tendencia en el desarrollo tecnológico [18].

1.1 Emociones y computación fisiológica

La computación afectiva basada sólo en respuestas fisiológicas es más reciente si se la compara, por ejemplo, con el uso de la voz o expresiones faciales. Aún así, existen numerosos aportes tanto en el estudio de las relaciones entre la fisiología y los estados mentales (área denominada psicofisiología) como en su implementación en la tecnología y sus interacciones. Diversos estudios han demostrado que existen cambios fisiológicos relacionados con las emociones, capturados con diferentes señales biomédicas [19]. Se puede diferenciar por un lado las señales provenientes del sistema nervioso central, como las que se podrían registrar con un electroencefalograma (EEG) [20, 21]. La información sobre el estado mental de una persona se encuentra enmascarada en los registros de la actividad cerebral, y se han obtenido interesantes resultados analizando señales de EEG [23]. Por otro lado, las señales fisiológicas *periféricas* registran la influencia de las redes simpáticas y parasimpáticas del SNA en diferentes sistemas, como la circulación sanguínea, la respiración y la regulación de las glándulas de la piel [22]. Entre éstas se pueden citar diferentes estudios que utilizaron patrones de respiración [24], electrocardiograma (ECG) [25], fotopletismografía (PPG, del inglés *photoplethysmography*) [26], actividad electrodérmica (EDA, del inglés *electrotermal activity*) [27], temperatura de piel (TP) [28] y respuesta pupilar [29]. Además, la actividad del sistema nervioso somático está presente en movimientos y reflejos musculares involuntarios [19], o activaciones musculares características como las que suceden en situaciones de estrés [17], medidas mediante electromiograma (EMG). Esta última se ha utilizado también para registrar la activación de

músculos faciales [30], aunque esto estaría relacionado en mayor medida con en análisis de la expresión facial y no la respuesta fisiológica como tal.

Se han propuesto diferentes combinaciones multimodales para mejorar la capacidad discriminativa de los reconocedores [1]. En computación afectiva el concepto de *invasividad* es más estricto del que se utiliza normalmente, que proviene de la medicina. Algunos procedimientos, como medir la actividad cerebral usando 32 electrodos superficiales en el cuero cabelludo, son definidos como *no invasivos* porque se colocan por fuera del cuerpo, pero pueden ser *moderadamente invasivos* en experimentos con emociones en el laboratorio, considerando que el participante debería sentirse cómodo durante el registro. Esta misma técnica probablemente sería *altamente invasiva* en una aplicación de campo, cuando el usuario tiene que conectar todos los electrodos para realizar una actividad en su ambiente cotidiano. El hecho de tener que instrumentarse con múltiples sensores impone una limitación práctica no menor. Esto hace que sea muy difícil conseguir condiciones de prueba en las que el comportamiento del usuario no se vea afectado por todos los sensores (y cables) que tiene conectados. Desde esta perspectiva es que en el presente trabajo se busca minimizar la cantidad de señales necesarias para el reconocedor, poniendo toda la complejidad del sistema en los algoritmos de extracción de características y clasificación.

Los resultados del reconocimiento de emociones a partir de señales fisiológicas en trabajos previos indican que es una tarea factible pero aún deben mejorarse. Si se comparan resultados recientes [110], puede verse que la tasa de reconocimiento a partir del análisis audiovisual de los sujetos es cercana al doble que la de las señales fisiológicas utilizadas en conjunto. Más aún, es deseable reducir al mínimo la cantidad de sensores requeridos y lograr tasas de reconocimiento aceptables para que, junto con el avance de las tecnologías en sensores corporales, el registro de señales fisiológicas de baja invasividad permita desarrollar aplicaciones funcionales fuera del laboratorio [31]. También es necesario desarrollar métodos que permitan analizar las relaciones entre las variables fisiológicas y los modelos afectivos. Un esquema típico de reconocimiento se encuentra ilustrado en la Figura 1.1. La medición, lograda de la forma menos invasiva posible, podrá obtener información de un usuario en todo momento y reportar su estado afectivo en una representación conveniente. Los bloques de registro, extracción de características, adaptación a nuevas condiciones basales y clasificadores, así como también el tipo de representación de las emociones, son el objeto de estudio de esta tesis.

A partir del marco descripto, se plantea la siguiente hipótesis general de investigación: es posible desarrollar arquitecturas de aprendizaje y modelos guiados por los datos que mejoren las tasas de reconocimiento actualmente alcanzadas para estados afectivos, registrando un mínimo de señales fisiológicas, idealmente sólo una.

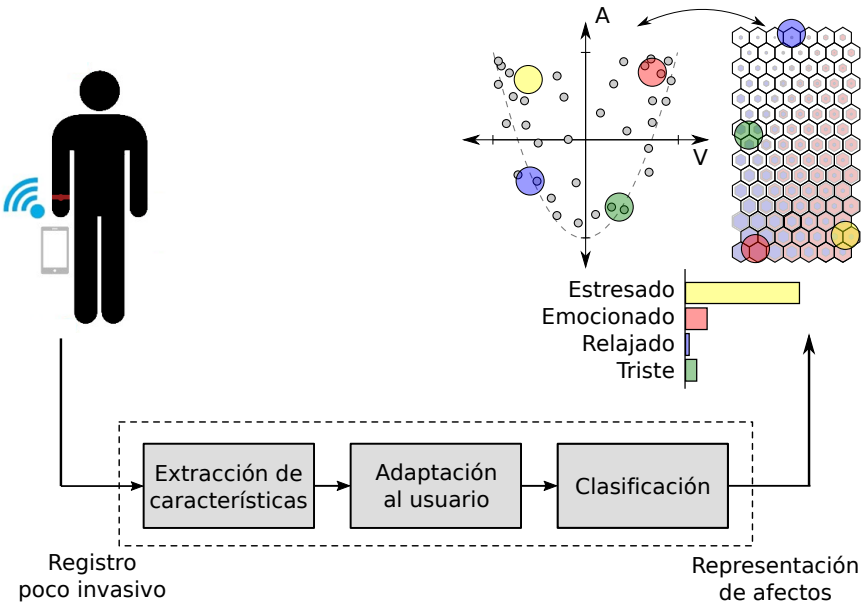


Figura 1.1: Diagrama de bloques de un sistema elemental de registro basado en sensores corporales. A partir del registro, las etapas de extracción de características, adaptación al usuario y clasificación, el sistema hace una estimación del estado afectivo, en alguna representación conveniente.

1.2 Objetivo general

Desarrollar nuevos algoritmos para el reconocimiento de estados emocionales a partir de señales fisiológicas y proveer una mejora significativa de las tasas de reconocimiento actualmente alcanzadas en situaciones realistas, con requerimientos mínimos en cuanto a la instrumentación, para alcanzar una baja invasividad al usuario.

1.3 Objetivos específicos

- Analizar las características de las señales fisiológicas desde la perspectiva de su capacidad discriminativa para el reconocimiento de emociones, con independencia del usuario y robustez a los cambios en las condiciones del registro.
- Identificar la menor cantidad posible de señales que permita discriminar los estados afectivos, con el menor grado de instrumentación para el usuario.
- Desarrollar sistemas de reconocimiento continuo de diferentes tipos de emociones en base a información de las señales fisiológicas, que puedan utilizarse cercano a *tiempo real*.
- Desarrollar nuevas técnicas dirigidas por datos que permitan encontrar relaciones subyacentes entre la fisiología y los modelos de emociones, de forma

tal de que puedan ser analizadas luego por un experto.

- Validar los sistemas de reconocimiento propuestos en condiciones realistas de funcionamiento.

1.4 Organización de la tesis

Este documento se presenta bajo el formato de tesis por compilación, estructurada de la siguiente manera. En la Sección 2 se hace una revisión del estado del arte en relación a la hipótesis y los objetivos propuestos. En la Sección 3 se presenta la metodología desarrollada para abordar los desafíos mencionados, destacando los aportes originales de esta tesis. En la Sección 4 se presentan y discuten los resultados obtenidos y en la Sección 5 se destacan los principales aportes realizados al estado del arte. Finalmente, en la sección de Anexos se encuentran las publicaciones que sustentan el aporte original de la tesis.

2 Estado del arte

En esta sección se tratan los principales aportes en la bibliografía sobre el reconocimiento de estados afectivos a partir de señales biomédicas. Es necesario hacer una breve revisión de los modelos que se utilizan para representar los estados afectivos; tanto el tipo de representación como los métodos experimentales que se utilizan para el registro son importantes para el desarrollo del reconocedor. Luego se detallan las señales y características importantes de dos subsistemas que son de interés para el reconocimiento de los afectos y tienen gran potencial para su uso en la práctica. Finalmente, se mencionan los diferentes métodos de aprendizaje maquinal que se ha investigado en la bibliografía.

2.1 Emociones y desafíos actuales

En el enfoque tradicional de computación afectiva se han utilizado modelos categóricos para describir las emociones, definidos como clases no ordenadas [17]. El conjunto más conocido son las seis emociones básicas (alegría, tristeza, miedo, enojo, disgusto y sorpresa), que serían comunes a todas las personas sin importar su cultura o idioma, también llamadas emociones universales [32]. También se han utilizado otras clases definidas para campos de aplicación particulares. Por ejemplo, en educación y entretenimiento, es interesante modelar como se siente el usuario respecto a la tarea que está realizando. Una sobrecarga de esfuerzo puede aumentar los sentimientos negativos como la ansiedad, mientras que una carga baja (durante un trabajo muy fácil) podría convertirse en una tarea aburrida, y así en un condicionamiento negativo para desarrollar cualquier actividad [33]. Es

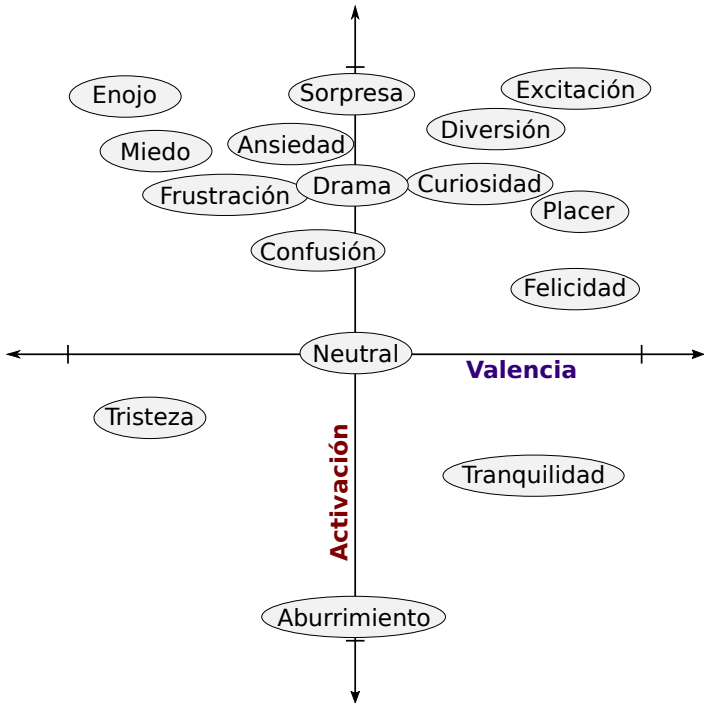


Figura 2.1: El modelo de afectos centrales, definido por los ejes cartesianos de activación y valencia, establece un plano basado en percepciones intuitivas. Se representan algunas emociones categóricas según se definieron para aplicaciones en educación [38], categorización de videos [29] y música [39].

por esto que las aplicaciones buscan que el usuario esté en un estado de entusiasmo para poder afrontar los desafíos [34]. En este caso, los estados negativos de aburrimiento o frustración, y los positivos como el entusiasmo, resultan más descriptivos para la aplicación particular [7, 35].

Otras teorías han planteado representaciones afectivas en forma de variables continuas, definiendo de esta forma espacios n-dimensionales donde cada punto representaría un estado afectivo particular. Una de las más utilizadas describe toda la gama de emociones y estados de ánimo utilizando estados neurofisiológicos principales llamados *afectos centrales*, cuya experiencia consciente puede ser representada en un plano (Fig. 2.1) [36]. Las componentes de este espacio son la *valencia* (nivel de placer) y la *activación* (nivel de excitación), definiendo un modelo muy intuitivo ya que se basa en percepciones básicas de las personas. También se han propuesto otras dimensiones para caracterizar los afectos en diferentes circunstancias, como por ejemplo el grado de control que uno siente sobre la situación o la sensación de novedad ante un evento [37].

Para definir la percepción de una emoción se requiere calificarla en términos absolutos, ya sea encasillándola en una clase o indicando un valor numérico. Esto muchas veces no es sencillo en la práctica, ya que los participantes tienen internalizadas diferentes escalas a las que harán referencia. Por esta razón algunos

autores [40, 41] prefieren describir los estados afectivos de una manera relativa (etiquetado por ranking) en lugar de una escala numérica absoluta (etiquetado por calificación).

Los trabajos actuales parecen no acordar sobre cuál es el modelo que representa más fielmente de las emociones. Se ha argumentado que las etiquetas categóricas no son convenientes porque las definiciones entre culturas e idiomas pueden variar, siendo un desafío para la comunicación e interpretación [31]. Por otro lado, muchas características humanas varían más bien de forma continua, y restringir los afectos a categorías puede llevar a perder información del modelo de fondo. Aún así, como sugirió Picard [42], más allá de la exactitud de las teorías desarrolladas hasta ahora, éstas podrían no llegar a explicar todo el espacio afectivo, o bien explicarlo de una forma poco práctica, por lo que los modelos deberían seleccionarse para cada situación particular [1, 17]. Se llevaron a cabo estudios teóricos y empíricos sobre las relaciones entre los modelos emocionales mencionados, con el objetivo de validar teorías o identificar cuáles serían las más útiles para definir afectos complejos [43, 44], y también se utilizaron modelos en conjunto para mejorar las tasas de reconocimiento [45, 46]. Existen coincidencias en los conceptos de los diferentes modelos, y usualmente las emociones categóricas pueden ubicarse en regiones de los espacios dimensionales, como se representa en la Figura 2.1. Aún así, no se han explorado en profundidad técnicas de inteligencia computacional que modelen diferentes espacios emocionales (categóricos y dimensionales) y permitan analizar la información subyacente en la fisiología.

Para desarrollar un reconocedor en condiciones realistas, es necesario tener en cuenta como se incitan y registran experimentalmente los afectos, ya que son aspectos importantes para describir el alcance y limitaciones de la validación experimental. En los inicios de la computación afectiva, se realizaron registros sobre emociones actuadas [17]. Este diseño experimental permitió realizar pruebas de concepto en el reconocimiento de afectos a partir de señales fisiológicas y sentar bases sobre los modelos afectivos y el procesamiento de las señales. El siguiente nivel de realismo se obtiene al exponer al participante con estímulos de diferentes características y de esta forma inducir estados afectivos, o generar una respuesta empática. Entre estos estímulos existen algunos estándares, como el sistema internacional de imágenes afectivas (IAPS, del inglés *international affective images system*) [47], el sistema internacional de sonidos digitales (IADS, del inglés *international affective digitalized sounds*) [48], y otros ampliamente utilizados como fragmentos de películas [29] o videos musicales [37]. En estos casos, si bien los estados inducidos no son espontáneos, sí se pueden considerar realistas cuando se busca, por ejemplo, clasificar automáticamente el contenido multimedia. Los estudios más recientes se basan en emociones espontáneas, como aquellas que surgen naturalmente en una discusión [10] o al interactuar con sistemas inteligentes [49]. Esta nueva modalidad presenta ciertos desafíos, ya que las respuestas emocionales podrían ser más bien sutiles y cambiar en función de la dinámica de las interacciones, algo lejano a la distinción de emociones prototípicas que se es-

tudiaba anteriormente. Éste es el estado actual de los desafíos donde se registran señales fisiológicas.

2.2 Señales fisiológicas y extracción de características

Como se ha mencionado en la introducción, existen numerosos trabajos que hacen uso del EEG o las señales periféricas para el reconocimiento de afectos. Actualmente hay una importante tendencia al desarrollo de sistemas de reconocimiento multimodales, que incorporan información de diferentes fuentes [50, 51, 52]. Esto se debe a la robustez que brinda tener fuentes variadas de información, y además esta inspirado en la forma natural en la que las personas reconocen los afectos: a través de la integración de pistas en la voz, expresiones corporales y actitudes. Sin embargo, en el caso de las señales fisiológicas existe un compromiso importante entre las ventajas de utilizar múltiples señales y el problema real de la invasividad. Es por esto que, sobre todo para el reconocimiento de afectos a partir de la fisiología, mejorar las tasas de reconocimiento sobre pocas fuentes, e incluso una única señal, es relevante y no ha sido explorado profundamente en la bibliografía.

Las señales biomédicas poseen en común ciertas características propias que dificultan su procesamiento por medios artificiales. Entre ellas se pueden mencionar: su importante variabilidad intra o inter-individuo, el estar contaminadas con ruido de naturaleza e intensidad muy diversa y el hecho de ser producidas o percibidas por mecanismos intrínsecamente no lineales y variantes en el tiempo. Desde la perspectiva del reconocimiento de patrones, estas características resultan en patrones con importantes dinámicas temporales, complejas fronteras de decisión, alta dimensionalidad, redundancia de información y distribuciones no Gaussianas en los datos y en sus clases. Todo esto hace que el reconocimiento de emociones a partir de estas señales sea una tarea muy desafiante en la actualidad [31].

Se han realizado diversos estudios sobre la incidencia de las emociones sobre diferentes señales, encontrándose numerosos indicios de estos efectos [19]. Además, se han generado corpus de uso libre con gran variedad de desafíos, para que las nuevas contribuciones puedan evaluarse comparativamente en condiciones similares [17, 37, 29, 10]. En esta sección se detallará el estado del arte para dos fuentes que son de mucho interés para la computación afectiva fisiológica (ver Sección 3.1), y sobre los que se ha basado esta tesis: la actividad electrodérmica y las respuestas del sistema circulatorio. Se mencionan las características propuestas en la bibliografía para representar la señal con la menor dimensionalidad y redundancia, creando un conjunto de variables que sean representativas de los datos y destacando aquellas que puedan ser utilizadas para reconocimiento en tiempo real.

2.2.1 Actividad electrodérmica

La EDA es un término amplio para designar el efecto del SNA sobre las propiedades eléctricas de la piel, producto de la actividad glandular del sudor y otros mecanismos [53]. El método que actualmente se utiliza para medir estos cambios consiste en aplicar una corriente de baja intensidad entre un par de electrodos superficiales en la piel, por ejemplo en la muñeca o entre dos falanges [54]. La transpiración, asociada a reacciones de sorpresa o estrés, produce un puente iónico entre los poros, aumentando la conductividad aún sin producir cambios visibles. Esta señal se ha reconocido desde hace mucho tiempo como un parámetro importante que responde a la actividad de la rama simpática [55] y fue considerada entre las bioseñales más evidentes para la detección de eventos emocionales [56, 10].

La señal de EDA presenta una componente característica de cambio lento o nivel de conductancia (SCL, del inglés *skin conductance level*) y una de cambio rápido o respuesta de conductancia (SCR, del inglés *skin conductance response*). La SCL, entendida como la variación de la amplitud a baja frecuencia, se asocia con el estado basal, que puede estar influenciado por estados de ánimo, humedad ambiente o cambios fisiológicos diarios. Por otro lado, la SCR suele estar relacionada con los eventos de excitación instantáneos, aunque se ha encontrado que algunos componentes frecuenciales no corresponden a estímulos emocionales [2]. En Figura 2.2 se ilustra el efecto de un estímulo mientras se registra la EDA con diferentes pares de electrodos. Las particularidades del registro durante la respuesta emocional se pueden observar en la amplitud, cambios repentinos y regularidad de la señal. El análisis básico de la EDA consiste en filtrar las altas frecuencias (>10 Hz) y separar las componentes SCL y SCR, en general utilizando filtros pasa banda u otros métodos guiados por los datos [57].

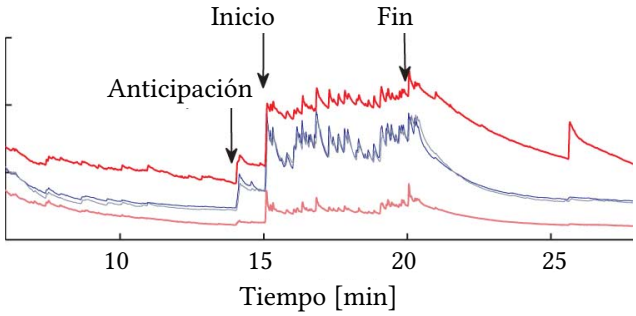


Figura 2.2: EDA registrada con diferentes sensores mientras se presenta un fragmento de película (género de terror). Se pueden observar respuestas que marcan el inicio y fin del estímulo, así como una activación inicial dada por la anticipación del participante al experimento [54].

Se han utilizado características básicas para medir los cambios en esta señal con buenos resultados. Por ejemplo, a partir de ventanas móviles sobre la SCL y SCR se extraen valores medios, desvíos y momentos estadísticos de diferente orden (curtosis y asimetría), derivadas y amplitud de las envolventes [58, 59, 37, 27].

El número de picos y cruces por cero se han utilizado para caracterizar las respuestas rápidas [60]. En el dominio de la frecuencia, se analizaron componentes espectrales, dividiendo el espectro de la señal en bandas arbitrarias [37, 3], descomposiciones ondita [61] y otras medidas usuales en señales en el tiempo [62]. También se combinaron características simples (valores medios y derivadas) a partir de ventanas de diferente tamaño, para estimar cambios a corto y largo plazo [63]. Por último, se emplearon medidas derivadas de teoría de la información, como la entropía aproximada, con el fin de cuantificar la regularidad de la señal [64].

2.2.2 Señales del sistema circulatorio

La actividad cardiovascular, además de su relevancia clínica, refleja cambios fisiológicos relacionados a los estados emocionales [19]. Estas respuestas presentan dinámicas no lineales y varían entre individuos, según el estado de actividad o momentos del día. Todo esto hace que el análisis de estas señales sea complejo para identificar eventos particulares, como las emociones.

El método de registro de referencia es el ECG, una medida poco invasiva de la actividad eléctrica del corazón. Al conectar un par de electrodos en dos puntos entre los que medie el corazón, por ejemplo en las muñecas o en el pecho, se puede registrar la actividad eléctrica durante cada latido. En las aplicaciones de computación afectiva suele ser suficiente utilizar un único canal o derivación, lo que también resulta en menor invasividad. El ECG tiene una morfología cuasicíclica, compuesta por el complejo PQRST (Fig. 2.3). Cada segmento del período corresponde a una etapa específica del ciclo de contracción y relajación del corazón, relacionada con la despolarización progresiva y la repolarización de las fibras musculares. Se ha encontrado que la morfología de la señal puede cambiar con los estados mentales; por ejemplo en el estrés, se encuentra incrementada la distancia Q-T aún manteniendo un ritmo cardíaco constante [65]. Se han utilizado diferentes medidas entre los puntos característicos del ECG [25] y medidas clásicas como los momentos estadísticos.

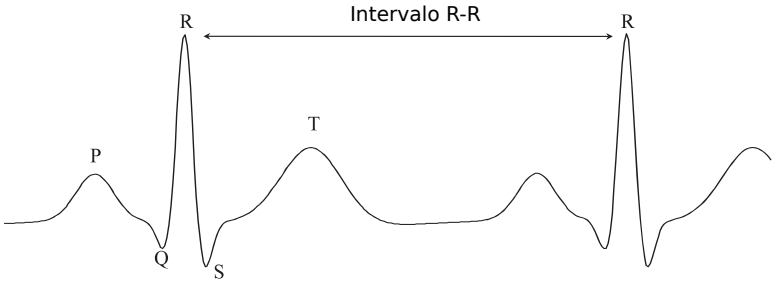


Figura 2.3: Etapas de despolarización y repolarización del músculo cardíaco reflejadas en la señal de ECG, definidas por los puntos característicos PQRST [25].

Otro método muy utilizado para registrar la respuesta del sistema circula-

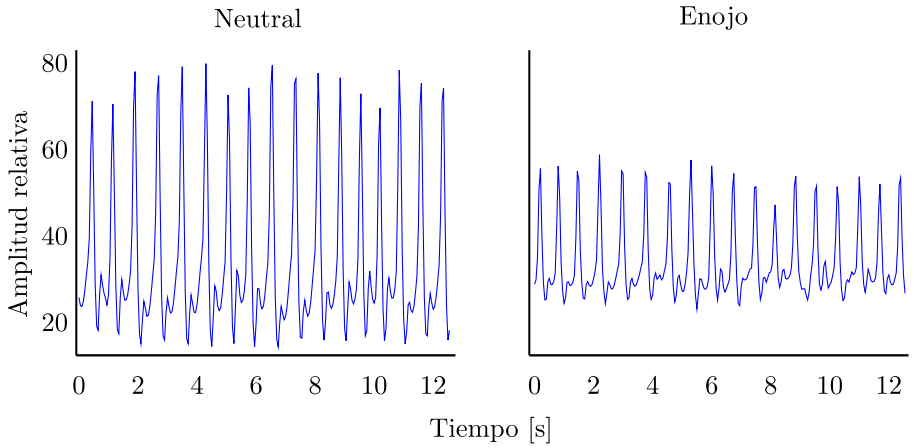


Figura 2.4: Segmentos de PPG en estados neutral y de enojo [17]. Las señales están compuestas por los picos de pulso sanguíneo frente al sensor, mostrando diferencias en amplitud y regularidad para los dos estados.

torio es la PPG (Fig. 2.4). Esta señal representa el flujo sanguíneo a través de pequeños vasos, y es medida utilizando un dispositivo con un led emisor y un fotosensor. La medición puede ser por transmisión o reflexión. La transmisión es la utilizada por los oxímetros de pulso, detectando la absorción de la luz a través de una falange. En cambio la reflexión es utilizada en las bandas deportivas o las aplicaciones para celular [66]. Los avances en el procesamiento de señales han permitido también aproximar esta señal a partir de sutiles variaciones en el color de la piel, utilizando una cámara web [67]. La señal obtenida por cualquiera de estos métodos permite aproximar el ritmo cardíaco (HR, del inglés *heart rate*) y se espera que tenga una buena precisión cuando los sujetos están en posiciones de reposo [68, 69]. La amplitud del pulso es proporcional con el flujo sanguíneo y refleja el nivel de vasoconstricción, que se encontró incrementado en situaciones de alta activación [70]. En la Figura 2.4 se pueden comparar señales correspondientes a un episodio de ira y un estado neutro, donde aparecen diferencias en la amplitud y la regularidad de la señal, que se han explotado para el reconocimiento de afectos [17] y ansiedad [26].

Un análisis derivado de estas señales que merece un tratamiento especial es la variabilidad del ritmo cardíaco (HRV, del inglés *heart rate variability*), basado en la medición de la distancia entre latidos en el tiempo. Actualmente la HRV de referencia es la que se obtiene del ECG, ya sea con equipos de grado médico o cintas pectorales deportivas. Los latidos están marcados por los picos R, y se detectan usualmente con el algoritmo de Pan-Tompkins [71], que es adaptativo en el tiempo y tiene robustez al ruido, fluctuaciones de baja frecuencia y latidos ectópicos. Luego, se mide la distancia R-R en cada punto y se interpola la señal a una tasa de muestreo uniforme. Se encontró que se puede obtener una buena aproximación de la HRV a partir del PPG en reposo [72], mientras que los sensores corporales más recientes reportan una buena calidad de señal en el uso diario [73].

Comparando el uso de HRV y ECG para reconocimiento de emociones [74], se encuentra que la HRV ha llevado a los mejores resultados, lo que podría explicarse en que las características del complejo PQRST estarían más bien relacionadas a los estados anatomo-fisiológicos basales.

En el dominio frecuencial se diferencian tres bandas de importancia de la HRV: una componente de alta frecuencia (HF, del inglés *high frequency*) (0.15-0.40 Hz), una región de baja frecuencia (LF, del inglés *low frequency*) (0.04-0.15 Hz) y una de frecuencia muy baja (<0.04 Hz), aunque esta última se cree más relacionada con la termorregulación o el ciclo circadiano [75]. Para estimar estas componentes, que varían a través del tiempo, el método más utilizado es la transformada de Fourier en tiempo corto (STFT, del inglés *short-time Fourier transform*) [57]. También se propusieron modelos paramétricos, como los modelos auto-regresivos, con una mejor resolución de frecuencia y robustez respecto al espectrograma para el reconocimiento de emociones en registros muy cortos (<5 min) [76, 77, 78, 79]. En el control del ritmo cardíaco influyen ambas ramas del SNA; la red parasimpática se asocia a la zona HF, mientras que la LF estaría regulada por ambas ramas [80]. De esto se desprende que la relación de energías entre las bandas LF y HF se use como índice de la regulación simpatovagal [43, 37], pudiendo reflejar cambios emocionales.

Se ha encontrado que la amplitud de la HRV y el HR máximo se modifican con estímulos placenteros [81]. Durante el esfuerzo mental y el estrés, se observan cambios en el HR [82] y las componentes frecuenciales de la HRV [77]. En experimentos con meditación se encontró que el HR no varía significativamente pero sí lo hacen las componentes LF y HF [83]. La influencia respiratoria también aparece en la región HF debido a la compresión del nervio vagal durante el ciclo respiratorio, conocida como arritmia sinusal respiratoria. Esta componente podría proveer información de los patrones respiratorios sin tener que medir directamente estos movimientos. Se encuentra que las oscilaciones respiratorias pueden ser más o menos regulares, por lo que en el espectro del HRV se pueden ver picos o señales de banda ancha [84]. Esta componente se ha utilizado, por ejemplo, para identificar emociones en interacciones con agentes virtuales [85].

Se propusieron también características diseñadas *ad hoc* para aplicaciones en medicina, como los métodos basados en el histograma de la HRV [77]. Se usaron diferentes métricas temporales para medir el estrés, como el promedio de intervalos R-R de más de 50 ms (pNN-50), la raíz cuadrática media de las diferencias de intervalos R-R sucesivos (RMSSD, del inglés *Root mean square of the successive differences*), desvío estándar de latidos normales (SDNN, del inglés *standard deviation of NN intervals*), entre otros índices [43, 25]. Entre estos, el análisis de Poincaré relaciona la variabilidad de corto y largo plazo de los intervalos R-R, describiendo su dinámica [86]. Muchas de estas características fueron diseñadas para registros relativamente largos (usualmente 5 minutos como mínimo [77]), por lo que en muchos casos carece de fundamento utilizarlos para identificar respuestas en el orden de los segundos.

Por último, se exploraron también las características no-lineales de la HRV, como el análisis de ciclicidad [87] y diferentes técnicas para identificar el modelo no-lineal de fondo [88]. Se propuso un análisis basado en exponentes de Hurst, una aproximación no lineal que proporciona información sobre la auto-similitud, y se capturaron cambios emocionales en seis categorías emocionales [89]. El comportamiento recurrente, donde se alternan ciclos periódicos e irregulares, se exploró en ECG y EDA, utilizando tres algoritmos de extracción de características diferentes [59, 90]. También se caracterizó el comportamiento caótico característico del HRV mediante los exponentes de Lyapunov [91]. Tomando el exponente dominante, se puede aproximar el modelo a partir de registros cortos.

2.2.3 Selección, transformación y aprendizaje de características

La gran cantidad de características obtenidas puede ser redundante (la información contenida en varias de ellas se repite) o irrelevante (no son discriminativas para el problema), debido a la falta de conocimiento sobre el modelo que genera los datos o la inespecificidad de los métodos de extracción de características. En algunos trabajos se utilizaron algoritmos para seleccionar subconjuntos de características, evaluando iterativamente el error de clasificación [58, 9, 78, 86], utilizando diferentes técnicas de optimización [92, 17, 93] o algoritmos genéticos [94, 21]. Otra estrategia es transformar el espacio de características para reducir la dimensionalidad de datos, por ejemplo usando análisis de componentes principales (PCA, del inglés *principal component analysis*) [95, 59] y tomando las componentes que representan la mayor variabilidad de los datos. Una desventaja de las transformaciones de características es que se pierden las relaciones entre las características originales y los resultados del clasificador.

Más recientemente se han utilizado métodos basados en redes convolucionales, que permiten extraer información directamente de la señal temporal, evitando la definición de características y con alta flexibilidad para combinar señales de diferente naturaleza [40, 96]. En estas redes se entrena la etapa de extracción de características junto al clasificador, método conocido como *end-to-end*. Estos sistemas requieren cierto volumen de datos de entrenamiento o el uso de técnicas de aprendizaje por transferencia de información de otros dominios.

2.3 Reconocimiento de emociones

En esta sección se discutirán las diferentes tareas de reconocimiento de emociones que son de interés actual, y los principales enfoques de aprendizaje automático que se han empleado para tal fin.

2.3.1 Tareas de reconocimiento

Los modelos y aplicaciones de la Sección 2.1 definen diferentes tareas a resolver en cuanto al reconocimiento de patrones. El primer desafío planteado fue el reconocimiento de emociones categóricas en señales segmentadas, partiendo de fragmentos de registros donde se conoce previamente que deben contener información de una clase particular [17]. Los modelos dimensionales, aunque son de naturaleza continua, se utilizaron originalmente en forma cuantizada, reduciendo el problema al de clasificación en cuadrantes de plano AV [37]. Sin embargo, al reducir las etiquetas a clases categóricas se pierde información del modelo subyacente [97]. También se ha evaluado el reconocimiento conjunto (multitarea) de modelos categóricos y dimensionales, aunque las anotaciones según diversos modelos afectivos para el mismo experimento no son comunes [46].

Una situación más desafiante consiste en partir de un registro continuo en el tiempo, que no esta segmentado. Si el procesamiento de señales y clasificación se realizan una vez finalizada la etapa de registro de los datos, el proceso se puede definir como reconocimiento en tiempo diferido (del inglés *offline*), y es utilizado ampliamente en muchas aplicaciones, por ejemplo para investigar respuestas a estímulos multimedia [5]. En aplicaciones interactivas, una situación más realista supone que un nuevo usuario se conecta a un dispositivo, comienza el registro y, luego de una breve etapa de adaptación, el clasificador debería poder realizar estimaciones con una demora y error razonables. Como contraposición al caso anterior, el reconocimiento en tiempo real (del inglés *online*) supone restricciones sobre el sistema, tanto en el costo computacional como en los datos que se tienen disponibles al momento del reconocimiento. Estos desafíos se han explorado sólo superficialmente [17, 98], ya que en la mayoría de los casos se utilizan señales pre-segmentadas, o haciendo el análisis de datos con posterioridad al registro.

En el avance del análisis sobre señales continuas en tiempo real, una de las primeras tareas fue la identificación de eventos, en la cual se busca identificar el instante en el que el participante pasa por algún cambio importante [99]. Más recientemente, el reconocimiento continuo de los afectos se ha vuelto de gran interés en el área [31, 100]. Aquí el término *continuo* hace referencia por un lado al uso de los registros en tiempo real, y por el otro a que la variable de referencia es una variable continua en lugar de las emociones *prototípicas*, definiendo un problema de regresión (en lugar de clasificación).

2.3.2 Métodos de aprendizaje automático

En gran parte de la bibliografía se realizan registros de datos particulares para responder a las preguntas planteadas. En este sentido, si bien se han comparado numéricamente algunos métodos básicos de aprendizaje maquinal [101, 9], la comparación de algunos modelos avanzados es limitada ya que parten de métodos experimentales, modelos afectivos y sistemas de registro que son muy diferentes. A continuación se describe, en términos generales, los métodos utilizados para el

reconocimiento de afectos a partir de las señales estudiadas, con algunas de sus ventajas y limitaciones.

Para el reconocimiento categórico se han comparado diferentes métodos clásicos de aprendizaje maquinal. K-vecinos más cercanos (kNN, del inglés *k-nearest neighbours*) es un método simple y robusto, ya que no requiere definir un modelo para los datos sino que se etiquetan por vecindad [17, 102, 60]. Los árboles de decisión se utilizaron para clasificar cuadrantes del plano AV [86], estados de estrés [103] e identificar respuestas rápidas en juegos [104]. En este método, las funciones de separación de nodos están basada en la información mutua [9] o el índice de diversidad de Gini [89]. En numerosas publicaciones se utiliza el perceptrón multicapa (MLP, del inglés *multi-layer perceptron*), con varias arquitecturas y algoritmos de entrenamiento [12, 105, 101, 106, 107]. Finalmente, se utilizaron también ideas de lógica difusa para mapear datos multimodales al espacio AV. Las funciones de pertenencia (de las muestras a los respectivos cuadrantes AV) se generaron a partir del histograma de cada señal, agregando reglas de expertos del dominio [6].

Las máquinas de soporte vectorial (SVMs, del inglés *support vector machines*) son métodos de aprendizaje supervisado que buscan minimizar el error de clasificación maximizando el margen geométrico entre las clases en el espacio de características, reportando buenos niveles de generalización en conjuntos de características grandes y pocas muestras de entrenamiento. Las SVMs lineales se utilizaron en aplicaciones de computación afectiva en diversos conjuntos de características y modelos de emoción [43, 108, 109]. También se exploraron funciones de núcleos no lineales, como las de base radial [2, 107, 29] y polinomial [21]. El objetivo en estos casos es transformar las características fisiológicas a un espacio de mayor dimensión, en el que se presume que el problema se puede separar linealmente. Para el reconocimiento de afectos continuos, las SVM para regresión (SVR) fueron ampliamente utilizadas, con buenos resultados [110, 111, 112].

Los modelos gráficos probabilísticos permiten representar las interacciones entre variables fisiológicas y estados afectivos. Se utilizaron clasificadores basados en *naïve* Bayes sobre diferentes conjuntos de datos debido a su capacidad para tratar con las clases desequilibradas en conjuntos de entrenamiento reducidos [37, 59, 9, 56]. Para tratar la dinámica de la fisiología ante diferentes estados emocionales se propusieron los modelos ocultos de Markov (HMM, del inglés *hidden Markov models*) [113, 11, 114]. El registro continuo de señales biomédicas permite el modelado de esta dinámica y las transiciones entre estados afectivos. También se exploraron técnicas de ensambles, utilizando numerosos clasificadores de baja complejidad para conformar una frontera de decisión de gran complejidad [22]. Usando algoritmos como AdaBoost [95], sistemas de votación [8] o integración de errores [93] con clasificadores estándar (SVM, kNN, árboles de decisión, entre otros), se obtuvieron buenos resultados en conjuntos reducidos de estados emocionales. Los métodos que reportan las mejores tasas de reconocimiento son los ensambles basados en redes neuronales profundas (DNN, del inglés *deep neural*

networks) [115] y el de bosque aleatorio (RF, del inglés *random forest*) [64].

En cuanto al reconocimiento de variables continuas, se han hecho numerosos aportes recientes [100], con clasificadores tales como DNNs [116], redes con memoria a corto y largo plazo (LSTMs, del inglés *long short-time memory neural networks*) [117] y máquinas de relevancia vectorial (RVMs, del inglés *relevance vector machines*) [118]. Algunos métodos se han adaptado al problema continuo por regresión mediante clasificación [96], una técnica que consiste en cuantizar la etiqueta en N clases, lo que permite entrenar cualquier clasificador como si fuese un problema categórico. Analizando los resultados de los diferentes aportes, se encuentra que éstos aún deben mejorarse mucho para avanzar en aplicaciones sobre interacciones naturales.

3 Metodología

En esta sección se propone un criterio para elegir las señales fisiológicas, y se menciona el procesamiento de las señales y la etapa de extracción de características. Luego se presentan los métodos de aprendizaje maquinal desarrollados para esta tesis, seguido de la metodología de validación de las soluciones propuestas: materiales, medidas de desempeño y diseño de la validación experimental. Para facilitar la reproducibilidad de los resultados, se encuentra a disposición el código fuente[1] que genera los principales resultados de esta tesis. Para detalles adicionales sobre la metodología propuesta, ver la sección de Anexos.

3.1 Selección de las señales fisiológicas

Con el objetivo de lograr un sistema de reconocimiento afectivo que sea factible y práctico de usar, es deseable reducir las señales fisiológicas necesarias al mínimo. Para hacer esta selección, es necesario definir un criterio, contemplando las ventajas y limitaciones de cada señal. Además del análisis de los resultados previos, en la Figura 3.1 se proponen diferentes propiedades de las señales y métodos de registro, basadas en el estudio de la bibliografía y las tecnologías actuales. En las columnas se separan las señales biomédicas usuales para los principales sistemas fisiológicos. En las filas se listan las propiedades de interés para la computación afectiva. El grado de invasividad de los sensores y su robustez al ruido y a cambios en las condiciones de registro son factores importantes para lograr métodos prácticos, que puedan ser utilizados fuera del laboratorio. El tiempo que demora en aparecer un cambio significativo en la señal ante un estímulo afectivo es relevante para reconocer cambios instantáneos, en aplicaciones interactivas. Finalmente, dos aspectos que no son centrales pero sí relevantes para la investigación son el grado de disponibilidad de las señales en los corpus públicos y el grado de novedad en cuanto a la cantidad de trabajos y profundidad que se ha alcanzado

[1] `https://sourceforge.net/projects/sourcesinc/files/emoHR`.

en ellos. El puntaje cualitativos que se asigna a cada propiedad se indica con el area sombreada de cada celda.

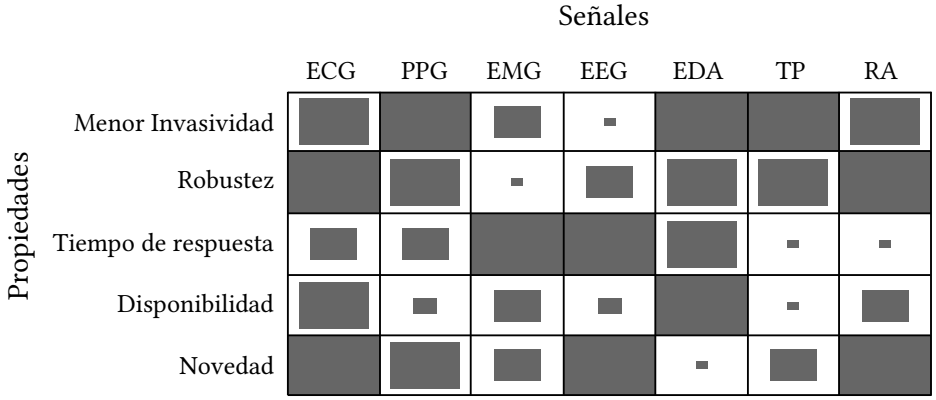


Figura 3.1: Propiedades relevantes de las señales fisiológicas para ser utilizadas en contexto de una aplicación de computación afectiva. Para cada una de las propiedades descritas, se asigna un puntaje a cada señal fisiológica, indicado con el area pintada en cada celda.

Los registros menos invasivos son los basados en un único sensor, como el de temperatura, o en un par de electrodos puntuales como EDA, PPG y ECG, mientras que en el extremo opuesto se encuentra el EEG. En particular, el ECG requiere actualmente una cinta pectoral, al igual que para los patrones respiratorios, mientras que EDA y PPG se podrían obtener de una pulsera. En cuanto a la robustez al ruido, el ECG se encuentra en mejor posición, siendo ampliamente utilizado para aplicaciones médicas y deportivas en diversos contextos. El retraso de las respuestas podría ser significativo para los patrones de respiración y temperatura de la piel, que pueden demorar decenas de segundos en registrar cambios, mientras que las otras fuentes registran cambios en pocos segundos (ECG y PPG) o prácticamente instantáneos (EEG). Para analizar la disponibilidad, se contabilizaron la cantidad de apariciones en corpus de acceso libre, donde se puede apreciar que la EDA y el ECG son las señales de preferencia. Finalmente, el análisis de la dinámica compleja de los patrones respiratorios y la HRV [119], así como el análisis necesario para identificar componentes relativas a las emociones en el EEG [37] indican que existe información subyacente a los modelos que no está establecida en trabajos clásicos de sicofisiología.

La bibliografía muestra que las señales listadas en la Figura 3.1, comparadas en diversos experimentos de computación afectiva, reportan en todos los casos algún grado de capacidad para discriminar afectos [19, 57]. Como se resume en la Sección 2.2, las señales cardiovasculares y EDA han sido ampliamente usadas en la investigación sicofisiológica y en diversas aplicaciones. A partir del análisis de la figura y los resultados ya mencionados del estado del arte, la señal de ECG (en particular la HRV) y la EDA constituyen buenos candidatos para un sistema de reconocimiento de afectos en tiempo real, que posibilite el uso fuera

del laboratorio, y por tanto se utilizarán para los experimentos siguientes.

3.2 Procesamiento de las señales

A partir de los registros crudos se realiza un pre-procesamiento estándar para eliminar artefactos y componentes espectrales que no tienen información de los sistemas fisiológicos estudiados [57]. Para cada registro, se realizó un espejado de la señal en los extremos (10 s) para evitar el efecto de borde en el procesamiento posterior y estimar las características desde el instante inicial. Para la identificación de los picos QRS en ECG se utilizó el algoritmo de Pan-Tompkins [71]. Luego el ritmo cardíaco se obtiene a partir de la inversa del intervalo R-R y la señal se interpola a 25 Hz. En la señal de PPG los máximos de los picos de pulso tienen una tasa de cambio mucho más lenta que los QRS, por lo que la detección del instante de cada pico es sensible al ruido y conlleva un error que no es despreciable cuando se analiza la variabilidad del ritmo cardíaco. Un método más preciso es identificar el máximo de la pendiente ascendente de los pulsos, debido a que tienen un cambio más rápido [120]. A partir de estos puntos de máxima pendiente, se puede estimar la posición de los picos y valles evaluando la tasa de cambio en la señal, lo que luego se usa para determinar la envolvente de la señal y la amplitud. Para la EDA, se estimaron las componentes SCL (0– 0.5 Hz) y SCR (0.5–1 Hz) con un filtro Butterworth de 3er orden [10].

3.2.1 Extracción de características

Se realizó la extracción de características con una ventana móvil, simulando registros en tiempo real. La ventana debe tener la longitud suficiente para calcular características significativas y a la vez poder capturar las variaciones rápidas y lentas. Se utilizó una ventana Hamming de 20 s y un paso de 0.5 s. La longitud utilizada permite obtener suficientes datos temporales y una adecuada resolución frecuencial para las bandas de interés sin perder resolución temporal [121]. A partir de las características descriptas a continuación, se adicionaron características de las variaciones temporales a partir de las derivadas y concatenando características de ventanas adyacentes (información de contexto temporal).

Para cada señal se calculó una serie de características básicas: valores medios, desviación, extremos, y momentos estadísticos de alto orden (curtosis y asimetría). Además, para el ECG se evaluaron características propias como las distancias entre puntos característicos del complejo PQRS, pero fundamentalmente se extrajeron características de la HRV. Para el análisis espectral se utilizó la transformada de Fourier sobre cada ventana. Se utilizó la energía de las bandas LF y HF, su radio LF/HF y la energía total de la variabilidad (descontando el valor del HR). La componente respiratoria se estimó identificando el punto máximo en la zona de frecuencia respiratoria (0.18-0.5 Hz). También se consideró la tasa de decaimiento frecuencial, que se encuentra relacionada con el control autónomo

[122]. Para esto, se modeló una tasa de decaimiento cuadrática y se utilizaron los coeficientes como características. De la señal de PPG se obtuvo el valor medio y desvío de la amplitud de las envolventes, y los mismos métodos de la HRV se aplicaron a la variabilidad de los picos de pulso.

3.2.2 Adaptación a nuevos registros

Para la mayoría de los métodos de aprendizaje maquinal es necesario normalizar las características a un rango similar, con el fin de que todas las variables tengan el mismo peso en el clasificador. Como se mencionó anteriormente, el enfoque que da mejores resultados es normalizar cada sesión de registro por separado, antes de pasar a la etapa de clasificación (en diferido), para minimizar el efecto de la variabilidad entre sesiones. El enfoque tradicional que permite el uso del clasificador en tiempo real es estimar parámetros de los datos de entrenamiento (como la media y la varianza) y luego utilizarlos para normalizar los datos nuevos que llegan al clasificador. Este método parte del supuesto que la distribución de los datos de prueba será similar a la de los de entrenamiento. Sin embargo, las diferencias entre participantes, condiciones fisiológicas basales o sistemas de registro llevan a que cada registro tenga distribuciones diferentes.

Para adaptar la etapa de normalización de características a cada sesión de registro en tiempo real, se propone un método simple que consiste en estimar el estado fisiológico basal de cada individuo a partir de un segmento inicial del registro [123]. Con una longitud suficiente T, se espera aproximar los parámetros de la distribución fisiológica de base, en términos de la media y amplitud de cada característica, y de esta forma poder ajustar un clasificador ya entrenado a nuevos usuarios o condiciones de registro. Si $\mathbf{f}(t)$ es el vector de características para el tiempo t de una sesión de registro particular, y $C = [0, T]$ es el intervalo de tiempo definido para la calibración, el vector de características normalizado para esa misma sesión quedará definido como

$$\hat{f}_i(t) = \frac{f_i(t) - \mu_{f_i(t|t\in C)}}{\sigma_{f_i(t|t\in C)}} \tag{3.1}$$

siendo μ un estimador central y σ un estimador de dispersión. Se eligió la mediana y la distancia intercuartil por la robustez a valores extremos. De esta forma, se busca reducir la variabilidad entre sesiones respetando las restricciones que tienen las aplicaciones en tiempo real.

3.3 Reconocimiento y modelado de emociones

Se evaluaron diferentes métodos para el reconocimiento y modelado de estados afectivos. Como estudio preliminar, se entrenaron clasificadores conocidos, como el MLP, RF y SVM [123, 74]. Luego se propusieron dos métodos novedosos, que

se describen a continuación: un modelo auto-organizativo, que permite modelar la similaridad entre las características fisiológicas y las emociones, y otro basado en ELMs, una familia emergente de redes neuronales artificiales.

3.3.1 Mapas auto-organizativos supervisados

Los mapas auto-organizativos supervisados (sSOM) son un nuevo método motivado en la capacidad de los SOMs para encontrar relaciones en los datos y representarlas en una forma compacta [124]. Para entrenar el modelo, las características fisiológicas y las etiquetas emocionales se concatenan en un vector de características extendido. Los principios que organizan topológicamente el mapa permiten realizar el agrupamiento de los datos multidimensionales, considerando en un mismo espacio tanto a las características como las etiquetas. El entrenamiento, que es no supervisado, es robusto a los valores extremos y datos faltantes, y pueden combinarse tanto etiquetas de emociones categóricas como dimensionales. Una vez entrenado, las predicciones se obtienen identificando la neurona ganadora a partir de las características fisiológicas y luego devolviendo las etiquetas aprendidas por esta neurona.

Dado un conjunto de características fisiológicas $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$, con $\mathbf{x}_n \in \mathbb{R}^F, n = 1, \ldots, N$ y de etiquetas $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$, con $\mathbf{y}_n \in \mathbb{R}^P$, el conjunto de entrenamiento es la matriz $\check{\mathbf{X}} = [\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_N]$, que resulta de concatenar los vectores de características y etiquetas en $\check{\mathbf{x}}_n = [\mathbf{x}_n, \lambda \mathbf{y}_n] \in \mathbb{R}^{F+P}$. El factor de escala λ se incorpora para modificar la influencia de las etiquetas en la topología del mapa.

El sSOM está compuesto por un arreglo rectangular de unidades s_j. Cada unidad genera una respuesta a las entradas $\check{\mathbf{x}}_n$ dada por

$$h_j = ||\check{\mathbf{x}}_n, \check{\mathbf{w}}_j||_2, \tag{3.2}$$

donde $\check{\mathbf{w}}_j = [\mathbf{w}_j^x, \mathbf{w}_j^y] \in \mathbb{R}^{F+P}$ es el vector de pesos sinápticos, que tiene dos componentes: una correspondiente a las características $\mathbf{w}_j^x$ y otra a las etiquetas $\mathbf{w}_j^y$. Los pesos se inicializan con un método basado en PCA, maximizando la varianza inicial de los datos en el mapa, a lo largo de los ejes principales.

El entrenamiento del sSOM es iterativo. Para cada tiempo t, un ejemplo $\check{\mathbf{x}}(t)$ se presenta al mapa y se determina la neurona con mayor similaridad (neurona ganadora), mediante

$$s^*(t) = \arg\min_j ||\check{\mathbf{x}}(t) - \check{\mathbf{w}}_j||_2. \tag{3.3}$$

El método de aprendizaje no supervisado premia a s^* ajustando $\check{\mathbf{w}}_{s^*}$ para aumentar la similitud al ejemplo, y derramando la recompensa a las unidades vecinas, de esta forma induciendo un ordenamiento topológico en el mapa. Los pesos se ajustan por

$$\check{\mathbf{w}}_j(t+1) = \begin{cases} \check{\mathbf{w}}_j(t) + \alpha[\check{\mathbf{x}}(t) - \check{\mathbf{w}}_j(t)], & \text{si } s_j \in \mathcal{N}_{s^*} \\ \check{\mathbf{w}}_j(t), & \text{si } s_j \notin \mathcal{N}_{s^*} \end{cases}, \tag{3.4}$$

donde $0 < \alpha < 1$ es el factor de aprendizaje y $\mathcal{N}_{s^*}$ es la función de vecindad de s^*. El radio de $\mathcal{N}_{s^*}$ y el valor α toman valores grandes al inicio, para hacer un ordenamiento topológico grueso del mapa. Con el paso de las iteraciones estos valores disminuyen, favoreciendo con la recompensa a una pequeña cantidad de neuronas y de esta forma se hace un ajuste fino, conservando las zonas topológicas ya definidas en el mapa.

Una vez finalizado el entrenamiento, entradas similares llevaran a neuronas ganadoras cercanas. Como la entrada al entrenamiento estaba compuesta por características fisiológicas y etiquetas, las regiones del sSOM modelan ambos espacios en conjunto. Usando sólo la entrada fisiológica, con los pesos $\mathbf{w}_j^x$, la neurona ganadora se obtiene con

$$s^* = \arg\min_j ||\mathbf{x} - \mathbf{w}_j^x||_2, \tag{3.5}$$

y la salida estimada será

$$\tilde{\mathbf{Y}}_{s^*} = \frac{\mathbf{w}_{s^*}^y}{\lambda}. \tag{3.6}$$

Para mejorar las estimaciones, que pueden ser ruidosas dependiendo del tamaño del mapa y los datos, se propone un método de suavizado en función de una lista de neuronas ganadoras. Dado $\mathbf{x}$, se determinan las K neuronas más cercanas $[s_1^*, \ldots, s_K^*]$ y la salida es un promedio pesado dado por

$$\bar{\mathbf{Y}} = \sum_{k=1}^{K} \gamma_k \tilde{\mathbf{Y}}_{s_k^*}, \tag{3.7}$$

donde

$$\gamma_k = \frac{||\mathbf{x} - \mathbf{w}_{s_k^*}^x||_2^{-1}}{\sum_{j=1}^{K} ||\mathbf{x} - \mathbf{w}_{s_j^*}^x||_2^{-1}} \tag{3.8}$$

es la distancia normalizada inversa de s_k^* en el espacio de las características. De esta forma, las neuronas ganadoras más cercanas, según las características fisiológicas, reciben un peso mayor en el promedio.

En los experimentos se optimizaron los hiperparámetros más relevantes del método. La cantidad de neuronas y la relación entre el alto y ancho de la red definen la complejidad del modelo. El factor λ se evaluó en el intervalo $[1, 5]$ para dar mayor o menor peso a las etiquetas respecto a las características. Finalmente, se ajusta la cantidad de neuronas ganadoras que se consideran para definir la salida (Ec. 3.7). Se pueden encontrar detalles adicionales en la Sección B.4.1.

Una vez entrenado, el sSOM permite explorar relaciones subyacentes en los datos, a diferencia de otros clasificadores de tipo *caja negra*. El descubrimiento de nuevas relaciones estará guiado por los datos, y luego podrá ser interpretado por un experto.

3.3.2 Máquinas de aprendizaje extremo

Las ELMs son una familia de clasificadores que ha demostrado tener buen rendimiento en diferentes aplicaciones [125]. Se utilizarán dos métodos ELM, cada uno motivado en diferentes ventajas. El primero, la máquina de aprendizaje extremo neuronal (nELM) se define como una red neuronal cuya capa oculta es generada aleatoriamente. La capa oculta realiza una proyección del espacio de las características en un nuevo espacio, de mayor dimensión, donde se espera que el problema sea linealmente separable. De esta forma, se evita la mayor complejidad computacional del entrenamiento, ya que sólo se entrena la capa de salida.

A partir de un conjunto de N vectores de características, la proyección en la capa oculta, definida por pesos aleatorios, se puede expresar con la matriz $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^T$. Ésta será la entrada a la capa de decisión, definida por los pesos $\mathbf{W}$. La salida de la red se puede escribir como

$$\tilde{\mathbf{Y}} = \mathbf{HW}. \tag{3.9}$$

Se ha demostrado que estas redes pueden aproximar cualquier función no lineal si disponen de la suficiente cantidad de neuronas en la capa oculta y la función no lineal de las neuronas cumplen ciertas condiciones [126]. Para encontrar $\mathbf{W}$, se puede definir un problema de optimización de mínimos cuadrados como

$$\underset{\mathbf{W}}{\text{minimizar}} \quad ||\mathbf{HW} - \mathbf{Y}||_2. \tag{3.10}$$

La solución con norma mínima, más robusta al ruido en las características, está dada por

$$\hat{\mathbf{W}} = \mathbf{H}^{\dagger}\mathbf{Y}, \tag{3.11}$$

donde $\mathbf{H}^{\dagger}$ es la seudo-inversa de Moore-Penrose [127]. Esta solución es extremadamente rápida cuando se la compara con otros algoritmos de entrenamiento de redes neuronales artificiales y otros métodos como las SVMs. En este método, el único hiperparámetro es la cantidad de neuronas en la capa oculta.

El segundo método es una variante que utiliza una función de núcleo para representar la proyección de características (kELM), de una forma similar a las SVM. Se introduce una regularización en el problema de optimización

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimizar}} \quad & \frac{1}{2}||\mathbf{W}||_2 + C\frac{1}{2}\sum_{n=1}^{N}||\epsilon_n||_2 \\ \text{sujeto a} \quad & \mathbf{h}_n\mathbf{W} = \mathbf{y}_n^T - \epsilon_n^T, \end{aligned} \tag{3.12}$$

donde ϵ_n es el error de entrenamiento y C el factor de regularización. La solución está dada por

$$\hat{\mathbf{W}} = \mathbf{H}^T\left(\frac{1}{C}\mathbf{I} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{Y}. \tag{3.13}$$

Reemplazando (3.13) en (3.9), se obtiene

$$\tilde{\mathbf{Y}} = \mathbf{H}\hat{\mathbf{W}} = \mathbf{H}\breve{\mathbf{H}}^T \left(\frac{1}{C}\mathbf{I} + \breve{\mathbf{H}}\breve{\mathbf{H}}^T\right)^{-1} \breve{\mathbf{Y}}, \tag{3.14}$$

donde $\breve{\mathbf{H}}$ y $\breve{\mathbf{Y}}$ son las matrices de proyección y etiquetas de los datos de entrenamiento. Seleccionando una función de núcleo $\mathcal{K} : (\mathbb{R}^F, \mathbb{R}^F) \to \mathbb{R}$, la matriz para los dos conjuntos de características $(\mathbf{X}, \mathbf{X}')$ esta dada por

$$\mathbf{\Omega}(\mathbf{X}, \mathbf{X}') = \mathbf{H}\mathbf{H}'^T : \Omega_{i,j} = \mathbf{h}_i \cdot \mathbf{h}_j = \mathcal{K}(\mathbf{x}_i, \mathbf{x}'_j). \tag{3.15}$$

Luego, (3.14) se puede escribir como

$$\tilde{\mathbf{Y}} = \mathbf{\Omega}(\mathbf{X}, \breve{\mathbf{X}}) \left(\frac{1}{C}\mathbf{I} + \breve{\mathbf{\Omega}}\right)^{-1} \breve{\mathbf{Y}}, \tag{3.16}$$

donde $\breve{\mathbf{\Omega}}$ es la matriz núcleo de entrenamiento. De (3.16), se puede ver que la función de núcleo reemplaza a las matrices de proyección. Este método es más costoso a nivel computacional que el nELM, pero se reporta con mejores capacidades para modelar relaciones no lineales en los datos. Estos métodos pueden utilizarse tanto para clasificación como para regresión, adaptándose a los diferentes escenarios de reconocimiento de emociones. El desarrollo detallado de estos métodos se encuentra en la Sección B.4.2. La función de núcleo utilizada es la de base radial, y se optimiza para el coeficiente exponencial γ y el factor de regularización C.

3.4 Diseño experimental

3.4.1 Materiales

Los métodos propuestos se evaluaron sobre diferentes corpus, con desafíos y limitaciones que permiten explorar diversos aspectos del reconocimiento de emociones. En una evaluación preliminar de características, clasificadores y métodos de normalización (Anexo A) se usó el conjunto de 8 emociones de Picard y col. [17]. Éste consta de 20 registros de un participante en diferentes días, lo que permite estudiar la capacidad de adaptación de los modelos a nuevos contextos. En cada registro, el participante expresa una serie de 8 emociones, seguidas una de la otra, por un lapso de 20 minutos. Se utilizó el registro continuo de PPG para caracterizar emociones categóricas.

Para el reconocimiento de emociones categóricas también se utilizó el corpus de Monkaresi y col. (Sección B.5.1) [128]. Este corpus sigue los lineamientos de los registros tradicionales en computación afectiva. Consiste en 16 sesiones de diferentes participantes, donde las emociones fueron inducidas utilizando el IAPS y las anotaciones son registros de auto-percepción. En cada sesión se presentaron

aproximadamente 75 imágenes, a intervalos de 10 s, mientras se registra la señal de ECG y el participante reporta sus emociones en la escala SAM. Luego, el registro es segmentado para cada imagen y las anotaciones se binarizan en clases negativa y positiva para valencia, y alta y baja para la activación.

La sección más relevante los experimentos se realizó con el corpus RECOLA de Ringeval y col. [10], que consiste en registros de interacciones diádicas entre participantes mientras discuten como resolver un problema en teleconferencia. Los primeros 5 minutos de interacción fueron anotados por 6 evaluadores externos, según como ellos perciben los afectos expresados en el plano AV. Las anotaciones para cada instante se promedian según la concordancia media entre los evaluadores [100], obteniéndose de esta forma una etiqueta cuadro a cuadro para activación y otra para valencia, lo que permite evaluar el reconocimiento de afectos continuos (Sección B.5.2). De los 27 participantes que tienen registros fisiológicos, 18 fueron liberados públicamente y 9 se reservaron para tareas de validación ciega.

Finalmente, se utilizó el corpus MAHNOB de Soleymani y col. [5] para evaluar cualitativamente las representaciones del sSOM sobre diferentes modelos emocionales. Este corpus consta de registros multimodales para emociones espontáneas en respuesta a fragmentos de películas. Se registraron 27 participantes, los cuales etiquetaron cada estímulo en categorías (9 clases) y en modelos dimensionales (activación, valencia, dominancia y predecibilidad). Además, se dispone de las etiquetas de referencia de los estímulos (7 clases) a partir de encuestas en un conjunto diferente de participantes.

3.4.2 Validación

Los métodos fueron validados con señales reales, en interacciones naturales, y utilizando los corpus de acceso libre mencionados anteriormente. Se utilizó un esquema de validación cruzada anidada, contemplando siempre tres particiones: una para entrenamiento, otra para la optimización de hiperparámetros del modelo, y otra para validación. De esta forma, se evita el sobreajuste en la selección de hiperparámetros y se reduce el sesgo del error reportado en validación. Para que el método de validación sea realista, las particiones se definieron por agrupación de participantes, de forma tal que los registros de los participantes que se utilizan para el entrenamiento no se utilizan para la validación. El número de particiones fue seleccionado previamente, tomando entre 3 y 5 particiones según la cantidad de participantes y datos disponibles.

Para verificar si la diferencia de rendimiento de los clasificadores es estadísticamente significativa, se utilizaron test no paramétricos, debido a que son relativamente pocas sesiones. Para comparar más de dos clasificadores a la vez se utilizó el test de Friedman, y de resultar diferencias significativas, se utilizó el test de rangos con signo de Wilcoxon como método post-hoc, para encontrar que pares de clasificadores son los que reportan diferencias significativas.

Para realizar comparaciones justas con otros métodos del estado del arte que no se hayan podido reprooducir (la mayoría no presentan código fuente ni parámetros completamente detallados), los métodos propuestos se evaluaron también con los mismos datos y particiones que los métodos del estado del arte. Esto es particularmente relevante sobre la partición de validación de RECOLA, ya que esta partición es ciega (no se disponen de las etiquetas) y por tanto no es posible sobreajustarse a esta.

3.4.3 Medidas de desempeño

Se utilizaron medidas acordes a cada tarea para cuantificar el rendimiento. Para los experimentos de clasificación se utilizaron dos medidas. El coeficiente Kappa de Cohen [129], definido como

$$\kappa = \frac{A_0 - A_c}{1 - A_c}, \tag{3.17}$$

donde A_0 es la exactitud de clasificación y A_c la probabilidad de acierto por azar dada la distribución de las clases. Esta medida considera un nivel de base para la clasificación, siendo $\kappa = 0$ cuando no hay evidencia de que el clasificador funcione mejor que una predicción aleatoria, y $\kappa = 1$ para una clasificación perfecta. La segunda medida es la UAR

$$\text{UAR} = \frac{\sum_i^{n_c} S_i}{n_c}, \tag{3.18}$$

donde n_c el número de clases y $S_i = \text{VP}/(\text{VP} + \text{FN})$ es la sensibilidad definida a partir de los verdaderos positivos (VP) y falsos negativos (FN) para la clase i. Ambos coeficientes son más robustos al desequilibrio de clase que la simple exactitud, ya que consideran una ponderación de la tasa de acierto observada respecto a la distribución de clases en los datos.

Para la estimación de afectos dimensionales se utilizó el coeficiente de concordancia de Lin ρ_c [130], definido como

$$\rho_c(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\rho\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}^2 + \sigma_{\hat{\mathbf{y}}}^2 + (\mu_{\mathbf{y}} - \mu_{\hat{\mathbf{y}}})^2}, \tag{3.19}$$

donde $\mathbf{y}$ es el vector de etiquetas de referencia para cada instante de un registro completo, $\hat{\mathbf{y}}$ el vector de etiquetas estimado, ρ es el coeficiente de correlación de Pearson, μ_y es la media y σ_y es la desviación estándar. El rango de ρ_c es [-1,1], tomando valores alrededor de 0 si no hay evidencia de concordancia entre las etiquetas, y 1 para una concordancia perfecta. Este coeficiente es una mejora de ρ, ya que considera la correlación de las etiquetas en el tiempo junto con el error

cuadrático medio.

4 Resultados

En esta sección se presentan y discuten los principales resultados que soportan la hipótesis y objetivos propuestos para esta tesis. Las tablas, figuras y discusiones adicionales se encuentran en la sección de Anexos.

4.1 Adaptación al usuario en tiempo real

La variabilidad fisiológica entre participantes y entre sesiones de registro es un factor importante para el reconocimiento de emociones. Esto se evidencia claramente cuando los resultados del reconocimiento con la normalización estándar, sólo a partir de los datos de entrenamiento, es notablemente superada por los resultados de normalizar cada sesión de forma independiente [74]. Sin embargo, esta normalización en tiempo diferido no podría emplearse en aplicaciones interactivas.

El método de adaptación propuesto para este problema (Sección 3.2.2) se evaluó en dos corpus, con diferentes modelos de inducción y representación de emociones. En primer lugar se lo hizo sobre una sucesión continua de emociones categóricas [17], donde la primera es siempre un estado neutral. De este estado inicial se tomó una parte para estimar los parámetros de normalización, definiendo un segmento de calibración de hasta 30 s. Normalizando el registro restante a partir de estos parámetros, se evaluaron las tasas de reconocimiento para diferentes longitudes de registro: 3 min (2 clases de emociones), 6 min (3 clases), 15 min (5 clases) y 25 min (8 clases). En Figura 4.1 se presentan los resultados comparando el efecto de la adaptación a cada sesión de registro, donde se encuentra que el método propuesto mejora el funcionamiento de los clasificadores para emociones prototípicas, sobre todo para registros de corta duración.

Una pregunta que surge de estos resultados es si el estado neutral durante el período de adaptación es necesario, ya que en la mayoría de los casos el usuario podría comenzar con cualquier otro estado basal. Además, no es claro que longitud es la óptima para el segmento de adaptación. Para superar las limitaciones del estudio anterior, se evaluó el método de adaptación propuesto en el corpus RECOLA, que contiene registros de afectos espontáneos en interacciones realistas, a diferencia de los resultados reportados en [123]. Los participantes además comienzan con diferentes estados basales, con anotaciones dispersas en el plano AV [10].

Siguiendo la metodología experimental de [74], con las características de la HRV y los clasificadores sSOM y nELM, se toman los primeros N segundos de cada registro para estimar la media y varianza de cada característica. Para hacer comparaciones justas entre los métodos de normalización, este segmento se descarta antes de entrenar el modelo para todos los casos. Se realizaron evaluaciones

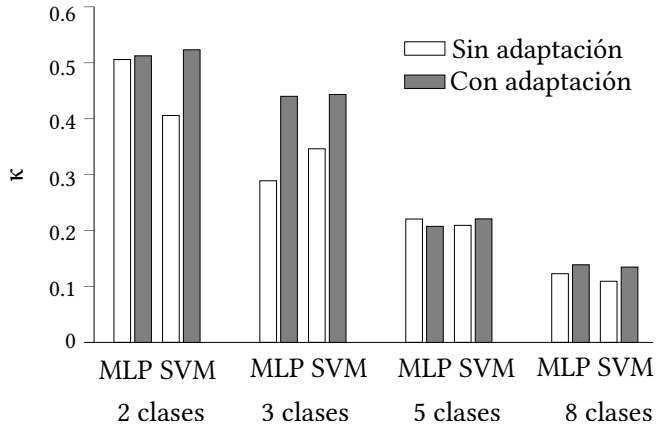


Figura 4.1: Resultados de la validación cruzada (κ promedio) con el método de adaptación en reconocimiento de emociones categóricas. Se comparan registros de 2 a 25 minutos, conteniendo 2, 3, 5 y 8 emociones diferentes, con los clasificadores MLP y SVM.

para los tres tipos de normalización (por sesión, por adaptación y la normalización estándar) variando $N \in \{20, 40, 60, 120, 180\}$ s.

Se puede observar en la Figura 4.2 que el método propuesto arroja resultados intermedios entre la normalización estándar y la normalización por sesión. Para segmentos muy cortos (20 s), las estimaciones no llegan a modelar la distribución correspondiente a cada registro y los resultados son en general inferiores a los otros métodos. Sin embargo, se obtienen mejoras significativas para períodos de adaptación entre 40 y 60 s, lo que indicaría que los parámetros generales del estado basal fisiológico pueden estimarse con sólo un minuto de adaptación de los modelos en cada sesión de registro, y obtener resultados prácticamente equivalentes a los del reconocimiento en tiempo diferido.

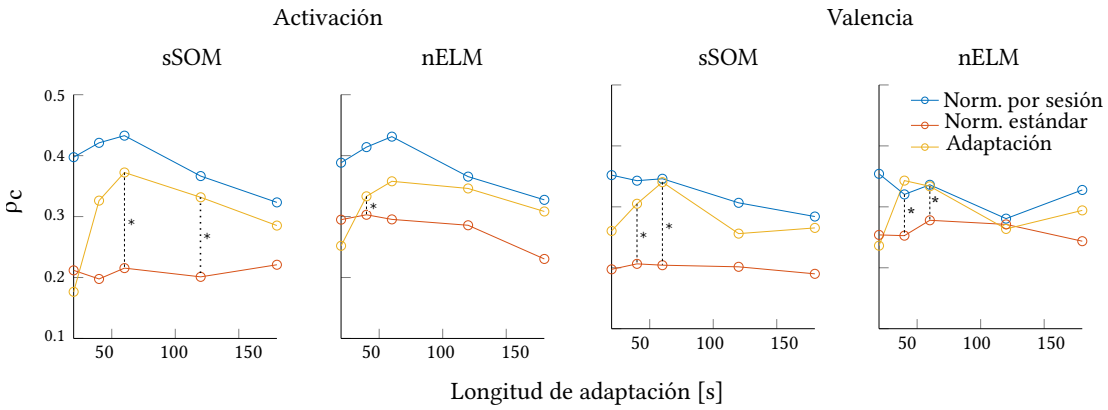


Figura 4.2: Comparación de métodos de normalización de características. Se muestran los resultados del reconocimiento de activación y valencia (ρ_c promedio) con sSOM y nELM. Los métodos de normalización estándar, por sesión y por adaptación se evalúan para diferentes períodos de adaptación en cada registro. Las diferencias significativas se indican con (*).

4.2 Representación de respuestas fisiológicas y modelos afectivos

La investigación en métodos guiados por los datos que llevó al desarrollo del sSOM está motivada en la búsqueda y representación de relaciones subyacentes entre las características fisiológicas y los modelos afectivos. El ordenamiento topológico del mapa, una vez entrenado, permite visualizar el espacio n-dimensional de las características y etiquetas en imágenes simples. En la Figura 4.3 se muestra un subconjunto de dos características de la HRV y las etiquetas de activación y valencia. Cada imagen es una representación de los pesos del mapa para una característica, aprendidos durante el entrenamiento. Esto permite ver por ejemplo como el cambio de la relación entre las componentes de baja y alta frecuencia de la variabilidad cardíaca (LF/HF') está relacionado a los cambios de activación, mientras que el ritmo cardíaco medio (μ_{HR}) varía con los de valencia. Este tipo de análisis provee una herramienta para la búsqueda de relaciones importantes en los datos, que puede ser utilizado para encontrar nuevo conocimiento, y es una ventaja del sSOM respecto otros clasificadores de tipo caja negra.

El sSOM se evaluó también como un método novedoso para representar relaciones entre diferentes espacios emocionales. El método propuesto permite incorporar variables numéricas y categóricas en la misma representación de manera directa. Para evaluar este aspecto, se utilizaron las etiquetas disponibles en el corpus MAHNOB, que se pueden separar en tres grupos: etiquetas dimensionales y etiquetas categóricas, ambas reportadas por los participantes según su percepción, y las etiquetas categóricas de los estímulos, generadas a partir de un grupo disjunto de participantes [5]. Entrenando el sSOM con estos datos, se obtuvieron las representaciones de la Figura 4.4, donde se separan las componentes de los

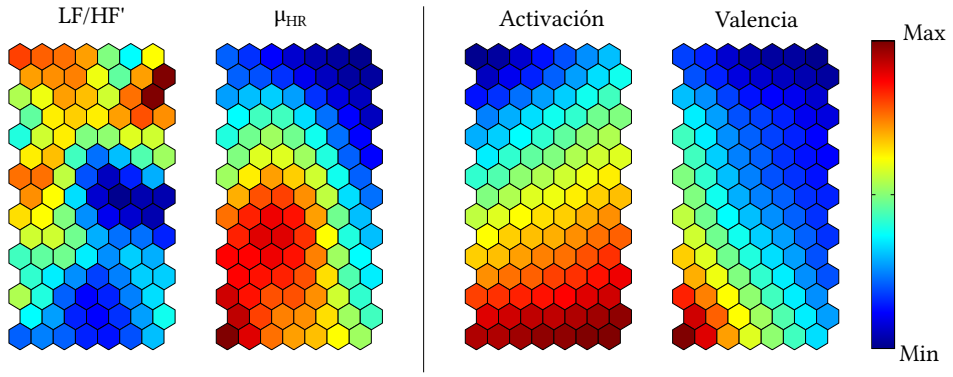


Figura 4.3: Representación de características fisiológicas y etiquetas emocionales basada en sSOM. A la izquierda, dos características de la HRV: la derivada de la relación entre energía de bandas de baja y alta frecuencia (LF/HF') y el valor medio del HR (μ_{HR}). A la derecha, las componentes relacionadas a las etiquetas: activación y valencia.

tres grupos de etiquetas. Del análisis de los agrupamientos generados se pueden inferir rápidamente algunas particularidades del corpus. Por ejemplo, el estímulo de *diversión* se percibe en categorías similares, como *alegría/felicidad*, pero también refleja respuestas un tanto diferentes a lo esperado, donde se puede ver que parte del agrupamiento se interpreta como *disgusto* y otra parte como algo *neutral*. También se observa que el conjunto de estímulos catalogados como *miedo* se percibe en diferentes categorías relacionadas: *miedo*, *ansiedad* y *sorpresa*. Comparando luego con los planos de activación y valencia, se pueden referenciar las clases categóricas al plano dimensional. Más aún, se pueden caracterizar diferentes subgrupos dentro de las emociones categóricas, por ejemplo el agrupamiento de *entretenimiento* se percibe con diferentes valores de activación, tomando desde valores medios a valores altos, indicando diferentes intensidades dentro de la misma categoría. Como se ilustra para esta aplicación, el sSOM podría dar nuevos detalles sobre la efectividad de los estímulos en términos de como se configuran los grupos de emociones en los participantes respecto a las etiquetas que se espera que manifiesten.

4.3 Reconocimiento de estados afectivos

Los métodos se evaluaron para el reconocimiento de afectos en clasificación y regresión. En cuanto al reconocimiento de clases categóricas, se utilizaron los datos registrados por Monkaresi y col. [128]. La Tabla 4.1 muestra los resultados de clasificación de sSOM, nELM y kELM para niveles binarios de activación y valencia, en comparación con clasificadores de base (SVM y RF) y resultados del estado del arte [128]. Como se puede ver en la tabla, se encontró que tanto el sSOM como las ELMs dieron buenos resultados, superando los del estado del arte en las mismas condiciones experimentales.

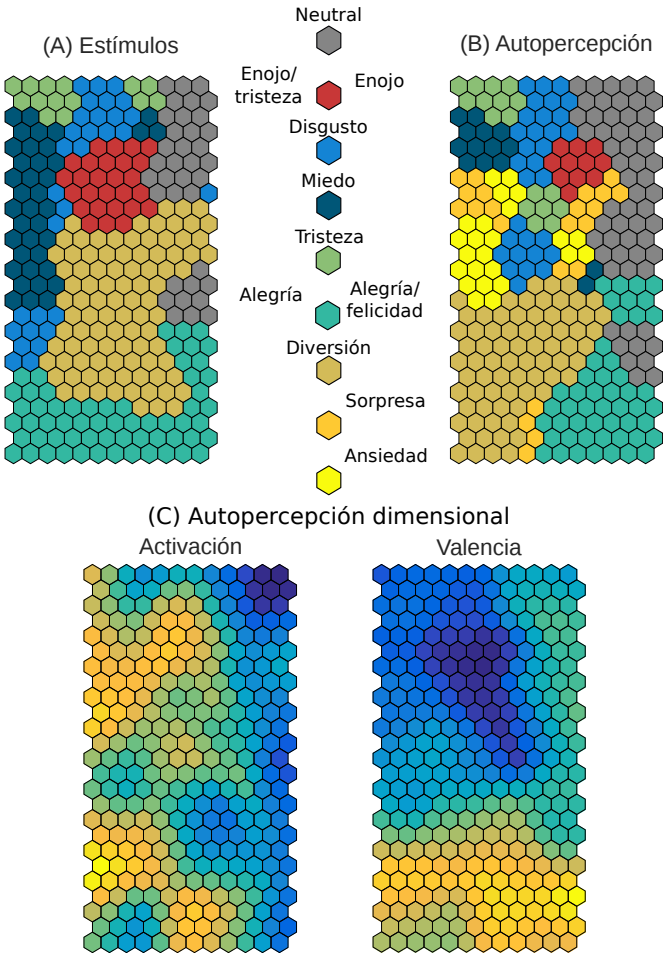


Figura 4.4: Representación de modelos afectivos con sSOM a partir de un corpus de películas [5]. Las películas utilizadas como estímulos están definidas en un conjunto de 7 clases (A). Los reportes de los participantes, según su propia percepción, están definidos en un conjunto de 9 clases (B) y en términos de activación y valencia (C). Cada categoría se representa en un color; la mayoría de las clases aparece en ambos conjuntos (estímulo y percepción).

Clasificador	Activación κ	Activación UAR	Valencia κ	Valencia UAR
Ensamble [128]	.071	-	.191	-
SVM	.143	.570	**.213**	**.607**
RF	.109	.558	.163	.582
sSOM	.137	.566	.202	.597
nELM	.068	.541	.119	.559
kELM	**.148**	**.576**	.208	.603

Tabla 4.1: Resultados de validación cruzada (κ y UAR promedio) para clasificación binaria de afectos con el corpus de Monkaresi y col. [128]

En cuanto al reconocimiento de afectos dimensionales, se realizaron una serie de experimentos con el corpus RECOLA. En la Tabla 4.2 se presentan los resultados en validación cruzada (ρ_c medio) para los métodos propuestos, utilizando los clasificadores SVR y RF como referencia. Se presentan primero los resultados utilizando la normalización por sesión, donde el kELM resultó el mejor método para activación y nELM para valencia. Es interesante destacar que kELM muestra resultados altos con poca varianza, indicando estabilidad y capacidad de generalización entre sesiones de prueba. Por otro lado, nELM es extremadamente rápido para entrenar y la búsqueda de hiperparámetros es muy simple, requiriendo sólo definir el número de neuronas de la capa oculta.

Como se discutió en la Sección 4.1, una tarea más desafiante es el reconocimiento de afectos en tiempo real, en donde la normalización de características no es posible como en tiempo diferido. Usando la normalización estándar, a partir de los parámetros de normalización del conjunto de entrenamiento, se obtuvieron los resultados de la segunda parte de la Tabla 4.2. En este caso, con kELM se obtienen los mejores resultados, con diferencias significativas respecto a los demás clasificadores. Esto sugiere que el método es más robusto a la variabilidad de las características entre registros. Estos resultados confirman un mejor desempeño que otros modelos reportados con el mismo método de normalización [131], sobre todo para la predicción de activación.

Se realizaron pruebas para evaluar la relevancia de la señal de EDA, que es otra de las fuentes de respuesta fisiológica interesantes para el reconocimiento con los modelos propuestos. La Tabla 4.3 muestra comparativamente los resultados con el mismo diseño experimental, para HRV y EDA. Los resultados muestran que el reconocimiento de valencia en EDA es más efectivo que el de activación, coincidiendo con otros reportes [110]. La comparación respecto a HRV indica que esta última resulta mucho más descriptiva del nivel de activación y valencia que la señal de EDA, según las características evaluadas. Esto también condice con la bibliografía, donde esta diferencia de rendimiento es similar [132, 110, 111, 118], dejando a la HRV como la señal fisiológica de preferencia para el reconocimiento

	Clasificador	Activación	Valencia
Normalización por sesión	SVR	.378 (.030)	.243 (.064)
	RF	.369 (.018)	.282 (.028)
	sSOM	.362 (.032)	.313 (.059)
	nELM	.366 (.039)	**.322** (.054)
	kELM	**.388** (.009)	.320 (.049)
Normalización estándar	SVR	.155 (.019)	.104 (.046)
	RF	.119 (.048)	.126 (.018)
	sSOM	.165 (.051)	.141 (.060)
	nELM	.217 (.025)	**.230** (.034)
	kELM	.**260** (.002)	.223 (.047)

Tabla 4.2: Resultados de validación cruzada (ρ_c promedio y desvío estándar entre paréntesis) utilizando normalización por sesión y normalización estándar en el corpus RECOLA. Se comparan los métodos propuestos y clasificadores de base

	Activación		Valencia	
Clasificador	EDA	HRV	EDA	HRV
sSOM	.123	.362	.187	.313
nELM	.151	.366	.253	**.322**
kELM	**.166**	**.388**	**.257**	.320

Tabla 4.3: Comparación de EDA y HRV para el reconocimiento de afectos dimensionales normalizando cada registro individualmente. Se muestra el ρ_c medio en validación cruzada y desvío estándar entre paréntesis.

de afectos.

En la Tabla 4.4 se presentan los resultados obtenidos sobre la partición de *optimización* de RECOLA, donde se comparan los métodos propuestos con una gran lista del estado del arte, que usan sólo la señal de ECG o la HRV para el reconocimiento. Para evitar sobreajustar el modelo a esta partición, se utilizó el mejor conjunto de hiperparámetros de experimentos anteriores. En la tabla se listan métodos como LSTM [133, 132, 134], SVR [112, 110, 111] y un modelo DNN *end-to-end* [96].

Como evaluación final, una vez determinados los mejores hiperparámetros por validación cruzada, los modelos propuestos se utilizaron por única vez en la partición de validación de RECOLA, que consiste en 9 sesiones de registro con etiquetas ocultas. En la Tabla 4.5 se comparan los resultados de los métodos propuestos con los del estado del arte, reportados en la misma partición. En estos resultados se puede ver que los métodos propuestos superan las otras contribuciones. Además, la metodología experimental empleada permitió que se preserve la capacidad de generalización de los modelos, evitando el sobreajuste. Esto se

Clasificador	Referencia	Activación	Valencia
Ensamble	Ringeval y col. 2015 [100]	.275	.183
LSTM	Chao y col. 2015 [133]	.222	.182
LSTM	Chen y col. 2015 [132]	.333	.314
DNN-LSTM	He y col. 2015 [117]	.297	.293
DNN	Cardinal y col. 2015 [116]	.262	.124
Ensamble	Kachele y col. 2015 [64]	.344	.256
RVM	Manandhar y col. 2016 [118]	.293	.274
SVR	Weber y col. 2016 [112]	**.468**	.413
PCA + LR	Povolny y col. 2016 [135]	.391	.388
SVR	Valstar y col. 2016 [110]	.379	.293
SVR	Sun y col. 2016 [111]	.392	.264
LSTM	Brady y col. 2016 [134]	.357	.364
End-to-end	Keren y col. 2017 [96]	.426	**.419**
sSOM		**.402**	.354
nELM		.399	**.375**
kELM		.388	.338

Tabla 4.4: Resultados de los métodos propuestos (ρ_c promedio) en comparación a los del estado del arte en la partición de *optimización* de RECOLA.

Clasificador	Referencia	Activación	Valencia
Ensamble	Ringeval y col. 2015 [100]	.192	.139
DNN	Cardinal y col. 2015 [116]	.161	.121
SVR	Valstar y col. 2016 [110]	.334	.198
End-to-end	Keren y col. 2017 [96]	.360	.225
sSOM		.404	.273
nELM		**.421**	**.321**
kELM		.367	.293

Tabla 4.5: Resultados para los métodos propuestos (ρ_c medio) en comparación a los métodos del estado del arte en la partición de validación de RECOLA.

puede observar al comparar los resultados de validación cruzada (Tabla 4.2), y las particiones de optimización (Tabla 4.4) y validación (Tabla 4.5). Este aspecto, en ocasiones ignorado, es muy importante y podría explicar por qué otros trabajos obtienen mucho mejores resultados en la Tabla 4.4 que en la Tabla 4.5.

5 Conclusiones

En esta tesis se ha enfrentado el desafío del reconocimiento de estados afectivos únicamente a partir las respuestas fisiológicas, medidas a través de señales biomédicas. Estas señales tienen diversas ventajas, como la privacidad y la posibilidad de registrarse continuamente, pero presentan desafíos que deben resolverse para llegar a desempeños aceptables para aplicaciones prácticas. En este sentido, se hicieron aportes en diferentes etapas.

Por un lado, el uso de señales fisiológicas está generalmente restringido a la combinación multimodal de fuentes para obtener resultados aceptables, para lo cual es necesario instrumentar gran número de sensores en el usuario. Para reducir la invasividad requerida se hizo el estudio sistemático de diferentes señales biomédicas disponibles, considerando factores como la invasividad, la relación con respuestas afectivas y la tecnología actual de los sensores. De este análisis se encontró que las señales del sistema cardiovascular, ECG y PPG, y la EDA tienen un gran potencial para el reconocimiento de emociones en aplicaciones realistas. Se evaluaron diferentes tipos de características de estas señales, resultando en un conjunto reducido de características que reportaron una buena capacidad discriminativa.

Para mejorar el funcionamiento de los clasificadores para aplicaciones en tiempo real, donde se presentan diferentes usuarios, condiciones de registro y estados fisiológicos basales, se desarrolló un método de adaptación para que un clasificador, una vez entrenado, realice mejores estimaciones en estos nuevos registros. Este método, a partir de 1 min de registro, estima los parámetros de normalización de las características que ingresan al clasificador. Los resultados obtenidos en dos corpus de emociones son alentadores y permiten aproximar los resultados a los de reconocimiento en tiempo diferido.

Se propuso un novedoso modelo auto-organizativo supervisado (sSOM) para mejorar las tasas de reconocimiento y proveer una representación gráfica de las complejas relaciones entre características fisiológicas y espacios emocionales. A diferencia de otros métodos de tipo caja negra, es posible visualizar las relaciones aprendidas por el sSOM en imágenes compactas, haciéndolo un método versátil para diversas aplicaciones. Además, se emplearon dos métodos basados en un nuevo paradigma en redes neuronales: las nELMs, que tienen un muy bajo costo computacional, y las kELMs, que ha demostrado gran capacidad de generalización. Estos métodos fueron evaluados en un corpus de interacciones afectivas espontáneas y realistas, en desafíos que son de interés actualmente para la computación afectiva. En estos experimentos, se lograron superar los resultados del estado del arte, mejorando considerablemente las tasas de reconocimiento alcanzadas a partir de la HRV. Los aportes originales realizados en diferentes etapas del proceso, desde el análisis y visualización de características fisiológicas hasta

el reconocimiento de estados afectivos, contribuyeron a avanzar con el estado del arte para este desafío.

Anexos

El apartado de anexos se organiza de la siguiente manera. En el Anexo A se presenta el artículo publicado por Bugnon, L. A., Calvo, R. A., y Milone, D. H, "A Method for Daily Normalization in Emotion Recognition", *15vo Simposio Argentino de Tecnología, AST 2014, 43 JAIIO*, pp. 48–59, 2014. En este trabajo se propuso un método para adaptar los clasificadores a la variabilidad fisiológica independiente de las emociones, y se lo evaluó en un corpus de emociones categóricas.

En el Anexo B se presenta el artículo de Bugnon, L. A., Calvo, R. A., y Milone, D. H., "Dimensional Affect Recognition from HRV: an Approach Based on Supervised SOM and ELM", *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2017.2763943, enviado en Julio de 2016, revisado en febrero de 2017, en prensa el 17 de octubre de 2017. En este trabajo se desarrollaron métodos para el reconocimiento y representación de afectos dimensionales a partir de la señal de HRV y se validaron en condiciones realistas. Los resultados muestran que el estado del arte para el problema en cuestión fue superado con los métodos propuestos.

El autor de la tesis propuso los métodos originales, los desarrolló, realizó la puesta a punto y la evaluación en ambos trabajos. La discusión de las ideas, el análisis de resultados y redacción de los documentos fueron realizados en colaboración con los co-autores.

Aval de los Directores:

.....	.....
Dr. Diego H. Milone	Dr. Rafael A. Calvo
Director	Co-Director

Appendix A

A method for daily normalization in emotion recognition

Leandro A. Bugnon[1], Rafael A. Calvo[2] and Diego H. Milone[1]

[1] sinc(i) - Research Institute for Signals, Systems and Computational Intelligence, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, CONICET, Santa Fe, Argentina

[2] Department of Electrical Engineering, University of Sydney, Sydney, NSW, 2006, Australia

Abstract Affects carry important information in human communication and decision making, and their use in technology have grown in the past years. Particularly, emotions have a strong effect on physiology, which can be assessed by biomedical signals. This signals have the advantage that can be recorded continuously, but also can become intrusive. The present work introduce an emotion recognition scheme based only in photoplethysmography, aimed to lowerinvasiveness. The feature extraction method was developed for a realistic real-time context. Furthermore, a feature normalization procedure was proposed to reduce the daily variability. For classification, two well-known models were compared. The proposed algorithms were tested on a public database, which consist of 8 emotions expressed continuously by a single subject along different days. Recognition tasks were performed for several number of emotional categories and groupings. Preliminary results shows a promising performance with up to 3 emotion categories. Moreover, the recognition of arousal and emotional events was improved for larger emotion sets.

keywords: Emotion recognition, Daily variability, Photoplethysmography, Biosignal pattern recognition

A.1 Introduction

Emotional states are relevant not only in a social context, but also influence directly on cognitive process and take a role in decision making [1]. For this reason, by including affect in human-computer interfaces, the communication performance may be improved. In the early theoretical developments, several discrete categorizations of emotions have been described, for instance the six basic emotions of Ekman [136]: *joy, anger, fear, boredom, sadness, disgust* and *neutral*. Later, particular affective states were proposed for certain research fields, for example *confusion, boredom* and *flow* in educational applications [137]. Furthermore, different continuous scales of emotions were described, such as the core affect theory of Russell [36], which support a model of basic neurophysiological reactions that one can feel like energized/not-energized (*arousal*) and pleasant/unpleasant (*valence*).

The ground in psychophysiology have demonstrated that the affects experienced by a subject have several implications in his physiology (or vice versa), modifying its behaviour at different levels [55]. In the central nervous system, the cortical and sub-cortical activity present electrical variations that can be measured [21]. More frequently, affect has been assessed through the influence of the autonomous nervous system; the sympathetic and parasympathetic branches control different systems, as blood circulation, respiration patterns and skin glands regulation. Additionally, the effects of emotions over the somatic nervous system is present on voluntary muscle activity and involuntary reflex responses [19]. Even though physiological recordings can be intrusive, those signals can be recorded continuously (unlike voice and facial expressions), are more difficult to mask and provides an alternative source in the case of communicational disorders. As more portable and less invasive biosignals acquisition systems are developed, it become more feasible to use them in real world applications.

Affective computing based on physiological variables have advanced over several applications. For example autism-disorders research [2], learning technologies [8], gamer experience [7], multimedia automatic tagging [5] and anti-stress therapy [3] are recent developments. In addition, major efforts have been done to collect representative data, as the experimental design to elicit and measure emotions is complex. Singularly, the *Eight Emotion Sentic Data*[1] was compiled to study the physiological variations on a single subject, acquiring 4 biosignals: facial electromyogram, photoplethysmography (PPG), electrodermal activity and respiration amplitude, over 20 daily sessions. For each one, the subject tried to pass through 8 emotional states: *neutral, anger, hate, grief, platonic love, romantic love, joy* and *reverence*. This dataset make possible to test how biosignals vary in short and long term for different emotions.

The first analysis on the mentioned dataset was performed with an offline classification scheme [138]. In that approach, the signals were previously seg-

[1]Public access: http://affect.media.mit.edu/share-data.php

mented, obtaining a *recognition rate* (RR) of 46% for the whole emotion set. In their following research, the results were improved taking into account the heart rate and other physiologically relevant features, rising the RR to 81% [139, 17]. In addition, a more recent publication report similar results employing different classifiers [60]. Unfortunately, segmented signals are not readily available in real applications. Furthermore, as segments length is about 3 minutes, an instantaneous emotion estimation cannot be obtained. As a first approach towards an online classifier, a feature extraction method with moving average window was employed on this data, achieving a RR of 48.98% for all the emotion set, using all the 4 signals [138]. In another interesting application, it is required to detect an emotional event (this is, a *wake-up call*) from a neutral state, using an online approach. This has been attempted employing an auto-associative neural network [140, 99], training it with neutral patterns, and using the difference between new samples and the model estimates for classification.

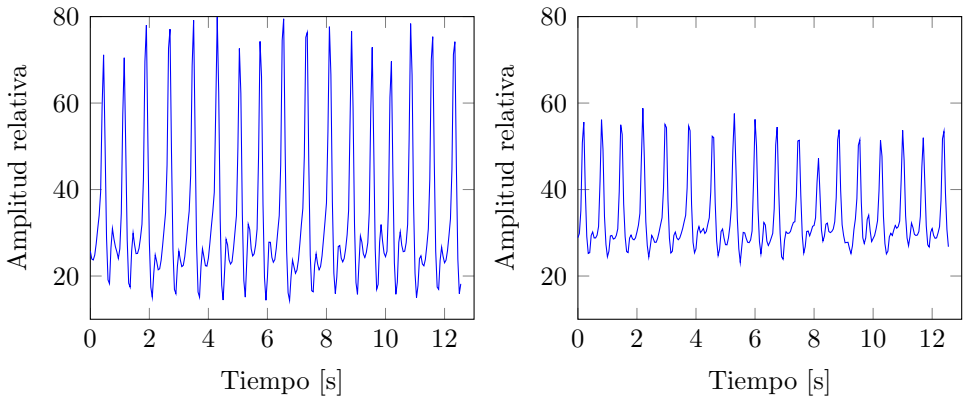
Previous works on affective computing supports the existence of correlation between biosignals and emotions, but the obtained RR still need to be improved significantly. Moreover, it is desirable to minimize the required sensors that allow an acceptable RR. In this direction, the goal of this work is to evaluate the PPG for affect estimation, extracting features related with psychophysiological regulation, and developing a classification scheme oriented to practical online applications.

In the next section, the proposed feature extraction and post-processing methods are discussed, along with the employed classifiers. In Section 3, the experiment design to evaluate the models is presented, followed by relevant results and discussion. Ultimately, interesting conclusions and future work are mentioned in Section 4.

A.2 Feature extraction and classification

Cardiovascular measurements are highly available in public databases, presents low invasiveness and seems to be highly related with several emotional dimensions and categories [19]. Specifically, the heart rate reflect arousal by the influence of sympathetic and parasympathetic branches, in response to approach/withdrawal instincts [55]. One of the sources used to capture circulatory activity is the PPG, which measure the blood flow between an infrared led and a sensor, in particular on a finger (Fig. A.1). While the heart is pumping, the blood flow depicts peaks in a quasi-periodical signal, and the distance between the peaks is found to be highly correlated with the heart rate [68]. Additionally, the pulse amplitude is related to vasoconstriction; when a subject is under stress, the vessels muscle is activated by sympathetic control [141] and the blood flow is reduced (compare Fig.A.1.a and Fig.A.1.b). Taking into account the low invasiveness of finger PPG, it results in a well suited source to estimate emotional information from physiology.

To obtain a heart rate estimate, the signal peaks were detected using a low



(a) Neutral (b) Anger

Figure A.1: Segments of PPG signal on a finger. In (a) a neutral state; and (b) an anger episode, recorded in the same day. This difference is not so evident in a wider scope of the signal. Data was extracted from *Eight-Emotion Sentics Dataset* [17].

pass filter and a windowed automatic threshold. The consequent distance between peaks was interpolated to the signal sample rate (20 Hz) using a piecewise cubic Hermite polynomial, which is less oscillating than spline polynomials for this data. Moreover, the PPG amplitude was estimated using the signal envelope, by interpolation of ascending and descending peaks. Over the obtained peak rate and amplitude, a moving window of width W was displaced with a fixed step of 1 s. For each window step, 4 simple features were extracted: the local mean and standard deviation of these two signals. The feature vector is associated with the central point of the window and the emotion label on that point. On the one hand, small values of W allow to detect short time events, but with a cost of more variance. On the other hand, longer lengths tends to provide a better estimate of mean values.

The regulation of the circulatory system is affected by humoral factors, circadian cycle and subject mood during the data registration, resulting in a significant data variance between days. Previous research accounted different normalization techniques for segmented signals [17], but these approaches cannot be used in online applications. Hereby, a new feature normalization is proposed, which only uses statistical parameters of a neutral segment of length X. It is expected that this neutral sample can approximate the daily baseline, and it can be used to normalize the features in the same day. Furthermore, this approach can be seen as a daily calibration that does not require to acquire many emotion examples everyday. Thus, if $\mathbf{f}(t, d)$ is the feature vector for the time t of day d, and C is the set of the time points used for calibration, the proposed normalization method is applied element by element as:

$$\hat{f}_i(t,d) = \frac{f_i(t,d) - \frac{1}{|C|}\sum_{t \in C} f_i(t,d)}{\max_{t \in C}\{f_i(t,d)\} - \min_{t \in C}\{f_i(t,d)\}} \quad . \tag{A.1}$$

For classification, two well-known models were tested. The MLP is an artificial neural network which structure is composed by forward full-connected perceptrons. Feature inputs feeds a first layer of perceptrons, and their outputs are connected with subsequent layers (hidden layers) until a final output layer, which have as many neurons as classes. This structure can resolve non-linearly separable problems, and were used with various architectures and training algorithms in categorical emotion problems [12, 101, 107] and arousal-valence [105] sets. In this work, the MLP was trained with the standard back-propagation algorithm [142].

SVMs are supervised learning methods which minimize the empirical classification error and maximizes the geometrical margin between the classes in the feature space. This margin maximization provides good generalization properties on high-dimension feature sets and few training samples. Even though SVMs are intrinsically binary classifiers, it can be extended to problems of several classes, for example, by *one-vs-all* method. SVMs have been applied to non-linear classification problems transforming the original feature space into a higher dimensional one, where is presumed that the problem becomes linearly separable. Moreover, it is possible to operate in the transformed feature space computing only the inner products between the projected versions of features pairs by using kernel functions. The Radial Basis Function (RBF) kernels are one of the most popular among them [2, 107, 29] and therefore the Gaussian kernel, a particular RBF, is used in the experiments.

In several datasets for emotion recognition, the classes are imbalanced. This may produce a significant bias in the mentioned classifiers, that advantage the majority class to minimize the overall fitting error. To reduce this effect, the training instances were resampled to equalize the classes occurrences, but leaving the test set as it was. Additionally, it is important to remark that the calibration segments are not used further in the classification task.

A.3 Results and discussion

The experiments in this section were performed using signals from the *Eight Emotion Sentic Data*, particularly the *SetB*, which comprises one subject in sessions properly labeled, over 20 days. The 8 emotions elicited in each session are listed in Table A.1, with the categorical, arousal and valence labels. Each emotional record have a mean length of 3 minutes, composing a continuous signal of around 25 minutes per session.

For each of the following experiments, the performance of the models was estimated using 5-fold cross-validation on the complete features set. Specifically,

Emotion	Arousal	Valence
Neutral	Low	Neutral
Anger	High	Negative
Hate	Low	Negative
Grief	High	Negative
Platonic Love	Low	Positive
Romantic Love	High	Positive
Joy	High	Positive
Veneration	Low	Neutral

Table A.1: Dataset labels in categorical, arousal and valence dimensions. The order of appearance is the same as the register.

the folds were arranged such as each day was contained in a separate fold. In fact, this approach is the most likely in a practical sense, as the model is trained with some days and tested in others. Additionally, in each cross-validation step the parameters of the model were selected according to the best performance using only the 16 training days.

There are several measurements for classifier performance. In this work was used the Cohen's Kappa statistic,

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}, \tag{A.2}$$

which compare the observed accuracy $\Pr(a)$ in relation with the estimated probability of obtaining the same result by-chance $\Pr(e)$, based on the available data in the confusion matrix [143]. The κ statistic takes the value 0 when the evidence of classification accuracy is the same as the by-chance probability, and 1 for the perfect accuracy. This provide a useful measure to compare sets with different number of classes. The algorithms for signal processing, feature extraction and result analysis were developed in Matlab, using the Weka library [144] to implement the classifiers.

A.3.1 Categorical emotion discrimination

In the first place, the proposed methods were tested in the task of recognizing individual emotions, using sets of 2, 3, 5 and 8 emotions. For each group, the proposed feature normalization method was tested, comparing the performance using MLP and SVM classifiers. The best model parameters were selected between two values of W (30 and 60 s), the length of the calibration segments $X \in \{10, 20, 30\}$ s, and basic classifier parameters: the number of hidden neurons {2,4,8,12} for MLP, and the soft-margin coefficient C and kernel exponential $\gamma \in \{2, 4\}$ for SVM. In the case of W, it was found that smaller values than 30

s worsen the results significantly for the current classification scheme, agreeing with other research [138]. For MLP, the best performance was obtained for 2 to 4 neurons in all groups.

Results for categorical emotion recognition are summarized in Figure A.2. In general, all models performed significantly better than the baseline (the by-chance probability). It can be seen that the effect of feature normalization is positive for 2 and 3 emotions. Regarding the classifier model, it was found that MLP and SVM had a similar performance. Comparing with previous research, the present results does not overtake the ones accounted for 8 emotions in an online scheme [138]. However, there is no report of results using only one signal (versus all the four in the dataset), neither on sets smaller than the 8 emotions.

Going further in the evaluation of the normalization method, its effect on several consecutive emotion groupings was tested. It was found that the impact of normalization in RR seems to degrade in relation with the distance between the calibration and the evaluated segments. This explain why normalization does not improve the κ value for bigger set of emotions. Moreover, repeating the experiment with the whole neutral state as calibration segment (and therefore excluding it from the emotion groups) brings a better κ for those groups. This suggest that the size of the calibration segment have a significant effect to estimate a reference value for each day.

A.3.2 Arousal-valence discrimination

Emotions can be mapped in *arousal* (in this case discretized as low and high) and *valence* (negative, neutral and positive), as shown in Table A.1. This mapping allow to group emotions that shares a similar nature, which have a special meaning for different applications. Thereby, it is interesting to evaluate the performance of the PPG and the presented methods over arousal and valence dimensions.

Figure A.3 shows the results of testing again the proposed normalization method and the classifiers for the same emotion sets, but grouped into arousal and valence dimensions. In the first case (Fig. A.3.a), the normalization rise the score for several number of emotions, significantly better than the baseline in all cases. Moreover, the κ value for 3 and 5 emotions, using MLP and normalization, is close to the 2 emotion group, suggesting that the emotions can be effectively grouped in the arousal sets. In the case of valence (Fig. A.3.b), the first 3 emotions have only two classes (neutral and negative), so this should be a similar problem as for arousal. Nevertheless, there is an imbalance towards the negative class in 3 and 5 emotions sets, given by the elicitation ordering. However, the proposed normalization method improves the results significantly for the set of 3 emotions. Finally, the results obtained with the selected methods suggests that arousal discrimination is more feasible than valence with the PPG signal, probably explained by the strong effect of arousal in the circulatory system.

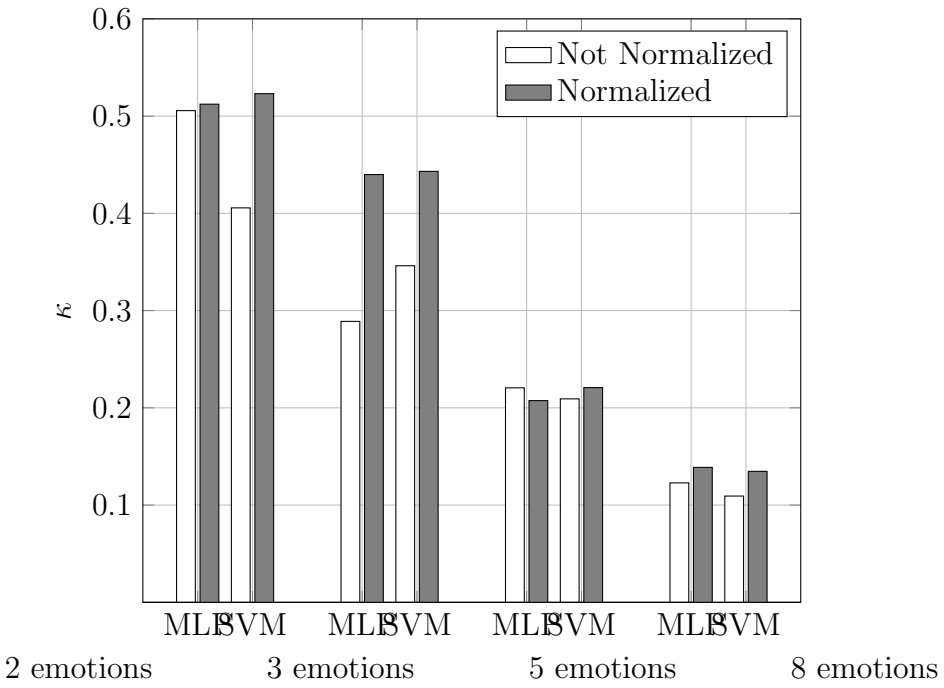
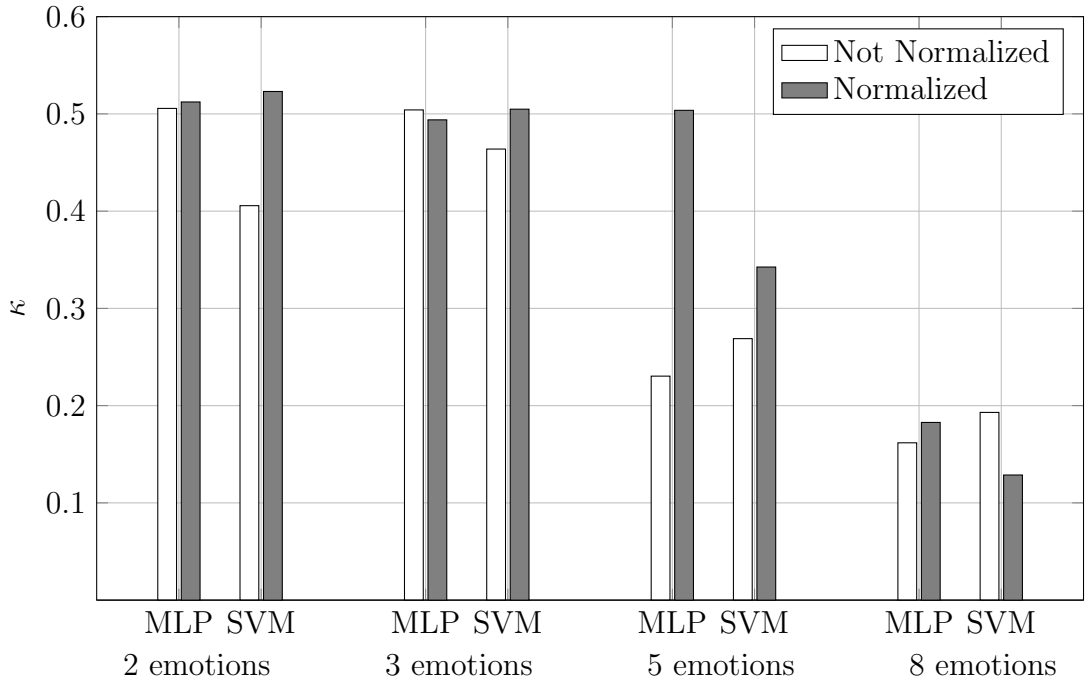
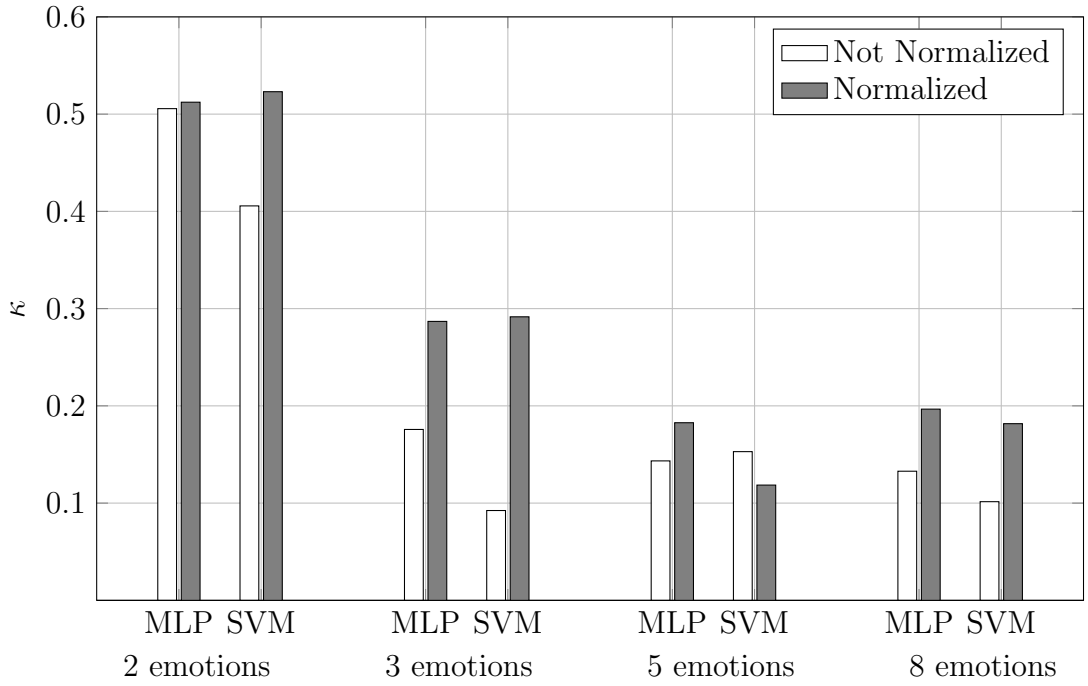


Figure A.2: Mean κ score obtained from cross-validation for individual emotion recognition. Results are shown for emotion sets of various sizes: 2 classes (Neutral-Anger), 3 classes (adding Hate), 5 classes (adding Grief and Platonic Love) and the whole 8 emotions set. The impact of the proposed feature normalization is depicted, along with the comparison of MLP and SVM classifiers.



(a) Arousal



(b) Valence

Figure A.3: Mean κ score from cross-validation over different emotion sets. The normalization method and classifiers are compared with the same sets as the categorical emotions experiment, but labeled in: (a) Arousal and (b) Valence dimensions.

A.3.3 Emotional event detection

The ability to detect an emotional event can be useful, for example, when users have imperative requests. In particular, the dyad neutral-anger is one of the simpler problems, because anger (or rage) has a very active effect in heart rate and vasoconstriction. However, it is not evident if the emotions of different nature, like anger-joy-reverence could be grouped in opposition to the neutral state. To evaluate this, the binary problem of emotion/non-emotion was set by grouping all states but *neutral* as an *emotion*, using a set of the first 5 emotions. Because the dataset is imbalanced for this task (neutral samples are about 1/5 of total, and even lower after removing the calibration segment), two considerations were made. In the first place, the training examples were resampled to equate the occurrences of the classes in training, as used in the valence case. Secondly, as the RR and κ can be misleading in imbalanced data, results were analyzed using the sensitivity (true positive rate) of neutral and emotion classes.

The comparison between the methods shown that, without normalization, MLP achieved a mean sensitivity of 61.7% ($\sigma = 8.2\%$) and SVM scored 58.0% ($\sigma = 7.4\%$). However, the feature normalization rose the mean sensitivity to 73.8% for MLP ($\sigma = 5.8\%$) and 74.0% for SVM ($\sigma = 6.8\%$). Besides the importance of a sufficiently large calibration segment, this factor is trade-off in this task, as higher values overly reduce the neutral instances.

A.4 Conclusions and future work

In this work was proposed a low invasive emotion recognition method using only PPG. A simple method for online feature extraction and a classification scheme were tested in different affective computing tasks, involving several categorical and clustered emotion sets. By taking a small segment of the source signal for calibration, a realistic feature normalization procedure was proposed, which reduce the effect of day variance for moderate periods of time.

It was shown that the employed methods performed significantly better than the baseline in all cases. Moreover, the proposed feature normalization improved the results by reducing daily variability. In particular, the categorical emotion recognition have reached acceptable results for small set sizes. Besides the smaller number of classes, the selected methods obtained a good result for arousal recognition, which is interesting for several applications. The current methods also had promising results in the emotional event detection task, despite the drawbacks of the imbalanced dataset. Additionally, it was found that the length of the calibration segment had an important role to reduce the high variability through different day recordings.

In future research, better feature extraction methods will be addressed, aiming to find more discriminant features. On the other hand, classifiers that properly models the time dynamics of biological signals will be developed. Finally, different

ways to combine the information of several labels, for example categorical and arousal sets, will be pursued.

Anexo A

Appendix B

Dimensional affect recognition from HRV: an approach based on supervised SOM and ELM

Leandro A. Bugnon[1], Rafael A. Calvo[2] and Diego H. Milone[1]

[1] sinc(i) - Research Institute for Signals, Systems and Computational Intelligence, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, CONICET, Santa Fe, Argentina

[2] Department of Electrical Engineering, University of Sydney, Sydney, NSW, 2006, Australia

Abstract Dimensional affect recognition is a challenging topic and current techniques do not yet provide the accuracy necessary for HCI applications. In this work we propose two new methods. The first is a novel self-organizing model that learns from similarity between features and affects. This method produces a graphical representation of the multidimensional data which may assist the expert analysis. The second method uses extreme learning machines, an emerging artificial neural network model. Aiming for minimum intrusiveness, we use only the heart rate variability, which can be recorded using a small set of sensors. The methods were validated with two datasets. The first is composed of 16 sessions with different participants and was used to evaluate the models in a classification task. The second one was the publicly available Remote Collaborative and Affective Interaction (RECOLA) dataset, which was used for dimensional affect estimation. The performance evaluation used the kappa score, unweighted average recall and the concordance correlation coefficient. The concordance coefficient on the RECOLA test partition was 0.421 in arousal and 0.321 in valence. Results shows that our models outperform state-of-the-art models on the same data and provides new ways to analyze affective states.

Physiological measures, affect sensing and analysis, supervised self-organization,

extreme learning machines, dimensional affect estimation.

B.1 Introduction

Affective states, including emotions, moods, and feelings have a key role in the communication and decision-making process of a person. To improve human-computer interactions (HCI) and human-human computer mediated interactions (as in teleconferences), emotions, engagement and even psychological well-being should be taken into account [13].

The first step to improve interactions is the affect recognition, which can be of two types: categorical or continuous [1]. If the target labels are categories, the recognition task is known as classification. For example, the classes can be the basic emotions summarized by Ekman [145] or those more commonly used in HCI [35]. On the other hand, when labels take continuous values, the task is a regression or estimation of those values. In affective computing, this happens when dimensional models with arousal and valence as continuous variables are used [36]. The dimensional model of affect has also been frequently used in a classification context [44, 31]. In those cases, the labels were the result of a quantization over discrete values. For example, by defining the low, medium and high labels for arousal. These approaches will be referred in this work as classification tasks, leaving the term dimensional affect estimation only when the target affects take the original continuous values.

Several works have shown that physiology is correlated with mental states [19], thus it has been used for affect recognition [1]. One advantage of using physiological signals in real world HCI is that signals can be recorded continuously and may be more unconscious (due to the autonomic response) than traditional sources like voice and facial expressions [146]. Moreover, as one feels less noticed during the sensing of physiology, it may be less invasive in terms of privacy [17]. This is interesting in applications where the user does not want (or does not need) to be recorded by a camera or a mic, such as when playing games [6] or selecting a song playlist [3]. Another relevant case is when people have communicational impairments that makes difficult to analyze other sources [2]. Still, the sensing intrusiveness, the recording noise and the natural variations unrelated to emotions are challenging [31, 147]. The challenges have been addressed in studies using multiple physiological signals: electroencephalography (EEG) [108, 148, 149]; respiration patterns (RP) [24]; skin derived signals such as superficial temperature [28] or electrodermal activity (EDA) [54, 150, 151]; pupillary response [29]; and heart related signals such as electrocardiography (ECG) [25, 89] or photoplethysmography (PPG) [26]. Multimodal combinations of different sources have been addressed to improve recognition rates [1, 152, 153, 154]. Also, several efforts have been made to develop multimodal datasets; for example the DEAP dataset [37] combines EEG, PPG, EDA, RP, facial electromyography and skin temperature, along with audiovisual channels to analyze the impact of multimedia content on

users.

The search for a minimally intrusive method is important for real-world applications. In this work, we use Heart Rate Variability (HRV) that has received attention for being related to the autonomic nervous system [147] and basic emotional processes [19]. HRV is the evolution of changes in the beat-to-beat interval over time, and can be acquired using only one ECG lead (for example a chest strap) or it can be estimated from PPG using a specialized wristband. The advances of sensor engineering have lowered the costs and improved precision of HRV wearable devices, enabling to obtain this signal in a natural environment [155]. Studies have shown that the mean HR can be estimated from a smartphone accelerometer [156] and remotely from video [67], widening the possibility of a HCI system to include physiological analysis in their framework.

Current affective computing techniques can be improved in a number of areas. For example, most of the techniques require the use of multiple physiological signals, which necessitates more sensors [157] and intrusiveness [31] to the user. There is evidence that affect classification using a source with low intrusiveness like HRV is feasible [110], yet it is not accurate enough for real-world applications. Furthermore, unlike classification approaches, the dimensional estimation of affects has not been widely explored yet. Currently, improving the classifiers performance with novel approaches is an important challenge that should be addressed. Nevertheless, methods usually focus on performance estimation, but omit analyzing the hidden relations in the data. Novel methods for identification and visualization of the subjacent models of affect and its relation with the inputs should be evaluated.

In this work we approach physiological affect recognition with two different methods. For the first method, we propose a novel algorithm based on supervised self-organizing maps (sSOM) to improve recognition rates and also to provide a graphical representation of the underlying model. This representation can relate the features space with the target in a compact way. Opposed to a black-box, this type of models might allow an expert to find, in the trained model, new relations between physiology and affects. For the second method, we propose the use of extreme learning machines (ELM) [158]. ELM are emergent methods for pattern recognition which have shown improved recognition rates with low computational cost in different applications [125]. They have shown to be faster and more accurate than traditional multi-layer perceptrons and support vector machines (SVM) in several benchmarks [159]. ELM have been selected because of their theoretical capacity of dealing with the features non-linearity, the fast training algorithm and a simplistic computational framework.

Models were evaluated in two different datasets: one for classification and the other for dimensional affect estimation. The first dataset was recorded by Monkaresi et al. [128] and consists of multiple-subject recordings of emotions induced with pictures. The labels are binary self-reports in the four quadrants of the arousal-valence (AV) space. The other is the RECOLA dataset [10], composed

of multimodal recorded interactions between pairs of subjects during a problem solving task. For each interaction, this dataset contains a dimensional rating in the AV space, which was performed by six external annotators. These datasets have different experimental protocols including: type of interactions, emotional elicitation, spontaneity and labeling methods. In all cases, we use ECG features as input, with special interest in the HRV component. This provides a rich evaluation set for proposed methods. A web-demo [160] interface to rapidly test the methods is available[1]. The source-code of proposed methods is also freely available for academic purposes[2].

In the next section, related works on affect recognition using HRV are reviewed. In Section B.3 the datasets used in this work, the feature extraction stage and the experimental setup are presented. In Section B.4.1 a sSOM for affect recognition is presented. In Section B.4.2 different ELM classifiers are introduced. In Section B.5 the most relevant results are presented and discussed. Finally, the conclusions of this work are presented in Section B.6.

B.2 Related works

Several works used the HRV for AV classification. Valenza et al. [59] proposed a nonlinear method for feature extraction from HRV, along with EDA and RP, followed by principal component analysis (PCA) and a quadratic discriminant classifier. Authors obtained promising results using the standard International Affective Picture System (IAPS) as stimuli. Following a similar methodology, relevant improvements have been achieved with sound elicitation for classification in five classes in arousal and valence, using only HRV features [90]. Monkaresi et al. explored the binary classification in the AV space [161] and engagement recognition during a writing-reviewing process [162]. In these works, the authors combined remote HR sensing and facial expressions using a voting classifier composed by SVM, k-nearest neighbor (KNN), decision trees and logistic regression. These works have shown that affect classification using a source with low intrusiveness like HRV is feasible. Currently, improving the classifiers performance is an important challenge that can be addressed by research on novel methods.

Although the categorical approach to affect recognition has been employed successfully in several applications, many human states or traits vary continuously rather than in the rigid classes used in categorical approaches. In such cases the quantization into a few categorical labels might lead to a loss in model representativeness [31]. In comparison with the categorical problem, only a few publications have addressed the dimensional recognition challenges, yet it has become a trend in the affective computing community [163, 164, 31, 10, 165, 121]. Some works approximated dimensional affect indicators with fine-grained quan-

[1] `http://fich.unl.edu.ar/sinc/web-demo/dimensional-affect-recognition/`

[2] `https://sourceforge.net/projects/sourcesinc/files/emoHR`

tization scales on segmented data, as in [59]. Haag et al. [105] proposed an assessment of IAPS ratings using a multi-layer perceptron as regressor with multimodal inputs. Later, Bailenson et al. [166] used the same method to estimate levels of sadness and amusement with external raters. However, true dimensional affect estimation with physiological signals is quite recent. Ringeval et al. [131] used HRV, along with other physiological and audio-visual sources, to estimate arousal and valence levels during spontaneous interaction between humans. Several multimodal recognition systems have been tested with the dataset of this work, using for example PCA and linear regression (LR) [135], SVM for regression (SVR) [110, 111, 112], deep neural networks (DNN) [116], variations of the long-short-term memory recurrent neural network (LSTM) [117, 132, 133, 167, 134], relevance vector machines (RVM) [118], ensembles of random forests (RF) and neural networks [64], and more recently, an end-to-end approach using convolutional and recurrent networks [96]. The accuracy with physiological signals, particularly for HRV, is promising but should be improved to aim for naturalistic interaction applications.

B.3 Materials and evaluation setup

Datasets, preprocessing and further postprocessing of the output of classifiers are presented in this section. Then, the experimental setup is detailed for both classification and dimensional estimation tasks.

B.3.1 Classification

The dataset used for classification consisted of 16 laboratory sessions recorded by Monkaresi et al. [128]. Affects were elicited for different subjects using the IAPS and emotions were recorded as self-reports. Each session consisted of approximately 75 images, while one-lead ECG signal was being registered. Images were displayed sequentially in blocks with similar AV score. Each image was shown for 10 s, after which the subjects reported their affective state from 1 to 9 in the Self-Assessment Manikin (SAM) scale. Valence label was binarized in negative and positive classes. For arousal, low and high classes were defined. Our experimental setup with this dataset followed the same procedure of the authors. Sessions were segmented in chunks of one IAPS image. We used the same features provided by the authors: 84 classical features from ECG, including the distances between fiducial points of the PQRS complex, mean heart rate value and first order statistics. The RELIEF-F method was used for feature selection [168]. Each feature was normalized as $\check{x}_{n,j} = (x_{n,j} - \mu_j)/a_j$, where μ_j is the mean feature value and a_j one measure of deviation. The features of the validation partition were normalized using the training partition parameters.

The classifiers were tested for the classes defined above, using one classifier for each subject, one for arousal and another for valence (single-dimension approach)

as in [128]. A nested cross-validation was used, including partitions for training, parameter optimization and validation. The hyper-parameters optimization was performed without using the validation partition to get an unbiased performance estimation, including the model optimization. We used two performance measures. The first one is Cohen's Kappa [129],

$$\kappa = \frac{A_0 - A_c}{1 - A_c}, \tag{B.1}$$

where A_0 is the classification accuracy and A_c the by-chance probability observed in the confusion matrix. It takes into account a baseline reference for classification, being $\kappa = 0$ when there is no evidence that the classifier performs better than a random guess, and $\kappa = 1$ for perfect classification. The second measure is the unweighted average recall (UAR),

$$\text{UAR} = \frac{\sum_i^{n_c} A_i}{n_c}, \tag{B.2}$$

where A_i is the recall for class i data and n_c the number of classes. These coefficients are more robust to class imbalance than the simple accuracy.

B.3.2 Dimensional estimation

The RECOLA dataset [10] is a multimodal corpus that consists of recordings from dyadic interactions of subjects through an online communication channel (i.e. teleconference). Audio, video, ECG and EDA were registered while the participants were discussing how to solve a survival task. During the interactions affects were expressed spontaneously by the participants. Affects were tagged by six external raters as they perceived them. Their rating was based in a dimensional model of affects, using continuous values in arousal and valence. Also, the rating was annotated frame by frame for the first 5 minutes, thus all the variations in the AV space (according to the raters) are represented. A gold standard target was proposed by the dataset authors to convert the information of the six raters into a unique frame-by-frame rating. To do so, the target was defined as a weighted average of the raters based on their mutual agreement [100]. From the total of 46 subjects, 27 have a complete record of physiological signals. This set was divided by the authors into training, development and test partitions, containing 9 subjects each. In this work we use the 18 subjects that are publicly available (training and development partitions). Proposed methods with optimal hyper-parameters were also evaluated in the test partition.

The HRV signal was estimated from the ECG recording. First the R peaks were identified using the Pan-Tompkin method [71]. Then, the HR was estimated by taking the inverse of R-R distance and interpolating at ECG sampling frequency. Well-known HR features were obtained with a Hamming window of 20 s and a step of 0.5 s. This window length makes possible to have enough data

for feature extraction without losing time resolution [121]. In time domain, the HR mean and standard deviations were calculated. From spectral domain, low frequency band (0.04-0.15 Hz), high frequency band (0.15-0.4 Hz) and their ratio were used to estimate autonomous regulations. The spectral decay slope, modeled with a quadratic regression, provided more information of these regulations [122]. Additionally, the total spectral power, 5 fixed bands from 0.04-1 Hz and high order statistics (skewness and kurtosis) were included. The window length permits nearly continuous estimation with sufficient sample length for low frequency features [77]. The first and second derivatives of the features were computed to get information on how they changed in time. Contextual information was also considered by using frame stacking. Given a features vector, $\mathbf{x}_n$, a new set was constructed by adding the m frames before and after each frame, $\check{\mathbf{x}}_n = [\mathbf{x}_{n-m}, \ldots, \mathbf{x}_n, \ldots, \mathbf{x}_{n+m}]$. The features were normalized with the methods detailed above in a session-basis, as described in [100]. Two postprocessing methods were applied to the models outputs. First, a filter was applied to reduce the outputs noise. It was optimized from two common methods in time series processing, the moving average and the exponential smoothing. Then, an output correction factor was tested to adjust the output amplitude. This factor was defined as the mean ratio between filtered outputs and target amplitudes in the training set.

The set of sessions was divided in a 3-folds nested cross-validation scheme, so each session was used either for training or testing at a time. To compare the estimations with the targets, we used the Lin's concordance correlation coefficient ρ_c [130], which is the scoring metric used by other works on the RECOLA dataset and it is the official metric of the Audio/Visual Emotion Challenge and Workshop (AVEC) since 2015. This metric is defined as

$$\rho_c(\mathbf{y}, \hat{\mathbf{y}}) = \frac{2\rho\sigma_{\mathbf{y}}\sigma_{\hat{\mathbf{y}}}}{\sigma_{\mathbf{y}}^2 + \sigma_{\hat{\mathbf{y}}}^2 + (\mu_{\mathbf{y}} - \mu_{\hat{\mathbf{y}}})^2}, \tag{B.3}$$

where ρ is the Pearson's correlation coefficient, μ_y is the mean and σ_y is the standard deviation. The range of ρ_c is [-1,1], taking values around 0 if there is no concordance evidence, and 1 for a perfect concordance. This coefficient is an improvement of ρ, as it considers the correlation of the signals in time along with the mean square error.

The proposed methods were tested in different scenarios. In all cases, the gold standard rating was used as the estimation target. In first place, the methods were faced to the single-dimension estimation, training models for arousal and valence independently. However, it has already been shown that arousal and valence are dependent during emotional elicitation [47]. Thus, the two-dimensions AV estimation was conducted for comparison. In this experiment, the two outputs were estimated simultaneously by one model, whose parameters were optimized to maximize the mean ρ_c of both targets. As a baseline, two standard classifiers

were evaluated. One is a SVR with Gaussian kernel[3], the regularization factor (C) optimized in the range $[2^{-5}, 2^{25}]$ and the Gaussian exponential (γ) in the range $[2^{-30}, 2^{-5}]$. The other classifier is a RF[4], in which the number of trees was optimized in the range $[5, 150]$ and the size of the features subsets from 5 to the whole feature set for each tree. To compare the results with previous works, additional experiments were conducted. The best models were trained with the training partition, and validated on the development partition. In this case, the best hyperparameters were taken from cross-validation experiments as in [64], thus minimizing model overfitting. A final evaluation was made using the test partition. The optimal models from cross-validation experiments were trained with the whole public set (training and development partition) and labels were estimated on the test set of features. These estimations were made only once and sent to the authors of the RECOLA dataset for evaluation.

B.4 Methods

B.4.1 Supervised self-organizing maps

A self-organizing map is a neural network generally composed by one bi-dimensional layer of units. This model has been proposed for dimensionality reduction, clustering and classification [124]. In this section, we propose a novel method to train a sSOM for dimensional affect estimation. To this end, the inputs in the training stage will be the features extended with the target affects. Rather than minimizing an error function between the model output and the expected targets, an unsupervised method creates a map based on the similarity between the extended input vectors. Different regions are conformed on this map, associating the values of features and targets. When new unlabeled data is presented, the features are compared with the learned weights of all units in the map and the closest unit is chosen as output unit. Then, the affect learned by this unit is the estimated target, which was chosen based on the spatial structure of the map previously defined by the training data. An important advantage of this method is that the high-dimensional input space is mapped into a 2D representation. Therefore, new relations between the features and the affective space can be discovered by simple inspection.

Formally, given a set of N samples with F-dimensional features and P-dimensional continuous targets, lets define the input matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$, with $\mathbf{x}_n \in \mathbb{R}^F, n = 1, \ldots, N$, and the target matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T$, with $\mathbf{y}_n \in \mathbb{R}^P$. The features and targets in the training set are concatenated as a single input matrix. The new input matrix is given by $\check{\mathbf{X}} = [\check{\mathbf{x}}_1, \ldots, \check{\mathbf{x}}_N]$, with $\check{\mathbf{x}}_n = [\mathbf{x}_n, \lambda \mathbf{y}_n] \in \mathbb{R}^{F+P}$,

[3]Implemented with the quadratic programming functions of Matlab. It is included in the provided source code.

[4]Standard Matlab implementation: TreeBagger class.

where the scaling factor λ must be set to balance the influence of the targets in the map topology.

The sSOM have a rectangular array of units s_j, with $j = 1, \ldots, J$. For a given input $\check{\mathbf{x}}_n$, the output of s_j is given by

$$h_j = \varphi(\check{\mathbf{x}}_n, \check{\mathbf{w}}_j), \tag{B.4}$$

where $\check{\mathbf{w}}_j = [\mathbf{w}_j^x, \mathbf{w}_j^y] \in \mathbb{R}^{F+P}$ is the synaptic weight vector, composed by the feature weights $\mathbf{w}_j^x$ and the target weights $\mathbf{w}_j^y$. The operator φ is a similarity function, usually based in the euclidean distance. Weights are traditionally instanced at random [124]. However, the data distribution can be used to avoid local minima and speed up the training. Thus, an alternative to random initialization is to use PCA in the input space. First, the method finds the two greater eigenvalues and eigenvectors from the training set. Then, the weights of the map are generated by linear spanning in the two dimensions. In this way, the main data variability is initially arranged along the main axes of the map.

The sSOM training is an iterative procedure. At each time $t = 1, \ldots, T$, a sample $\check{\mathbf{x}}(t)$ is presented to the map. The best matching unit is the one with higher similarity with the input pattern. It is found by solving

$$s^*(t) = \arg\min_j ||\check{\mathbf{x}}(t) - \check{\mathbf{w}}_j||_2. \tag{B.5}$$

The method rewards the neuron s^* by adjusting $\check{\mathbf{w}}_{s^*}$ for a better matching with the sample. To induce a topological ordering in the map, the rewarding effect is scattered through the neighbouring units. The neighbours are defined in hexagonal shape, thus a 1-unit neighbourhood is a set of 6 units plus the central unit. Then, the weights are updated using the steepest-descent gradient optimization

$$\check{\mathbf{w}}_j(t+1) = \begin{cases} \check{\mathbf{w}}_j(t) + \alpha[\check{\mathbf{x}}(t) - \check{\mathbf{w}}_j(t)], & \text{if } s_j \in \mathcal{N}_{s^*} \\ \check{\mathbf{w}}_j(t), & \text{if } s_j \notin \mathcal{N}_{s^*} \end{cases}, \tag{B.6}$$

where $0 < \alpha < 1$ is the learning factor, and $\mathcal{N}_{s^*}$ is a neighbourhood function around s^*. For the early iterations, the $\mathcal{N}_{s^*}$ radius and α take large values. This configuration leads to a rough ordering of the map, defining the main topographic zones. In the later iterations, $\mathcal{N}_{s^*}$ and α are reduced until only s^* is affected by the optimization algorithm. This results in a fine-tuning of the map while the main topological structure is conserved. In addition to the traditional planar sSOM, a toroidal form can be defined by linking opposed border units of the plane in a same neighborhood, thus every unit will have the same amount of neighbours. Preliminary evaluations of toroidal model have not provided better estimations than the planar one, and graphical analysis of a toroid is more complicated. Thus, only the planar map will be used.

Once the sSOM training is complete, similar inputs will lead to closer winner units. As training inputs contains the features and targets, each region of the map

will model the spatial relations of both spaces. Then, in the recognition stage only the feature weights $\mathbf{w}_j^x$ are used. Thus, the best matching unit is obtained with

$$s^* = \arg\min_j ||\mathbf{x} - \mathbf{w}_j^x||_2, \tag{B.7}$$

and the output estimation is given by

$$\tilde{\mathbf{Y}}_{s^*} = \frac{\mathbf{w}_{s^*}^y}{\lambda}. \tag{B.8}$$

The smoothness of the outputs may depend on both the input data and the size of the map. Therefore, a spatial interpolation method is incorporated as last step. Given the input $\mathbf{x}$, the K closest units are determined, $[s_1^*, \ldots, s_K^*]$, with s_1^* the best matching unit as in (B.7). Then, the smoothed output is obtained with the weighted average

$$\bar{\mathbf{Y}} = \sum_{k=1}^{K} \gamma_k \tilde{\mathbf{Y}}_{s_k^*}, \tag{B.9}$$

where

$$\gamma_k = \frac{||\mathbf{x} - \mathbf{w}_{s_k^*}^x||_2^{-1}}{\sum_{j=1}^{K} ||\mathbf{x} - \mathbf{w}_{s_j^*}^x||_2^{-1}} \tag{B.10}$$

is the normalized inverse distance for each s_k^* in the feature space. With this expression, the closest s_k^* in the feature space receives the higher weight.

B.4.2 Extreme learning machines

The theoretical context of ELM includes several related methods [126]. In this section, two different ELM approaches are introduced: the original model as a neural network, and a later derivation based in kernels.

In the first conception of ELM, the classifier can be seen as a neural network with one hidden layer (nELM). The central paradigm is that the hidden units are randomly generated, thus the tuning of their parameters is avoided. As a direct consequence, the training time is dramatically reduced compared with other training methods. For a formal derivation, consider J hidden units with F inputs and P output units. The output of the hidden layer is given by

$$h_j = \Phi(\mathbf{v}_j^T \mathbf{x} + b_j), \tag{B.11}$$

where Φ is the activation function, $\mathbf{v}_j$ the input weights and b_j the bias for the j-th hidden unit. This can be expressed in a matrix form by defining the hidden-layer output matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^T$, also called the feature projection matrix, and $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_P]$ as the output layer weights, with $\mathbf{w}_p \in \mathbb{R}^J$ and $p = 1, \ldots, P$. Then, the nELM output can be written as

$$\tilde{\mathbf{Y}} = \mathbf{H}\mathbf{W}. \tag{B.12}$$

If Φ is an infinitely differentiable function, and $(\mathbf{v}_j, b_j)$ are randomly selected, it can be demonstrated that for any pair $(\mathbf{X}, \mathbf{Y})$ there exist a number $J < N$ such $||\tilde{\mathbf{Y}} - \mathbf{Y}|| < \epsilon$ for any small ϵ [126]. This means that the ELM can approximate the target $\mathbf{Y}$ of a given input $\mathbf{X}$ by adjusting only the number of hidden units and the output weights. To find $\mathbf{W}$, the problem can be stated as

$$\underset{\mathbf{W}}{\text{minimize}} \;\; ||\mathbf{HW} - \mathbf{Y}||_2, \tag{B.13}$$

which is a least square optimization problem. The smallest norm solution is given by

$$\hat{\mathbf{W}} = \mathbf{H}^{\dagger}\mathbf{Y}, \tag{B.14}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose pseudo-inverse [127].

A generalized ELM method based on kernels (kELM) can be derived from this theory [159]. To improve the solution stability and generalization, a regularized optimization problem was proposed as

$$\begin{aligned} \underset{\mathbf{W}}{\text{minimize}} \quad & \frac{1}{2}||\mathbf{W}||_2 + C\frac{1}{2}\sum_{n=1}^{N}||\epsilon_n||_2 \\ \text{subject to} \quad & \mathbf{h}_n\mathbf{W} = \mathbf{y}_n^T - \epsilon_n^T, \end{aligned} \tag{B.15}$$

where ϵ_n is the training error vector for the sample $\mathbf{x}_n$ and C a regularization factor. The solution is given by

$$\hat{\mathbf{W}} = \mathbf{H}^T\left(\frac{1}{C}\mathbf{I} + \mathbf{H}\mathbf{H}^T\right)^{-1}\mathbf{Y}. \tag{B.16}$$

Let be $\breve{\mathbf{H}}$ the feature projection of the training set $(\breve{\mathbf{X}}, \breve{\mathbf{Y}})$, and $\mathbf{H}$ the projection of any other set $(\mathbf{X}, \mathbf{Y})$. The estimation of $\mathbf{Y}$ is given by (B.12) and (B.16)

$$\tilde{\mathbf{Y}} = \mathbf{H}\hat{\mathbf{W}} = \mathbf{H}\breve{\mathbf{H}}^T\left(\frac{1}{C}\mathbf{I} + \breve{\mathbf{H}}\breve{\mathbf{H}}^T\right)^{-1}\breve{\mathbf{Y}}. \tag{B.17}$$

With the selection of a kernel function $\mathcal{K} : (\mathbb{R}^F, \mathbb{R}^F) \to \mathbb{R}$, the kernel matrix for the inputs $(\mathbf{X}, \mathbf{X}')$ is defined as

$$\mathbf{\Omega}(\mathbf{X}, \mathbf{X}') = \mathbf{H}\mathbf{H}'^T : \Omega_{i,j} = \mathbf{h}_i \cdot \mathbf{h}_j = \mathcal{K}(\mathbf{x}_i, \mathbf{x}'_j). \tag{B.18}$$

Thus, (B.17) becomes

$$\tilde{\mathbf{Y}} = \mathbf{\Omega}(\mathbf{X}, \breve{\mathbf{X}})\left(\frac{1}{C}\mathbf{I} + \breve{\mathbf{\Omega}}\right)^{-1}\breve{\mathbf{Y}}, \tag{B.19}$$

where $\breve{\mathbf{\Omega}}$ is the training kernel matrix. From (B.19), it can be seen that the kernel function replaces the projection matrices.

Several hyper-parameters detailed in Section B.3 and B.4 were optimized with a grid search. The order of feature derivatives, frame stacking size and post processing parameters were optimized for each case. The feature normalization factor a was the standard deviation for sSOM and the maximum amplitude of the data for ELM. For sSOM, we explored different map architectures (size and shape), the scaling factor λ, the spatial interpolation method and training length. For nELM, a hidden layer of variable size and standard activation functions were used, including sigmoid, hard-limit, and sinusoidal functions. The kELM was implemented with a radial basis function, with the exponential coefficient γ and associated regularization factor C being tuned in a grid with logarithmic scale.

B.5 Results and discussion

In the first part of this section we show the classification performance of the proposed methods in comparison with baseline classifiers and previous works. In the second part we show experimental results for dimensional affect estimation on the RECOLA database. We show the distinctive use of sSOM as a qualitative model to explore affects and their relations with the physiological features. Then we compare the proposed models and the baselines in a quantitative way with cross-validation results. In the last part, we make a comparison with state-of-the-art works by using the same dataset partitions.

B.5.1 Categorical affect classification

The results shown in Table B.1 are the average of 10 randomized cross-validation repetitions on the categorical dataset. The columns are the single-dimension classification tasks, while in the rows are listed our models, baseline classifiers (SVM with radial basis function kernel and RF) and the reference for comparison. In [128], a vote classifier was used to combine the decision of four standard classifiers: SVM, KNN, decision trees and logistic regression. From Table B.1 it can be seen the most effective classifiers are kELM for arousal and SVM for valence.

Our methods outperformed previous results on the classification task. Although that classes were balanced and selected to be contrasting (high versus low, and positive versus negative), the score was considerably higher for valence. This may suggest that the selected features from ECG had a better discriminability or the elicitation process was more effective in valence. This was consistent with the results reported by the authors of the database [128]. When comparing arousal results with the reference, our methods show a higher difference in the kappa score. By using a single classification model as proposed here, instead of several classifiers, the number of tuning parameters has been reduced, making a simpler model for the problem. As it was shown, by using the same features and experimental setup, the classifiers proposed here can improve the results for binary categorization of arousal and valence.

Comparing now the models with higher scores, SVM and kELM show similar results. Both methods share the theoretical objective of projecting data to a higher dimension where data may be easier to separate, in this case using the same kernel function. However, kELM is faster and uses less memory during the optimization. Differences between kELM, nELM and RF are significant ($p < 0.01$, one-way ANOVA) for arousal and valence in both performance measures. However, kELM required more resources as the algorithm uses the training data to provide estimations. The trained sSOM is represented in a small set of parameters, thus the memory usage for training is considerably low. The sSOM seems to be effective as well, which may be explained by the unsupervised association of features and affects, providing robustness to outliers. This model also shows a significant difference with nELM and RF for valence and nELM for arousal ($p < 0.01$). The better scores provided by kELM can probably be explained by its capacity for non-linear modeling of the feature space.

	Arousal		Valence	
Classifier	κ	UAR	κ	UAR
Vote classifier [128]	.071	-	.191	-
SVM	.143	.570	**.213**	**.607**
RF	.109	.558	.163	.582
sSOM	.137	.566	.202	.597
nELM	.068	.541	.119	.559
kELM	**.148**	**.576**	.208	.603

Table B.1: Mean κ and UAR for binary affect classification on Monkaresi et al. dataset.

Performance for the classification task is lower than the dimensional estimation (as will be detailed in the next section). The general differences between our framework in classification and dimensional estimation tasks could be explained by the effect of several factors. In the first place, emotion expression is different in the datasets. Emotions in both datasets are naturally expressed, this is not acted. However, the dataset used for classification involves induced emotions (using IAPS) and the dimensional estimation dataset involves spontaneous emotions. The report is also different; the use of affect reports of several raters in RECOLA may involve a better estimation. Last but not least, categories in the classification case are binarized from a dimensional model. This may restrain emotion expression, as extreme emotions are in the same category than near-neutral values. In this way, continuous labels may be more difficult to register but they have a richer expression that the classifiers can use.

B.5.2 Dimensional affect estimation

We analyze first the interesting visual information obtained from sSOM trained on the RECOLA dataset. The high-dimensional space of the features and targets can be reduced to intuitive bi-dimensional representations. Some of them are shown in Figure B.1. Each one represents the distribution of a coefficient of the synaptic weights $\check{w}_j$ in the map. Let us take for instance the *Arousal* plane. As explained in Section B.4.1, arousal is part of an extended input of the model during training. The hexagonal cells represent the sSOM units, placed with their neighbours as they are in the model. The value of the input, in this case the arousal level, is modeled with the $\check{w}_j$ of each unit. This value is indicated in the image with a color scale that ranges from blue at the minimum to red at the maximum value. Upon visual inspection, one consequence of the training algorithm is that similar values are arranged in neighbouring units. In the *Arousal* plane it can be seen that the low-arousal zone was ordered in the upper-left of the image, increasing approximately along with the vertical axis to the bottom, which is the high-arousal zone. In the same way, looking now at the *Valence* plane, we can see a smooth value progression in an almost perpendicular direction to the *Arousal* plane.

Comparing now the targets (arousal and valence) with the other input planes (the actual features), relations between them can be assessed in a qualitative form. Although all features are considered in a multidimensional way to obtain an accurate estimation, this analysis can contribute to finding relevant features. From Figure B.1, it seems that P_1' and LF/HF' are strongly related with arousal. This can be seen by observing that the high P_1' zone and the low LF/HF' zone are overlaid with the high-arousal zone. In a similar way, LF/HF, HF and the HR statistics (σ_{HR} and μ_{HR}) can be associated with the valence distribution. It has been argued that the HRV is related with valence and well-being, specifically HF being directly correlated with valence [169]. It can be seen in Figure B.1 that low valence has a coincident area with low and medium HF, as well as with σ_{HR} and μ_{HR}, thus adding empirical support to the argument. This type of analysis provides a tool to visualize the relationship between affects and important features from the data.

For a practical and compact representation of the emotion model learnt by sSOM, we can merge the arousal and valence maps from Figure B.1 in a unique map as in Figure B.2. The targets are shown in colors, red for arousal and blue for valence. Now their values are coded in the size of the hexagons instead of a color scale. This map provides an idea at a glance of the structure and relation of arousal and valence in the model. The sSOM units are represented in the same positions as in Figure B.1, so the topological relations between features can be related with this new target map. Even more, if it is used to estimate dimensional affects in real-time, the AV map could serve as a display to show graphical interactions between the variables, highlighting the winner unit at each moment.

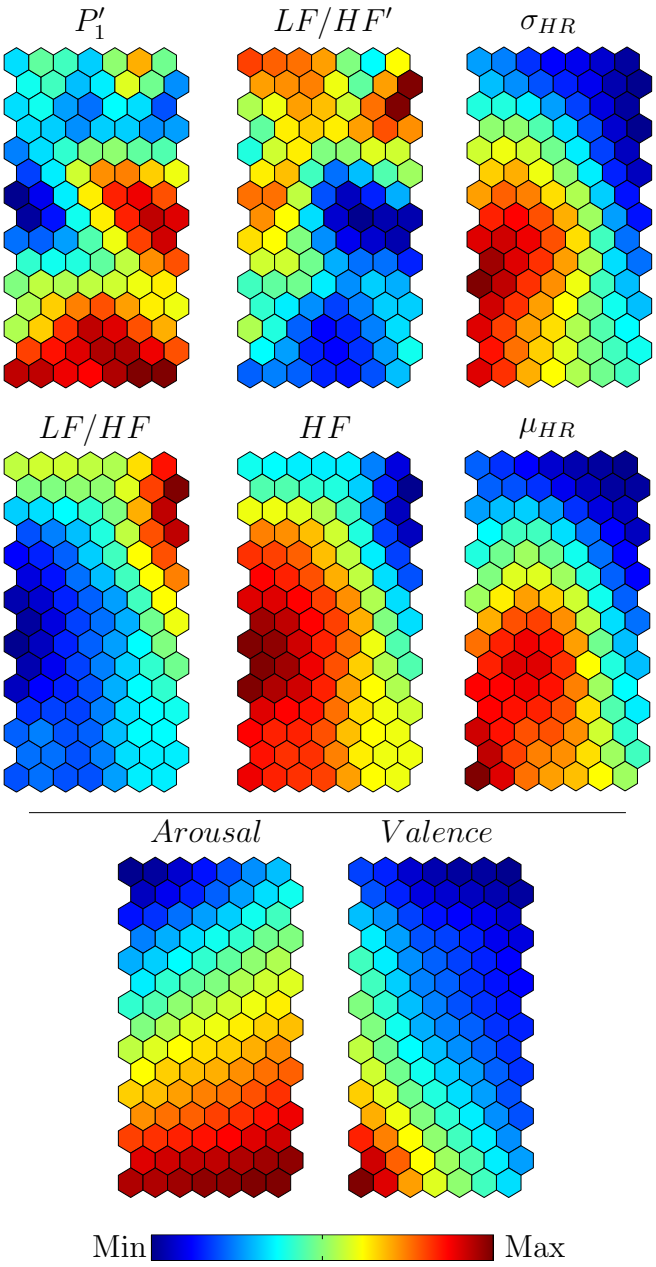


Figure B.1: Graphical representation of the inputs and targets in a trained sSOM. On the top, six features layers are shown: the first derivative of P_1 frequency band (P_1'), the low and high frequency bands ratio (LF/HF) and its first derivative (LF/HF'), the high frequency band (HF), the amplitude of HR (σ_{HR}) and its mean (μ_{HR}). On the bottom, the target layers for arousal and valence are displayed, along with the color-bar for reference.

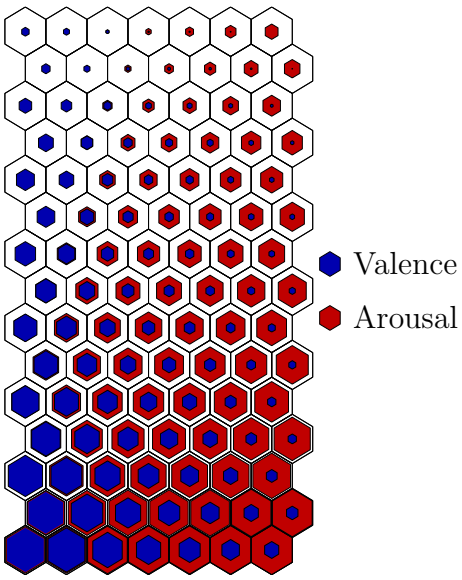


Figure B.2: Graphical representation of the AV space using sSOM. Output components of the trained map are superposed in the same plane. Arousal is represented in red and valence in blue. The level of these variables in each sSOM unit is coded by the size of each hexagon. Notice that this is the same information displayed by the trained model, now summarized in one single image.

The relationship between the traditional AV plane and the new data-driven representation is schematized in Figure B.3. The gray dots in the AV plane are theoretical affects that may be reported by a subject during an affect elicitation experiment. Some idealized cases (stressed, excited, relaxed and sad) are displayed in both representations. Thus, Figure B.3 illustrates how affects can be mapped from the theoretical AV plane to the sSOM by looking to the arousal and valence levels. Using this relationship, the sSOM graphical representation can be discussed using the following example case. It has been reported that affective stimuli (like audiovisual resources) are not equally effective to induce affects all over the AV plane [37, 47, 48]. In fact, in those works it was found that with low arousal levels there is very little possible variation in valence, which stays near the neutral point. However, for high arousal it can be reached the full spectrum of valence expression. That is, if the affects are plotted in a Cartesian AV space, the dots are inscribed in a parabolic shape, as seen in Figure B.3. If we now compare this results and the map of Figure B.2, we can see a similar behaviour in the affect representation of our experiment. The complete range of valence is only observable in high arousal zone (the bottom part of the map), while in low arousal zones the valence is between neutral and negative. This way, the theoretical relations between arousal and valence have been warped to the sSOM, using only the training data. This mapping provides an alternative representation of the AV space given the mutual closeness between the feature and target samples,

as a direct property of the sSOM training algorithm. Moreover, this example case of data-driven representation provides empirical evidence to support theoretical models described previously. This is indeed an advantage of the sSOM over black-box models, in that it provides graphical representations of data isles and may provide support for theoretical assumptions, based only on the training data.

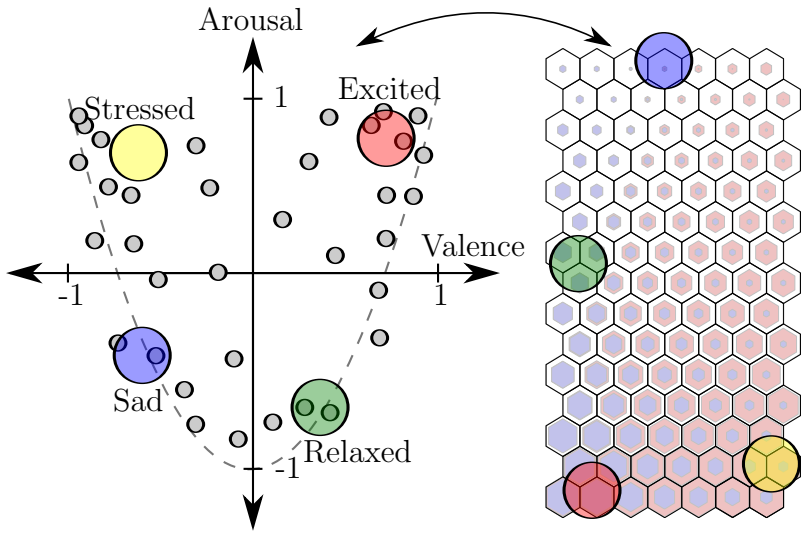


Figure B.3: Representation of the mapping between the theoretical AV space and the sSOM graphical model. The pictured affects (stressed, excited, sad and relaxed) are illustrated to exemplify the data-driven mapping.

The first cross-validation results on the RECOLA dataset are shown in Table B.2. In the columns are the estimated targets, and in the rows are the proposed methods along with standard models as SVR and RF for comparison. Methods are detailed as single-dimension when trained with either arousal or valence, and two-dimensions when trained with both outputs simultaneously. It can be seen that ELM achieve the higher concordance rates. On the one hand, nELM has a fast and low memory implementation. On the other, kELM provided the lowest variances, denoting a more stable model across the tested sessions. However, sSOM follows these results closely with the advantage of providing an explicit model for visual analysis, as mentioned above. In agreement with [100], arousal estimation was shown to be more accurate, with higher ρ_c and lesser variance. It is also interesting to note that nELM works very well with the RECOLA dataset but not so in the classification dataset (Section B.5.1). A possible explanation is that the generalization capacity of nELM improve with the amount of data available.

Estimating both targets with the same model seems to slightly improve sSOM and nELM performance in arousal. However, the general performance is similar compared to the single-dimension approach and it did not yield significant differences. The results show that the combination of arousal and valence with

	Classifier	Arousal	Valence
	SVR	.378 (.030)	.243 (.064)
	RF	.369 (.018)	.282 (.028)
Single-dimension output	sSOM	.362 (.032)	.313 (.059)
	nELM	.366 (.039)	**.322** (.054)
	kELM	**.388** (.009)	.320 (.049)
Two-dimensions output	sSOM	.364 (.033)	.312 (.059)
	nELM	.379 (.039)	.313 (.060)
	kELM	.388 (.010)	.321 (.044)

Table B.2: Mean ρ_c and standard deviation from 3-fold cross-validation on the RECOLA dataset. Each session features were normalized independently. Proposed models were trained using a single-dimension (one model for arousal and another for valence) and using the two dimensions togheter.

the current models does not seem to provide better recognition capabilities than individual target models. Measuring the concordance between arousal and valence targets for all sessions, it yields a mean of 0.29. This suggests concordance between the targets and may explain the lack of improvement, as the additional information provided when estimating one target along the other is not leveraging the results. However, the bi-dimensional outputs makes possible the analysis presented with sSOM about Figure B.2.

Although the proposed models are able to perform the recognition in real-time, an additional factor that should be considered with the recognition performance is the computational cost of training the models. Both sSOM and nELM models are compressed in low quantity of parameters and low training time. On the contrary, the training data is needed to compute every estimation with kELM, as seen in Section B.4.2. The training time for kELM was significantly higher than nELM and sSOM models, but these are much lower than SVR. This difference may be important for future applications in limited hardware, as wearable devices, or when using bigger datasets.

As described in [100], the gold-standard rating is composed by six human raters. Let us consider a rater as either one of these humans or one of the proposed models. It is interesting to evaluate the behaviour of the models in comparison to the humans. The agreement between a pair of raters can be measured with the mean ρ_c across the sessions. The agreement of each rater with the other human raters is shown in Table B.3. In the rows are listed the six human raters, the mean inter-rater agreement and the proposed models. The columns are arousal and valence as independent targets. These results show that the proposed models have a mean agreement superior to some raters (the rater 6 in arousal and rater 5 in valence). In the case of arousal, the ELM models even approximate the second least agreeing rater (the number 5). Although valence estimation is more

Raters	Arousal	Valence
Rater 1	.305	.386
Rater 2	.296	.361
Rater 3	.349	.327
Rater 4	.231	.259
Rater 5	.223	.181
Rater 6	.113	.298
Mean inter-rater	.253	.302
sSOM	.203	.192
nELM	.216	.214
kELM	.215	.208

Table B.3: Mean ρ_c between a rating (either from a human rater using audio-visual cues or a model using physiology) and all the human raters in the database.

challenging, consistent with the results discussed above, humans have a higher inter-rater agreement for valence. This may be explained by the natural human ability to identify valence states from face expressions [170]. Comparing the mean inter-rater concordance from Table B.3 and results from Table B.2, it can be seen that the models have competitive performance. Therefore, it can be said that the proposed models could be playing the role of an external rater with a consistent agreement with other raters and good rating towards the gold standard label.

An example of the estimated ratings for a validation session is shown in Figure B.4. The arousal and valence targets correspond to the session tagged as *dev-3* in RECOLA. These are compared with the sSOM, nELM and kELM outputs. The shaded area is the standard deviation of ratings given by the human raters. It can be seen that models yield close estimations and follows the main events in the target signal, like the peaks around 45 s and 150 s in arousal. Also, the estimations mainly remain in the shaded zone. For this session, the percentage of the arousal estimations inside the human rater deviation is 83%, 80% and 81% for nELM, kELM and sSOM respectively, while for valence is 71%, 67% and 78%. These relations are similar for the whole database, with 73%, 75% and 76% for arousal, 67%, 68% and 71% for valence. It can be seen that the sSOM outputs follow the reference more closely in general. However, Table B.2 shows that ELM models achieve higher ρ_c, which may be explained by its higher sensibility to small variations. This can be seen, for example, around the 260 s point in Figure B.4. Another aspect that have been discussed in previous worksx is the asynchrony between emotional expression and the emotional labels provided by the external raters [131]. It was reported [110] that when using the HRV signal, a delay on the training labels does not improve the performance of the classifier. As HRV responses are slower than audio-visual cues, it is possible that the rater delay may have been partially compensated with the physiological delay.

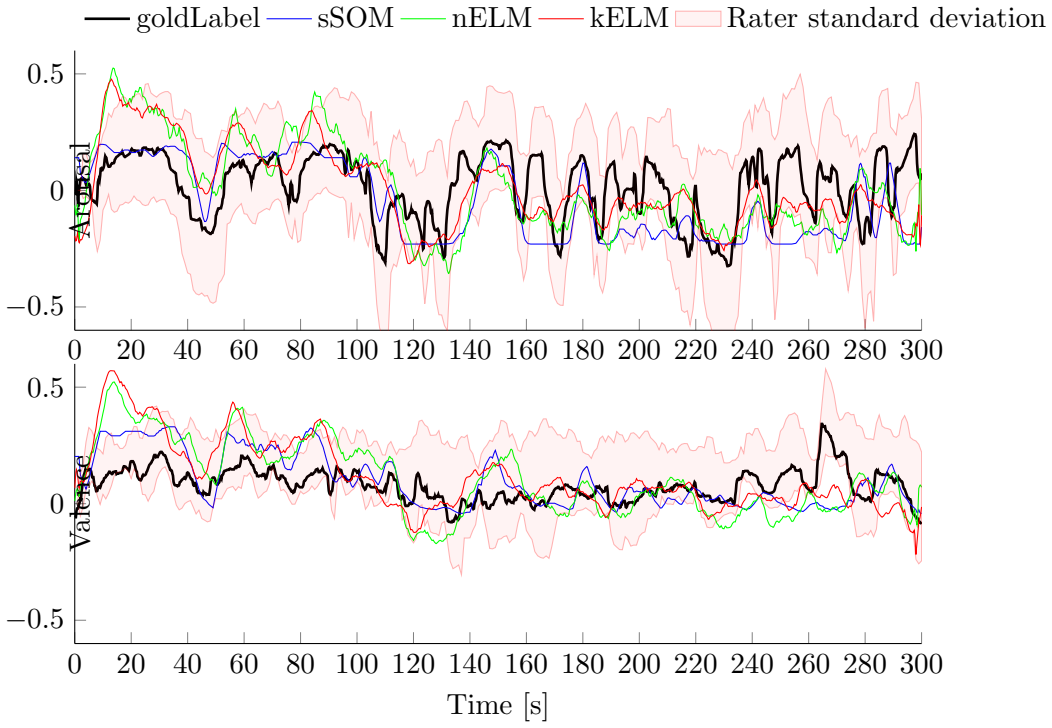


Figure B.4: Comparing the outputs of the models with their targets. Gold-standard for arousal and valence is in bold black line. Model estimations are represented with color lines. The shaded area correspond to the standard deviation of human raters.

As detailed in the previous section, the experiments described here involve features that were normalized for each session independently, as in [100]. However, a more challenging case is the task of estimating dimensional affects on a totally new subject in real-time, without any prior sensing on this new subject. This case can be evaluated with a small change in the feature normalization stage. Instead of normalizing the features in a session basis, the normalization parameters (features mean and deviation) are obtained from the training sessions only. Results for single and two-dimensions models are shown in Table B.4. It can be observed that ELM models can provide a better estimation and also seems more robust than sSOM to the feature normalization challenge. Differences between ELM methods and baseline classifiers are significant in all experiments ($p < 0.05$). Moreover, kELM achieved significant improvement against the other methods as well ($p < 0.05$). This suggest that feature complexity of multiple subject experiments is better handled with proposed methods.

B.5.3 Comparisons with state-of-the-art methods

Previous works on the RECOLA dataset reported their results on the partition called *development* (detailed in [100]). State-of-the-art models that used only the

	Classifier	Arousal	Valence
	SVR	.155 (.019)	.104 (.046)
	RF	.119 (.048)	.126 (.018)
Single-dimension output	sSOM	.165 (.051)	.141 (.060)
	nELM	.217 (.025)	.230 (.034)
	kELM	.260 (.002)	.223 (.047)
Two-dimensions output	sSOM	.181 (.038)	.148 (.056)
	nELM	.262 (.008)	**.253** (.029)
	kELM	**.263** (.007)	.227 (.046)

Table B.4: Mean ρ_c and standard deviation from 3-fold cross-validation on the RECOLA dataset. The normalization parameters (features mean and deviation) are obtained from the average of training sessions.

Classifier	Reference	Arousal	Valence
Using ECG features			
Vote classifier	Ringeval et al. 2015 [100]	.275	.183
LSTM	Chao et al. 2015 [133]	.222	.182
LSTM	Chen et al. 2015 [132]	.333	.314
DNN-LSTM	He et al. 2015 [117]	.297	.293
DNN	Cardinal et al. 2015 [116]	.262	.124
NN ensemble	Kachele et al. 2015 [64]	.344	.256
oaRVM	Manandhar et al. 2016 [118]	.293	.274
S-fusion SVR	Weber et al. 2016 [112]	**.468**	.221
Using HRV features			
PCA + LR	Povolny et al. 2016 [135]	.391	.388
SVR	Valstar et al. 2016 [110]	.379	.293
SVR	Sun et al. 2016 [111]	.392	.264
S-fusion SVR	Weber et al. 2016 [112]	.424	.413
LSTM	Brady et al. 2016 [134]	.357	.364
End-to-end	Keren et al. 2017 [96]	.426	**.419**
sSOM		**.402**	.354
nELM		.399	**.375**
kELM		.388	.338

Table B.5: Mean ρ_c of the proposed models and other works on the RECOLA *development* partition.

ECG recordings for the affect recognition are listed in Table B.5. We show these results together with our results in the same partition for comparison. Our results

Classifier	Reference	Arousal	Valence
Vote classifier	Ringeval et al. 2015 [100]	.192	.139
DNN	Cardinal et al. 2015 [116]	.161	.121
Linear SVR	Valstar et al. 2016 [110]	.334	.198
End-to-end	Keren et al. 2017 [96]	.360	.225
sSOM		.404	.273
nELM		**.421**	**.321**
kELM		.367	.293

Table B.6: Mean ρ_c of the proposed models and other works on the RECOLA *test* partition.

were achieved with the proposed methods optimized by cross-validation, thus minimizing the overfitting on development partition. As expected, our results are near the cross-validation results from Table B.2, with the sSOM performing better for arousal and nELM for valence.

The state-of-the-art methods were reported in two groups. The first group of publications used general features of the ECG signal. The second group uses only HR and HRV derived features, with an overall better performance, except for [112], which provides a better score for arousal using ECG features. Many of the revised works ([133, 132, 117, 134]) were based on the LSTM model, which was introduced for this database in [131] and it is considered a state-of-the-art model for dimensional affect estimation [117, 31]. In these models, interesting variations have been proposed by defining different loss functions, like the ϵ-insensitive loss in [133] and the concordance correlation in [132], instead of using square-error based functions. These networks use memory inputs to learn from the evolution of the features at different time scales, thus are suitable for estimating time series. In our models, time context information was introduced using the feature derivatives and frame stacking. This short time context provided enough information to achieve competitive results. Moreover, our methods may be more robust when long time recordings are not available. Other works use DNN models [116] and combinations of DNN and LSTM [117]. The SVR with linear kernels is also popular for this task, using L2 [110] and L1-regularized [111] loss functions. One of the outstanding results for arousal is a subject-level fusion (S-fusion) strategy using SVR, achieving a wide difference with other works [112]. However, authors state that the model lack of generalization capabilities, not being able to reach the baseline results of the test partition. This can be explained by the final stage of the model training, which was adjusted using the same developing partition (which was used to measure the performance), thus overfitting the model. On the contrary, works like [64] uses a cross-validation stage to perform hyper-parameter optimization and thus recognition rates for unseen test data are more predictable. The most recent work, which is also better than the others in valence prediction, is

an end-to-end approach [96]. In that work, convolutional and recurrent networks are used to learn features and time dynamics directly from HRV signal, thus avoiding hand-made features.

Related to the discussion on two-dimensions models (Section B.5.2), an output-associative RVM (oaRVM) was proposed in [118]. This model is trained using a feed-back of both arousal and valence estimations, as proposed in [171] for audiovisual features. They show that their two-dimensions approach achieved an improvement compared to other single-dimension models. However, authors from [132] could not find important differences in recognition performance between single and two-dimensions approaches for the reported modalities. In our work, the two-dimensions approach is of importance in two cases. First, using both target variables in sSOM makes it possible to visualize relationships with the multidimensional feature space. Secondly, we show that there is a small improvement in the case of a new session to be estimated in real-time, without normalization in a session basis. Moreover, there is practically no increase in the computational cost of training the proposed models for simultaneous arousal and valence estimation in contrast with the heavier computational cost of oaRVM.

The available results on the test partition are shown in Table B.6. It can be seen that our methods improved the state-of-the-art on this partition. Among them, nELM approach achieved the best estimations. This could be explained by their random hidden layer generation, providing optimal solutions with good generalization. The sSOM method also achieved competitive results. The patterns between features and targets were effectively modeled with the sSOM structure and the unsupervised training. Summarizing the results of Table B.6, estimation of both targets using only HRV was effectively improved with the proposed methods.

B.6 Conclusions and future work

In this work two new methods for affect recognition have been presented, using features extracted only from the HRV. A novel supervised self-organization model (sSOM) was proposed to improve the recognition accuracy but also to provide a graphical representation of relations between features and targets. Contrary to a black-box model, the sSOM represents a graphical superposition between sensed data and affects. Given the sSOM properties, numerical and categorical variables can be represented, making it a very versatile model for HCI applications. Two novel methods based on extreme learning machines (nELM and kELM), were also applied to these tasks. These models were evaluated in classification and dimensional affect estimation, providing competitive performance compared with state-of-the-art works. In classification, the best results were achieved by kELM in both arousal and valence. In the dimensional estimation task, proposed models outperformed state-of-the-art results in the RECOLA test partition. sSOM obtained very good performance according to the quantitative measures and also

provided an alternative way to represent multidimensional data. nELM achieved the best performance with a very low computational cost.

We already shown general properties of the methods and results for classification and regression tasks. Moreover, proposed methods can be used for several applications. sSOM can directly combine features and labels of different nature (categorical or numerical) making it a very versatile model. It could be trained for example to model target affects as engagement or boringness in conjunction with personal traits categories. The graphical representation provided by this model could be exploited in real-time for the communication of affects and personal states, where one can manage to see a relation between the input space and the labels. In addition, ELM proved to be as versatile as SVM with a simplistic framework, as the ELM algorithm for both regression and classification is very similar. It can also manage additional dimensions in the output just by adding an output unit. Moreover, these models have shown that it is possible to face affect recognition using only HRV. The possibility of using a physiological signal like this is promising for out-of-the-lab applications. With better performance on HRV signals and the advances of wearable technology, real-world HCI applications could be seen with such a simple equipment as a wrist-band or a distant web-cam.

In future works we will investigate ways to combine multi-rater information. The point-to-point agreement between raters may be an important clue to determine automatically the confidence around an affect estimation and improve training. Another topic for further research is to improve the dimensional estimation of valence, for which new methods for capturing the temporal dynamics should be explored.

Acknowledgments

This work was supported by the Agencia Nacional de Promoción Científica y Tecnológica (PICT 2015-0977, PICT 2014-1442), and Universidad Nacional del Litoral (PECAP 2014, CAID 2011-519/525), Argentina. RAC is supported by ARC Future Fellowship FT140100824.

GLOSARIO

Para mantener la consistencia terminológica de las publicaciones y facilitar la tarea al lector, en esta tesis se utilizaron las siglas de uso común en la bibliografía, en su mayoría proveniente del idioma inglés.

ASR arritmia sinusal respiratoria.

AV activación-valencia.

DNN red neuronal profunda.

ECG electrocardiograma.

EDA actividad electrodérmica.

EEG electroencefalograma.

ELM máquina de aprendizaje extremo.

EMG electromiograma.

HF alta frecuencia.

HMM modelos ocultos de Markov.

HR ritmo cardíaco.

HRV variabilidad cardíaca.

IADS sistema internacional de sonidos digitales.

IAPS sistema internacional de imágenes afectivas.

kELM máquina de aprendizaje extremo con función kernel.

kNN k-vecinos cercanos.

LF baja frecuencia.

LSTM redes con memoria a corto-largo plazo.

MLP perceptrón multicapa.

nELM máquina de aprendizaje extremo neuronal.

PCA análisis de componentes principales.

PPG fotopletismografía.

RF bosque aleatorio.

SAM maniquí de auto-apreciación.

SCL nivel de conductancia.

SCR respuesta de conductancia.

SNA sistema nervioso autónomo.

SOM mapa auto-organizativo.

sSOM mapa auto-organizativo supervisado.

STFT transformada de Fourier en tiempo corto.

SVM máquina de soporte vectorial.

SVR máquina de soporte vectorial para regresión.

TP temperatura de piel.

UAR sensibilidad media no ponderada.

REFERENCIAS

[1] R. A. Calvo and S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," *IEEE Transactions on Affective Computing*, vol. 1, pp. 18–37, jan 2010.

[2] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Physiology-based Affect Recognition For Computer-Assisted Intervention of Children with Autism Spectrum Disorder," *International Journal of Human-Computer Studies*, vol. 66, pp. 662–677, 2008.

[3] M. D. van der Zwaag, J. H. Janssen, and J. H. Westerink, "Directing Physiology and Mood through Music: Validation of an Affective Music Player," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 57–68, 2013.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[5] M. Soleymani and J. Lichtenauer, "A Multimodal Database For Affect Recognition and Implicit Tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[6] R. L. Mandryk, M. S. Atkins, and A. I. N. Press, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *International Journal of Human-Computer Studies*, vol. 65, pp. 329–347, apr 2007.

[7] L. E. Nacke, M. Kalyn, C. Lough, and R. L. Mandryk, "Biofeedback Game Design: Using Direct and Indirect Physiological Control to Enhance Game Interaction," in *ACM Conference on Human Factors in Computing Systems*, vol. 11, pp. 103–112, 2011.

[8] M. S. Hussain, O. AlZoubi, R. A. Calvo, and S. D'Mello, "Affect Detection from Multichannel Physiology during Learning Sessions with AutoTutor," in *International Conference on Artificial Intelligence in Education*, (Auckland, New Zealand), pp. 131–138, 2011.

[9] O. AlZoubi, S. D'Mello, and R. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 298–310, 2012.

[10] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions," *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, pp. 1–8, apr 2013.

[11] D. Kulić and E. Croft, "Affective state estimation for human–robot interaction," *IEEE Transactions on Robotics*, vol. 23, pp. 991–1000, oct 2007.

[12] F. Nasoz, K. Alvarez, C. L. Lisetti, and N. Finkelstein, "Emotion Recognition From Physiological Signals Using Wireless Sensors For Presence Technologies," *Cognition, Technology & Work*, vol. 6, no. 1, pp. 4–14, 2004.

[13] R. A. Calvo and D. Peters, *Positive computing: Technology for Wellbeing and Human Potential*. The MIT Press, 2014.

[14] E. M. Albornoz and D. H. Milone, "Emotion recognition in never seen languages using a novel ensemble method with emotion profiles," *IEEE Transactions on Affective Computing*, no. i, pp. 1–11, 2016.

[15] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods : Audio , Visual , and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[16] S. D'Mello and A. Graesser, "Automatic Detection of Learner's Affect From Gross Body Language," *Applied Artificial Intelligence*, vol. 23, no. 2, pp. 123–150, 2009.

[17] R. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[18] H. Gunes and H. Hung, "Is automatic facial expression recognition of emotions coming to a dead end? The rise of the new kinds on the block," *Image and Vision Computing*, vol. 55, no. 1, pp. 6–8, 2016.

[19] S. D. Kreibig, "Autonomic Nervous System Activity in Emotion: A Review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[20] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," *Lecture Notes in Computer Science*, vol. 4105 LNCS, pp. 530–537, 2006.

[21] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion Recognition from Brain Signals Using Hybrid Adaptive Filtering and Higher Order Crossings Analysis," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 81–97, 2010.

[22] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.

[23] M. Soleymani, G. Chanel, J. Kierkels, and T. Pun, "Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User's Physiological Emotional Responses," in *IEEE International Symposium on Multimedia*, vol. 5, pp. 228–235, 2008.

[24] C.-K. Wu, P.-C. Chung, and C.-J. Wang, "Representative Segment-Based Emotion Analysis and Classification with Automatic Respiration Signal Segmentation," *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 482–495, 2012.

[25] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG Pattern Analysis for Emotion Detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, 2012.

[26] W. Handouzi, C. Maaoui, A. Pruski, and A. Moussaoui, "Objective model assessment for short-term anxiety recognition from blood volume pulse signal," *Biomedical Signal Processing and Control*, vol. 14, pp. 217–227, nov 2014.

[27] S. Taylor, N. Jaques, Weixuan Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1934–1937, 2015.

[28] S. Ioannou, V. Gallese, and A. Merla, "Thermal infrared imaging in psychophysiology: Potentialities and Limits," *Psychophysiology*, jun 2014.

[29] M. Soleymani, M. Pantic, and T. Pun, "Multimodal Emotion Recognition in Response to Videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[30] B. Cheng and G.-Y. Liu, "Emotion recognition from surface EMG signal using wavelet transform and neural network," in *International Conference on Bioinformatics and Biomedical Engineering*, pp. 333–335, feb 2008.

[31] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *IMAVIS*, vol. 31, no. 2, pp. 120–136, 2013.

[32] P. Ekman and D. Cordaro, "What is Meant by Calling Emotions Basic," *Emotion Review*, vol. 3, no. 4, pp. 364–370, 2011.

[33] M. Csikszentmihalyi, "Beyond boredom and anxiety," *Book Reviews*, pp. 703–707, 1975.

[34] R. S. J. Baker, S. K. D'Mello, M. M. T. Rodrigo, A. C. Graesser, M. M. T. Rodrigoc, and A. C. Graesser, "Better to Be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners' Cognitive-Affective States during Interactions with Three Different Computer-Based Learning Environments," *International Journal of Human-Computer Studies*, vol. 68, no. 4, pp. 223–241, 2010.

[35] S. D'Mello and R. A. Calvo, "Beyond the basic emotions: what should affective computing compute?," in *CHI 2013 – Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Paris, France), pp. 2287–2294, 2013.

[36] J. Russell, "Core Affect and the Psychological Construction of Emotion," *Psychological Review*, vol. 110, no. 1, pp. 145–172, 2003.

[37] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A database for emotion analysis using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[38] S. Hussain, S. Chen, R. A. Calvo, and F. Chen, "Classification of Cognitive Load from Task Performance & Multichannel Physiology during Affective Changes," in *Conference on Multimodal Interaction*, pp. 1–4, 2011.

[39] K. Markov and T. Matsui, "Music genre and emotion recognition using Gaussian processes," *IEEE Access*, vol. 2, pp. 688–697, 2014.

[40] H. Martinez, Y. Bengio, and G. Yannakakis, "Learning Deep Physiological Models of Affect," *IEEE Computational Intelligence Magazine*, vol. 8, pp. 20–33, may 2013.

[41] G. N. Yannakakis, R. Cowie, and C. Busso, "The Ordinal Nature of Emotions," in *Affective Computing and Intelligent Interaction*, 2017.

[42] R. W. Picard, *Affective Computing*. MIT Press, 1997.

[43] A. Lichtenstein, A. Oehme, S. Kupschick, and T. Jürgensohn, "Comparing Two Emotion Models for Deriving Affective States from Physiological Data," *Affect and Emotion in Human-Computer Interaction*, pp. 35–50, 2008.

[44] R. A. Calvo and S. Mac Kim, "Emotions in text: Dimensional and categorical models," *Computational Intelligence*, vol. 29, no. 3, pp. 527–543, 2013.

[45] E. Albornoz, D. Milone, and H. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language*, no. i, pp. 1–38, 2011.

[46] R. Xia, S. Member, Y. Liu, and S. Member, "A Multi-task Learning Framework for Emotion Recognition Using 2D Continuous Space," vol. 3045, no. c, pp. 1–11, 2015.

[47] M. M. Bradley and P. J. Lang, "The international Affective Picture System (IAPS) in the Study of Emotions and Attention," in *Handbook of emotion elicitation and assessment*, ch. 2, pp. 29–46, Oxford University Press, 2007.

[48] M. Bradley and P. J. Lang, "The International affective digitized sounds (IADS): stimuli, instruction manual and affective ratings.," tech. rep., Center for the Study of Emotion and Attention, 1999.

[49] A. Deshmukh, S. Janarthanam, H. Hastie, M. Y. Lim, R. Aylett, and G. Castellano, "How expressiveness of a robotic tutor is perceived by children in a learning environment," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April, pp. 423–424, 2016.

[50] H. A. Osman and T. H. Falk, "Multimodal Affect Recognition: Current Approaches and Challenges," in *Emotion and Attention Recognition Based on Biological Signals and Images*, pp. 60–86, InTech, 2017.

[51] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[52] F. Mosciano, A. Mencattini, F. Ringeval, B. Schuller, E. Martinelli, and C. Di Natale, "An array of physical sensors and an adaptive regression strategy for emotion recognition in a noisy scenario," *Sensors and Actuators A: Physical*, vol. 267, pp. 48–59, nov 2017.

[53] M. J. Christie, "Electrodermal activity in the 1980s: a review," *Journal of the Royal Society of Medicine*, vol. 74, no. 8, pp. 616–22, 1981.

[54] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity.," *IEEE Transactions on Biomedical Engineering*, vol. 57, pp. 1243–52, may 2010.

[55] J. Cacioppo, L. G. Tassinary, and G. G. Berntson, *The Handbook of Phychophysiology*. Cambridge University Press, 3rd ed., 2007.

[56] J. Fleureau, P. Guillotel, and Q. Huynh-Thu, "Physiological-Based Affect Event Detector for Entertainment Video Applications," *Affective Computing, IEEE Transactions on*, vol. 3, no. 3, pp. 379–385, 2012.

[57] B. Cowley, M. Filetti, K. Lukander, J. Torniainen, A. Henelius, L. Ahonen, O. Barral, I. Kosunen, T. Valtonen, M. Huotilainen, N. Ravaja, and G. Jacucci, *The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human–Computer Interaction*, vol. 9. now publishers, 2016.

[58] D. Wu, C. G. Courtney, B. J. Lance, S. S. Narayanan, M. E. Dawson, K. S. Oie, and T. D. Parsons, "Optimal Arousal Identification and Classification for Affective Computing Using Physiological Signals: Virtual Reality Stroop Task," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 109–118, 2010.

[59] G. Valenza, A. Lanatà, and E. P. Scilingo, "The Role of Nonlinear Dynamics in Affective Valence and Arousal Recognition," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 237–249, 2012.

[60] I. T. Meftah and C. Ben Amar, "Multimodal Approach for Emotion Recognition Using an Algebraic Representation of Emotional States," in *International Conference on Signal Image Technology and Internet Based Systems*, pp. 541–546, 2012.

[61] V. Sharma, N. R. Prakash, and P. Kalra, "EDA wavelet features as Social Anxiety Disorder (SAD) estimator in adolescent females," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1843–1846, IEEE, sep 2016.

[62] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion Recognition: a review," *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, no. November 2014, pp. 410–415, 2011.

[63] F. Hönig, A. Batliner, and E. Nöth, "Fast Recursive Data-driven Multi-resolution Feature Extraction For Physiological Signal Classification," in *Russian-Bavarian Conference on Biomedical Engineering*, pp. 47–51, 2007.

[64] M. Kachele, P. Thiam, F. Schwenker, and M. Schels, "Ensemble Methods for Continuous Affect Recognition: Multi-modality, Temporality, and Challenges," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (New York, New York, USA), pp. 9–16, ACM, 2015.

[65] G. Andrássy, A. Szabo, G. Ferencz, Z. Trummer, E. Simon, and Á. Tahy, "Mental Stress May Induce QT-Interval Prolongation and T-Wave Notching," *Annals of Noninvasive Electrocardiology*, vol. 12, pp. 251–259, jul 2007.

[66] D. Lakens, "Using a Smartphone to Measure Heart Rate Changes during Relived Happiness and Anger," *IEEE Transactions on Affective Computing*, vol. 4, pp. 238–241, apr 2013.

[67] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics express*, vol. 18, pp. 10762–74, may 2010.

[68] N. D. Giardino, P. M. Lehrer, and R. Edelberg, "Comparison of finger plethysmograph to ECG in the measurement of heart rate variability," *Psychophysiology*, vol. 39, no. 2, pp. 246–53, 2002.

[69] A. Schäfer and J. Vagedes, "How accurate is pulse rate variability as an estimate of heart rate variability?: a review on studies comparing photoplethysmographic technology with an electrocardiogram.," *International journal of cardiology*, vol. 166, pp. 15–29, jun 2013.

[70] C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders," *Biomedical Signal Processing and Control*, vol. 6, pp. 261–268, jul 2011.

[71] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," *IEEE Transactions on Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, 1985.

[72] G. Lu, F. Yang, J. A. Taylor, and J. F. Stein, "A comparison of photoplethysmography and ECG recording to analyse heart rate variability in healthy subjects," *Journal of Medical Engineering & Technology*, vol. 33, pp. 634–641, nov 2009.

[73] J. Parak, A. Tarniceriu, P. Renevey, M. Bertschi, R. Delgado-Gonzalo, and I. Korhonen, "Evaluation of the beat-to-beat detection accuracy of PulseOn wearable optical heart rate monitor," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2015-Novem, pp. 8099–8102, 2015.

[74] L. A. Bugnon, R. A. Calvo, and D. H. Milone, "Dimensional Affect Recognition from HRV: an Approach Based on Supervised SOM and ELM," *IEEE Transactions on Affective Computing*, pp. 1–1, 2017.

[75] G. a. Reyes del Paso, W. Langewitz, L. J. M. Mulder, A. van Roon, S. Duschek, G. A. Reyes, D. E. L. Paso, W. Langewitz, L. J. M. Mulder, A. V. A. N. Roon, and S. Duschek, "The utility of low frequency heart rate variability as an index of sympathetic cardiac tone: a review with emphasis on a reanalysis of previous studies.," *Psychophysiology*, vol. 50, pp. 477–87, may 2013.

[76] A. Boardman, F. S. Schlindwein, A. P. Rocha, and A. Leite, "A study on the optimum order of autoregressive models for heart rate variability," *Physiological measurement*, vol. 23, pp. 325–36, may 2002.

[77] U. Rajendra Acharya, K. Paul Joseph, N. Kannathal, C. M. Lim, and J. S. Suri, "Heart rate variability: a review," *Medical & biological engineering & computing*, vol. 44, pp. 1031–51, dec 2006.

[78] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *International Journal of Human-Computer Studies*, vol. 66, no. 10, pp. 741–755, 2008.

[79] H. K. Lackner, E. M. Weiss, H. Hinghofer-szalkay, and I. Papousek, "Cardiovascular Effects of Acute Positive Emotional Arousal," *Applied psychophysiology and biofeedback*, vol. 39, pp. 9–18, oct 2013.

[80] B. M. Appelhans and L. J. Luecken, "Heart rate variability as an index of regulated emotional responding.," *Review of General Psychology*, vol. 10, no. 3, pp. 229–240, 2006.

[81] Y. Fu, H. V. Leong, G. Ngai, M. X. Huang, and S. C. Chan, "Physiological Mouse: Towards an Emotion-Aware Mouse," in *IEEE 38th International Computer Software and Applications Conference Workshops*, pp. 258–263, Ieee, jul 2014.

[82] J. A. Veltman and A. W. Gaillard, "Physiological workload reactions to increasing levels of task difficulty," *Ergonomics*, vol. 41, no. 5, pp. 656–69, 1998.

[83] A. Nesvold, M. W. Fagerland, S. Davanger, Ø. Ellingsen, E. E. Solberg, A. Holen, K. Sevre, and D. Atar, "Increased heart rate variability during nondirective meditation.," *European journal of preventive cardiology*, vol. 19, pp. 773–80, aug 2012.

[84] G. Parati, J. P. Saul, M. Di Rienzo, and G. Mancia, "Spectral Analysis of Blood Pressure and Heart Rate Variability in Evaluating Cardiovascular Regulation : A Critical Appraisal," *Hypertension*, vol. 25, pp. 1276–1286, jun 1995.

[85] T. Komatsu, S. Ohtsuka, K. Ueda, and T. Komeda, "Comprehension of Users Subjective Interaction States During Their Interaction with an Artificial Agent by Means of Heart Rate Variability Index," in *Affective Computing and Intelligent Interaction*, pp. 266–277, 2007.

[86] J. Kim and E. Andre, "Emotion-specific dichotomous classification and feature-level fusion of multichannel biosignals for automatic emotion recognition," *IEEE International conference on Multisensor Fusion and Integration for Intelligent Systems*, pp. 114–119, aug 2008.

[87] J. Bhattacharya, "Analysis and characterization of photo-plethysmographic signal," *... , IEEE Transactions on*, vol. 48, pp. 5–11, jan 2001.

[88] J. Choi and R. Gutierrez-Osuna, "Using heart rate monitors to detect mental stress," *Wearable and Implantable Body ...*, pp. 221–225, 2009.

[89] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob, "Classification of emotional States from electrocardiogram signals: a non-linear approach based on Hurst," *Biomedical Engineering Online*, vol. 12, p. 44, may 2013.

[90] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. Scilingo, "Recognizing Emotions Induced by Affective Sounds through Heart Rate Variability," *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–9, 2015.

[91] G. Valenza, P. Allegrini, A. Lanatà, and E. P. Scilingo, "Dominant Lyapunov exponent and approximate entropy in heart rate variability during emotional visual elicitation," *Frontiers in Neuroengineering*, vol. 5, 2012.

[92] E. Vyzas and R. W. Picard, "Affective pattern classification," in *Emotional and Intelligent: The Tangled Knot of Cognition*, pp. 176–182, 1998.

[93] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion Assessment From Physiological Signals for Adaptation of Game Difficulty," *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and humans*, vol. 41, no. 6, pp. 1052–1063, 2011.

[94] H. Martínez and G. Yannakakis, "Genetic search feature selection for affective modeling: a case study on reported preferences," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, pp. 15–20, 2010.

[95] B. Cheng, "Emotion Recognition from Physiological Signals Using AdaBoost," in *Communications in Computer and Information Science*, vol. 224, (Berlin, Heidelberg), pp. 412–417, 2011.

[96] G. Keren, T. Kirchstein, E. Marchi, F. Ringeval, and B. Schuller, "END-TO-END LEARNING FOR DIMENSIONAL EMOTION RECOGNITION FROMPHYSIOLOGICAL SIGNALS," No. edycja, pp. 4–6, 2017.

[97] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–13, 2015.

[98] O. Alzoubi, M. Hussain, S. D'Mello, and R. A. Calvo, "Affective Modeling from Multichannel Physiology: Analysis of Day Differences," in *Affective Computing and Intelligent Interaction*, pp. 4–13, 2011.

[99] E. Leon, I. Montalban, S. Schlatter, and I. Dorronsoro, "Computer-Mediated Emotional Regulation: Detection of Emotional Changes Using Nonparametric Cumulative Sum," in *International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 2010, pp. 1109–1112, jan 2010.

[100] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "The AV + EC 2015 Multimodal Affect Recognition Challenge: Bridging Across Audio, Video, and Physiological Data," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (New York, New York, USA), pp. 3–8, ACM, 2015.

[101] J. Wagner, J. Kim, and E. André, "From Physiological Signals to Emotions: Implementing and Comparing Selected Methods for Feature Extraction and Classification," in *IEEE International Conference on Multimedia and Expo*, pp. 940–943, 2005.

[102] A. Heraz and C. Frasson, "Predicting the three major dimensions of the learner's emotions from brainwaves," *Proceedings of World Academy of Science, Engineering and Technology*, vol. 25, pp. 323–329, 2007.

[103] A. Barreto, J. Zhai, and M. Adjouadi, "Non-intrusive Physiological Monitoring for Automated Stress Detection in Human-Computer Interaction," *Human–Computer Interaction*, vol. 4796, pp. 29–38, 2007.

[104] C. Liu, P. Agrawal, N. Sarkar, and S. Chen, "Dynamic Difficulty Adjustment in Computer Games Through Real-Time Anxiety-Based Affective Feedback," *International Journal of Human-Computer Interaction*, vol. 25, pp. 506–529, aug 2009.

[105] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion Recognition Using Biosensors: First Steps Towards an Automatic System," in *Affective Dialogue Systems*, pp. 36–48, 2004.

[106] E. Leon, G. Clarke, V. Callaghan, and F. Sepulveda, "A user-independent real-time emotion recognition system for software agents in domestic environments," *Engineering Applications of Artificial Intelligence*, vol. 20, pp. 337–345, apr 2007.

[107] S. Yang and G. Yang, "Emotion Recognition of EMG Based on Improved L-M BP Neural Network and SVM," *Journal of Software*, vol. 6, no. 8, pp. 1529–1536, 2011.

[108] O. AlZoubi, R. A. Calvo, and R. H. Stevens, "Classification of EEG for affect recognition: an adaptive approach," in *Advances in Artificial Intelligence*, pp. 52–61, 2009.

[109] P. A. Pour, M. S. Hussain, O. AlZoubi, S. K. D'Mello, and R. A. Calvo, "The Impact of System Feedback on Learners' Affective and Physiological States," in *Intelligent Tutoring Systems*, vol. 6094, pp. 264–273, 2010.

[110] M. Valstar, J. Gratch, F. Ringeval, M. T. Torres, S. Scherer, and R. Cowie, "AVEC 2016 – Depression, Mood, and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge.*, (Chicago), pp. 3–10, ACM Press, 2016.

[111] B. Sun, S. Cao, L. Li, J. He, and L. Yu, "Exploring Multimodal Visual Features for Continuous Affect Recognition," in *AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 83–88, 2016.

[112] R. Weber, V. Barrielle, C. Soladié, and R. Séguier, "High-Level Geometry-based Features of Video Modality for Emotion Prediction," *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pp. 51–58, 2016.

[113] J. Scheirer, R. Fernandez, J. Klein, and R. W. Picard, "Frustrating the user on purpose: a step toward building an affective computer," *Interacting with Computers*, vol. 14, no. 2, pp. 93–118, 2002.

[114] Z. Liu and S. Wang, "Emotion Recognition Using Hidden Markov Models from Facial Temperature Sequence," in *Affective Computing and Intelligent Interaction*, pp. 240–247, 2011.

[115] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer Methods and Programs in Biomedicine*, vol. 140, pp. 93–110, 2017.

[116] P. Cardinal, N. Dehak, A. Lameiras, J. Alam, and P. Boucher, "ETS System for AV+EC 2015 Challenge," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (New York, New York, USA), pp. 17–23, 2015.

[117] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 73–80, ACM, 2015.

[118] A. Manandhar, K. D. Morton, P. A. Torrione, and L. M. Collins, "Multivariate Output-Associative RVM for Multi-Dimensional Affect Predictions," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 10, no. 3, pp. 365–372, 2016.

[119] G. Valenza, A. Lanatà, and E. P. Scilingo, "Oscillations of heart rate and respiration synchronize during affective visual stimulation," *IEEE Transactions on Information Technology in Biomedicine*, vol. 16, pp. 683–90, jul 2012.

[120] J. Lázaro, E. Gil, J. M. Vergara, and P. Laguna, "Pulse rate variability analysis for discrimination of sleep-apnea-related decreases in the amplitude fluctuations of pulse photoplethysmographic signal in children," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 1, pp. 240–246, 2014.

[121] W. Schuller, "Acquisition of Affect," in *Emotions and Personality in Personalized Services*, pp. 57–80, Springer International Publishing, 2016.

[122] A. Accardo, G. Addio, D. Maestri, D. Vitale, G. Furgi, and F. Rengo, "Fractal dimension and power-law behavior reproducibility and correlation in chronic heart failure patients," in *Signal Processing Conference*, 2002.

[123] L. A. Bugnon, R. A. Calvo, and D. H. Milone, "A Method for Daily Normalization in Emotion Recognition," in *15th Argentine Symposium on Technology, AST 2014, 43º JAIIO*, pp. 48–59, 2014.

[124] T. Kohonen, "Essentials of the self-organizing map," *Neural Networks*, vol. 37, pp. 52–65, 2013.

[125] G. Huang, G. B. Huang, S. Song, and K. You, "Trends in extreme learning machines: A review," *Neural Networks*, vol. 61, pp. 32–48, 2015.

[126] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1-3, pp. 489–501, 2006.

[127] A. Ben-Israel and T. N. E. Greville, *Generalized inverses: theory and applications*. Springer, 2 ed., 2001.

[128] H. Monkaresi and R. A. Calvo, "Feasibility of a low-cost platform for physiological recording in affective computing applications," in *The 10th International Conference on Body Area Networks (BodyNets 2015)*., 2015.

[129] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 4 ed., 2014.

[130] L. I.-k. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility Author(s): Lawrence I-Kuei Lin Source:," *Biometrics*, vol. 45, no. 1, pp. 255–268, 1989.

[131] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-p. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data," *Pattern Recognition Letters*, 2014.

[132] S. Chen and Q. Jin, "Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (New York, New York, USA), pp. 49–56, ACM, 2015.

[133] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen, "Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition," in *AVEC '15 Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (New York, New York, USA), pp. 65–72, ACM, 2015.

[134] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-Modal Audio , Video and Physiological Sensor Learning for Continuous Emotion Prediction 1," in *AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 97–104, 2016.

[135] F. Povolny, P. Matejka, M. Hradis, A. Popkova, L. Otrusina, and P. Smrz, "Multimodal Emotion Recognition for Avec 2016 Challenge," in *AVEC '16 Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 75–81, 2016.

[136] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion," *Journal of Personality and Social Psychology*, vol. 53, no. 4, pp. 712–7, 1987.

[137] O. Alzoubi, *Automatic Affect Detection From Physiological Signals: Practical Issues*. PhD thesis, University of Sydney, 2012.

[138] E. Vyzas, *Recognition of Emotional and Cognitive States Using Physiological Data*. PhD thesis, Massachusetts Institute of Technology, 1999.

[139] J. A. Healey, *Weareable and Automotive Systems for Affect Recognition from Physiology*. PhD thesis, Massachusetts Institute of Technology, 2000.

[140] E. Leon, G. Clarke, V. Callaghan, F. Sepulveda, A. I. N. Press, E. Leon, G. Clarke, V. Callaghan, F. Sepulveda, A. I. N. Press, E. Leon, G. Clarke, V. Callaghan, and F. Sepulveda, "Real-time detection of emotional changes for inhabited environments," *Computers & Graphics*, vol. 28, pp. 635–642, oct 2004.

[141] D. Kahneman, *Attention and Effort*. Prentice-Hall, jun 1973.

[142] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[143] J. Cohen, "A Coefficiente of Agreement for Nominal Scales," *Educational and Psychological Measurement*, no. 1, pp. 37–46, 1960.

[144] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.

[145] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3, pp. 169–200, 1992.

[146] S. H. Fairclough, "Fundamentals of physiological computing," *Interacting with Computers*, vol. 21, no. 1-2, pp. 133–145, 2009.

[147] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, R. Barbieri, G. Valenza, L. Citi, and A. Lanata, "Revealing Real-Time Emotional Responses: a Personalized Assessment based on Heartbeat Dynamics," *Scientific reports*, vol. 4, no. 4998, p. 4998, 2014.

[148] R. Khosrowabadi, H. C. Quek, A. Wahab, and K. K. Ang, "EEG-based Emotion Recognition Using Self-Organizing Map for Boundary Detection," *International Conference on Pattern Recognition*, pp. 4242–4245, aug 2010.

[149] R. Khosrowabadi, C. Quek, K. K. Ang, and A. Wahab, "ERNN: A Biologically Inspired Feedforward Neural Network to Discriminate Emotion From EEG Signal," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 3, pp. 609–620, 2014.

[150] R. W. Picard, S. Fedor, and Y. Ayzenberg, "Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry," *Emotion Review*, vol. 8, no. 1, pp. 62–75, 2016.

[151] J. Hernandez, I. Riobo, A. Rozga, G. Adowd, and R. Picard, "Using electrodermal activity to recognize ease of engagement in children during social interactions," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2014.

[152] C. D. Katsis, N. S. Katertsidis, and D. I. Fotiadis, "An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders," *Biomedical Signal Processing and Control*, vol. 6, pp. 261–268, jul 2011.

[153] D. Novak, M. Mihelj, and M. Munih, "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing," *Interacting with Computers*, vol. 24, pp. 154–172, may 2012.

[154] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals," *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 126–140, 2014.

[155] Z. Zhang, Z. Pi, and B. Liu, "TROIKA: A General Framework for Heart Rate Monitoring Using Wrist-Type Photoplethysmographic Signals During Intensive Physical Exercise," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 2, pp. 522–531, 2015.

[156] J. Hernandez, D. J. McDuff, and R. W. Picard, "Biophone: Physiology monitoring from peripheral smartphone motions," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 7180–7183, IEEE, 2015.

[157] M. A. Quiros-Ramirez, S. Polikovsky, Y. Kameda, and T. Onisawa, "Towards developing robust multimodal databases for emotion analysis," in *6th International Conference on Soft Computing and Intelligent Systems, and 13th International Symposium on Advanced Intelligence Systems, SCIS/ISIS 2012*, pp. 589–594, 2012.

[158] G. B. Huang, "An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels," *Cognitive Computation*, vol. 6, no. 3, pp. 376–390, 2014.

[159] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification.," *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, vol. 42, no. 2, pp. 513–29, 2012.

[160] G. Stegmayer, C. Yones, L. Kamenetzky, and D. H. Milone, "High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, no. APRIL, pp. 1–1, 2016.

[161] H. Monkaresi, R. a. Calvo, and H. Yan, "A machine learning approach to improve contactless heart rate monitoring using a webcam.," *IEEE journal of biomedical and health informatics*, vol. 18, pp. 1153–60, jul 2014.

[162] H. Monkaresi, N. Bosch, R. A. Calvo, and S. K. D. Mello, "Automated Detection of Engagement using Video-Based Estimation of Facial Expressions and Heart Rate," *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–14, 2016.

[163] M. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, p. 933, 2011.

[164] B. Schuller and F. Weninger, "Ten recent trends in computational paralinguistics," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7403 LNCS, pp. 35–49, 2012.

[165] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. D. Natale, "Continuous Estimation of Emotions in Speech by Dynamic Cooperative Speaker Models," *IEEE Transactions on Affective Computing*, vol. 3045, no. c, pp. 1–14, 2016.

[166] J. N. Bailenson, E. D. Pontikakis, I. B. Mauss, J. J. Gross, M. E. Jabon, C. a.C. Hutcherson, C. Nass, and O. John, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *International Journal of Human-Computer Studies*, vol. 66, no. 5, pp. 303–317, 2008.

[167] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks," in *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH), to be published*, (San Francsico, CA), 2016.

[168] M. Robnik-Sikonja and I. Kononenko, "Theoretical and Empirical Analysis of ReliefF and RReliefF," *Machine Learning Journal*, vol. 53, pp. 23–69, 2003.

[169] J. Gruber, D. S. Mennin, A. Fields, A. Purcell, and G. Murray, "Heart rate variability as a potential indicator of positive valence system disturbance: A proof of concept investigation," *International Journal of Psychophysiology*, vol. 98, no. 2, pp. 240–248, 2015.

[170] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition," *International Journal of Synthetic Emotions*, vol. 1, pp. 68–99, jan 2010.

[171] M. A. Nicolaou, H. Gunes, and M. Pantic, "Output-associative RVM regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, pp. 186–196, mar 2012.