Teoría de la Educación. Educación y Cultura en la Sociedad de la Información
E-ISSN: 1138-9737
revistatesi@usal.es
Universidad de Salamanca
España

Diego, Isaac Martín de; Serrano, Ángel; Conde, Cristina; Cabello, Enrique
TÉCNICAS DE RECONOCIMIENTO AUTOMÁTICO DE EMOCIONES
Teoría de la Educación. Educación y Cultura en la Sociedad de la Información, vol. 7, núm. 2, diciembre, 2006, pp. 107-127
Universidad de Salamanca
Salamanca, España

TECNICAS DE RECONOCIMIENTO AUTOMÁTICO DE EMOCIONES

En este artículo presentamos un resumen de las principales técnicas para el reconocimiento automático de emociones. Repasaremos brevemente los principales tipos de emociones, haciendo hincapié en la necesidad de usar emociones específicas. Consideraremos los dos principales canales para el estudio de las emociones: las expresiones faciales obtenidas a partir de un video y las expresiones léxico-fonéticas obtenidas de un discurso de audio. A pesar de que las emociones recogidas en uno u otro medio pueden ser diferentes, las técnicas de investigación más novedosas se centran en la combinación de ambas fuentes de información. Presentamos un resumen de las técnicas de análisis estadístico utilizadas en la detección de emociones. El desarrollo de una base de datos apropiada para el estudio de las emociones es una tarea fundamental. El material de dicha base de datos debería de ser audiovisual y recogido en un entorno real, sin el uso de personas predispuestas a reflejar una emoción.

Palabras Clave: Reconocimiento, Emociones, Discurso, Expresiones Faciales, Combinación de Información.

TECHNIQUES FOR THE AUTOMATIC RECOGNITION OF EMOTIONS

In this paper we present a summary of the main techniques for the automatic recognition of human emotions. We study the most important kind of emotions, centering on the specific emotions. We focus on the two most relevant ways in which emotions are studied: facial expressions from a video, and phonetic expressions from a speech. Since the type of emotions obtained from these two approaches could be different, the fusion of these different sources of information is one of the most interesting areas of research. We present a review of the statistical techniques used in emotion recognition. A fundamental task is the development of an appropriate database, with real audio-visual data, focus on emotion learning.

Keywords: Recognition, Emotion, Speech, Facial Expressions, Combination of Information.

TECHNIQUES DE RECONNAISSANCE AUTOMATIQUES DES ÉMOTIONS

Nous présentons dans cet article un résumé des principales techniques relatives à la reconnaissance automatique des émotions. Nous résumerons brèvement les principaux types d'émotions en soulignant le besoin d'utiliser des émotions spécifiques. Nous considèrerons les deux principales voies de l'étude des émotions: les expressions faciales obtenues à partir d'un vidéo et les expressions léxico-phonétiques obtenues d'un discours de matériel audio. Bien que les émotions recueillies par l'un ou l'autre des procédés puissent être différentes, les plus récentes techniques de recherche se centrent sur la combinaison des deux sources d'information. Nous présentons un résumé des techniques d'analyse statistique utilisées lors de l'étude des émotions. Le développement d'une base de données adaptée á l'étude des émotions est un travail fondamental. Le matériel de cette base de données devrait être audiovisuel et recueilli dans un milieu réel, sans la présence de personnes prédisposées à refléter une émotion.

Mots-clé : Reconnaissance, Emotions, Discours, Expressions Faciales, Combinaison d'information

TÉCNICAS DE RECONOCIMIENTO AUTOMÁTICO DE EMOCIONES

Autores: Isaac Martín de Diego, Ángel Serrano, Cristina Conde, Enrique Cabello
Email: isaac.martin@urjc.es
Institución: Universidad Rey Juan Carlos (ESCET),

1.- INTRODUCCIÓN.

Uno de los últimos objetivos en la interacción hombre-máquina consiste en conseguir una comunicación natural y sin esfuerzo. Las nuevas tecnologías permiten que, más allá de la interacción existente a través del teclado y del ratón, surjan nuevas modalidades de comunicación entre un ordenador y el usuario que lo maneja. Dos de los canales más informativos para la percepción de emociones por parte de una máquina son las expresiones faciales obtenidas a partir de un video y las expresiones léxico-fonéticas obtenidas de un discurso. Para que el computador consiga establecer una interacción adecuada ha de ser capaz de tener alguna percepción del estado emocional del ser humano con el que interacciona, esto es, ha de comprender sus emociones.

Las emociones rigen casi todos los modos de comunicación humana: las expresiones faciales, los gestos, las posturas, el tono de voz, la elección de las palabras, la respiración, la temperatura corporal, etc. Las emociones cambian el mensaje transmitido, y en ocasiones lo importante no es el mensaje en sí, sino el modo en que este mensaje es transmitido. Sin embargo, no existe una teoría unificada que describa apropiadamente todo el amplio rango de emociones. Clasificar las emociones en un número pequeño de categorías primarias (emociones puras) es, obviamente, limitado. En general, es preferible caracterizar las emociones en dimensiones continuas (por ejemplo, negativo vs. positivo). De este modo se obtiene una representación más flexible y generalizable. La predicción de las posibles acciones a llevar a cabo a partir de estados emocionales debería de ser el centro de la tarea de comprensión de las emociones. Cualquier sistema cuyo fin último sea la interacción con humanos ha de basarse en esa predicción. En este artículo se presentan los aspectos a considerar en el desarrollo de un sistema automático de reconocimiento de emociones. En primer lugar consideramos los tipos de emociones, sus métodos de detección y de evaluación. Posteriormente nos centraremos en las emo-

ciones recogidas a través del rostro y en aquellas que lo son a través del discurso. Haremos un repaso a las principales técnicas de análisis matemático para el reconocimiento de emociones. Comentaremos brevemente las bases de datos disponibles para el entrenamiento de dichas técnicas. Finalizaremos con las principales conclusiones.

1.1 Definición y tipos de emociones.

La Real Academia Española de la lengua define emoción como la "alteración del ánimo intensa y pasajera, agradable o penosa, que va acompañada de cierta conmoción somática". El carácter intenso y momentáneo de estas alteraciones hace difícil su detección por sistemas automáticos. Hemos de considerar además, cuales son las emociones estamos interesados en detectar. Un sistema diseñado para detectar alteraciones independientemente del carácter de las mismas será mucho más sencillo de implementar y aplicar que un sistema cuyo fin sea la determinación efectiva del signo de la emoción, positivo, negativo, neutro... etc. Es posible clasificar las emociones en unas pocas categorías primarias, concretamente las seis grandes emociones (definidas por Cornelius, 1996 y previamente por Ekman, 1978 y Ekman y Friesen, 1978) son: felicidad, sorpresa, miedo, disgusto, cólera, tristeza. Para codificar las expresiones faciales, Ekman y Friesen (1978) desarrollaron un sistema de codificación de acciones faciales (FACS, "Facial Action Coding System") en el que los movimientos de la cara se describen mediante variables asociadas a los movimientos musculares. Los trabajos de Ekman motivaron el trabajo de un gran número de investigadores en el campo del análisis de expresiones faciales a partir de imágenes y video. Las seis grandes emociones definidas previamente, son consideradas primarias o básicas pues todas las otras emociones pueden ser derivadas a partir de ellas. Sin embargo, es lícito plantearse si esas emociones básicas son representativas de la realidad que se desea estudiar. Habitualmente es preferible emplear una caracterización basada en emociones definidas sobre una escala continua. En este último caso puede ser complejo establecer el punto exacto en el que se pasa de una emoción a otra en la escala continua.

1.2 Métodos de evaluación de las emociones.

Describir las relaciones entre el discurso o las expresiones faciales y las emociones depende de la identificación adecuada de las técnicas para describir los estados emocionales de los individuos estudiados. En principio existen varios modos de describir estos estados, sin embargo los métodos más sencillos no suelen ser los más eficientes. Cowie (Cowie, 2000) presentó varios métodos para describir de manera eficiente estas emociones. Se consideran las categorías emocionales básicas anteriormente descritas y se añaden las denominadas categorías emocionales secundarias. Estas últimas emociones son más complejas que las emociones básicas y sirven para extenderlas y complementarlas. Una de las opciones consiste en explotar el poder del lenguaje emocional al máximo mediante el uso de un amplio rango de emociones secundarias, incluso combinándolas dando lugar a emociones complejas imposibles de representar con las simples emociones básicas. Un método alternativo a esta representación es la llamada representación

biológica que consiste en la búsqueda de la estructura inherente a las diferentes categorías emocionales. Otra alternativa es el empleo de representaciones continuas. En este caso se asume que el dominio de estados emocionales corresponde a unas coordenadas en un espacio con un número pequeño de dimensiones. A partir de esta idea es posible derivar una representación simple y capaz de capturar un amplio rango de emociones. Dicha representación se denomina espacio de activación-evaluación. Además, se ha desarrollado un algoritmo para la evaluación del contenido emocional de un estímulo obtenido a partir de un discurso (Cowie et al., 2000).

Enunciemos algunas de las características del algoritmo: existe una translación natural entre los estados intermedios, las diferentes categorías emocionales están claramente representadas, el formato numérico obtenido hace que el tratamiento estadístico de las emociones sea mucho más sencillo, y por último, la habilidad para capturar variaciones emocionales en el tiempo, tanto de manera inmediata como a largo plazo. Otro modo alternativo de representación de las emociones se basa en la conceptualización de la comunicación del afecto a partir de tres dimensiones: activación o vigilia, placer y poder. Este concepto de las dimensiones de la emoción ha sido útil para describir y distinguir entre emociones (Pereira, 2000).

1.3 Identificación de variables para detectar emociones.

Idealmente la representación de una emoción no debería de ser una tarea meramente descriptiva, sino que debe tener una clara vocación predictiva. Es decir, a la hora de elegir las variables más adecuadas para representar emociones hemos de considerar como aspecto más relevante el objetivo final y éste no es otro que construir un mecanismo eficiente para la detección automática de emociones.

Se han realizado algunos trabajos que seleccionan, a partir de una colección de variables originales aquellas que mejor discriminan entre las diferentes emociones, como paso previo al empleo de un algoritmo de aprendizaje de las emociones (Bartlett, 2003). Además, la representación de una emoción debería de ser adaptable a lo largo del tiempo, influenciada por el propio aprendizaje.

Un aspecto a considerar cuando se recogen emociones es la espontaneidad de las mismas. La mayor parte de las bases de datos disponibles para el estudio de las emociones están construidas ad hoc, esto es, los gestos o el discurso asociado a cada una de las emociones están representados por personas que, por indicación, gesticulan o hablan de acuerdo a la emoción que les ha sido indicada. Obviamente, no existe garantía de que el discurso o la gesticulación de esas personas sea la misma información que recogeríamos de modo espontáneo (ver Campbell, 2000, para una revisión al respecto).

2.- ANÁLISIS DE AUDIO.

Existe evidencia de que un amplio rango de características lingüísticas tienen significado emocional. Estas características pueden ser tanto fonéticas (o acústicas) como sintácticas (o léxicas). Se necesita describir de modo adecuado los estados emocionales que aparecen en el discurso. Como indicamos anteriormente, uno de los problemas importantes en el reconocimiento de emociones es el modo de identificar de manera eficiente las características que definen las diferentes emociones y las relaciones entre ellas. Veamos a continuación cuales son los parámetros que determinan estas características en el discurso.

2.1 Parámetros del discurso y emociones específicas.

Existen ciertas variables que, por su propia definición, son específicas del discurso y no las encontraremos, o serán más difíciles de detectar en el gesto o en la representación facial del individuo. Por ejemplo, variaciones de la entonación pueden ser debidas a cambios en el estado emocional del individuo. La tabla 1 presenta un resumen de las relaciones entre las emociones y los parámetros del discurso (Murray y Arnott, 1993).

Tabla 1. Emociones y parámetros del discurso.

	Ira	Felicidad	Tristeza	Miedo	Disgusto
Velocidad	Ligeramente acelerada	Acelerada o retardada	Pausada	Muy acelerada	Mucho más acelerada
Variación	Muy alta	Alta	Ligeramente baja	Muy alta	Muy baja
Rango	Amplio	Amplio	Estrecho	Amplio	Amplio
Respiración	Acompasada	Acompasada	Resonante	Irregular	Refunfuñando
Intensidad	Alta	Alta	Baja	Normal	Baja
Articulación	Tensa	Normal	Pausada	Precisa	Normal
Calidad de la voz	Procedente del pecho	Estridente	Resonante	Irregular	Retumbante

Existe en general una relación conocida entre el discurso y las emociones primarias. Las medidas del discurso que parecen ser buenas indicadoras de estas emociones son medidas acústicas continuas, tales como las relacionadas con la variación del discurso, el rango, la intensidad y la duración del mismo. Sin embargo esta relación suele no ser suficiente. Una de las líneas de investigación en el reconocimiento automático de emociones es la mejora de nuestra capacidad para identificar la correlación entre las señales acústicas en el discurso y el amplio rango de emociones producidos por el hablante. Los sistemas diseñados para llevar a cabo esta tarea, por lo general, son extremadamente sensibles a la variabilidad introducida por el hablante. Esta variabilidad se debe, especialmente a variaciones en la voz y en estilo causadas por ejemplo por diferentes estados de ánimo del hablante.

2.2 Extracción automática de variables fonéticas.

Como se indicó, la percepción humana de las emociones asociadas al discurso proviene de múltiples variables. Una de las principales tareas en la detección automática de emociones es la recuperación automática de las variables más relevantes. El nivel de la voz es uno de los indicadores de la emoción. Un modo natural de medir este nivel de voz es mediante una función del voltaje recogido del micrófono. Sin embargo no suele existir una relación directa entre el nivel de voz y dicha función del voltaje, debido fundamentalmente a las numerosas variables que influyen en los valores de esta última. Es posible medir la variación en la frecuencia de la vibración. Este factor es un buen indicador estadístico de algunos tipos de discurso. Otro factor importante, y fácil de conseguir, es la duración del discurso. Las variaciones en la intensidad del discurso sirven como variables para determinar el tipo de discurso: por ejemplo, la intensidad se desplaza hacia las vocales en los discursos sonoros. La calidad de la voz puede marcar las diferencias entre unas emociones y otras. Existen numerosas variables fonéticas relacionadas con la calidad de la voz (Izzo, 1993): cociente de apertura de las cuerdas vocales, timbre de la voz, ruido, distribución de la energía, etc. Se ha intentado modelar el ritmo del discurso aunque esta tarea es más compleja. Es común el empleo de técnicas de reconocimiento de palabras en el discurso para la detección de aquellas palabras con claro significado emocional.

3.- ANÁLISIS DE VIDEO.

Los gestos faciales tienen un claro significado emocional. Sin embargo, el uso de imágenes estáticas limita la capacidad de transmisión de emociones. Por contra, las aproximaciones dinámicas han producido buenos resultados en la práctica.

3.1 Detección de la cara.

La gran cantidad de información geométrica disponible en una imagen hace posible la estimación de la posición de la cara y su posterior normalización, pero las diferencias entre los individuos, los cambios de expresión facial, las oclusiones o áreas con datos no adquiridos, etc. contribuyen a aumentar la complejidad de esta tarea. Estos factores hacen necesario que todo sistema de reconocimiento emocional a través del rostro incluya en primer lugar una etapa de normalización de la posición facial. Esta tarea ha sido a menudo realizada manualmente, pero el incremento del tamaño de las bases de datos, y la necesidad de un alto grado de precisión, han provocado que actualmente exista un gran interés en esta área. Existen varias aproximaciones para la detección de la cara (ver por ejemplo: Crowley y Berard, 1997; Collobert et al., 1996; Graf, 1996). Recientemente se han presentado varios métodos de normalización (Conde, 2006).

3.2 Representación de la cara

Existen diferentes tipos de representación de la cara, por ejemplo: imagen en color (2D), imagen de rango (2.5D) y mallado tridimensional (3D). La gran mayoría de los sistemas de verificación facial tanto a nivel comercial como de investigación usan imágenes de color o de intensidad. Estas imágenes son referenciadas como imágenes 2D o de textura, en contraposición a las representaciones faciales tridimensionales, a las que se denomina modelos o imágenes 3D. Existen diferentes maneras de representar la información tridimensional. Por un lado simplemente pueden representarse como una nube de puntos en el espacio. Estos puntos pueden representarse también en forma de mallado, donde cada punto es un vértice y a su vez los vértices están unidos por artistas. El polígono o geometría a utilizar para realizar esta unión es habitualmente un triángulo, ya que es la manera más simple de aproximar una nube de puntos a una superficie. Otra aproximación diferente muy utilizada son los mapas de profundidad o imágenes de rango. Son imágenes en escala de grises donde el nivel de intensidad de cada píxel está relacionado con la coordenada de profundidad z de un punto tridimensional. Como se trata de una representación bidimensional de información tridimensional se denomina a menudo imagen 2.5D. También es muy habitual disponer de información tridimensional registrada con la imagen de textura. En la Figura 1 puede verse un ejemplo de estos diferentes tipos de información facial tridimensional.

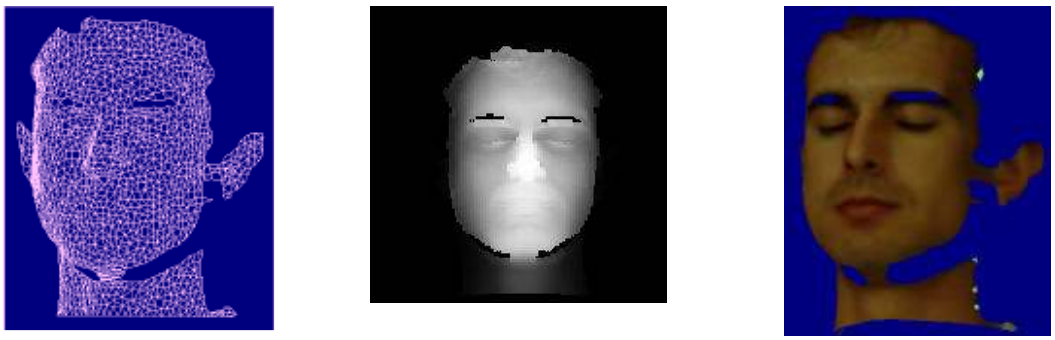

Figura 1. Diferentes modos de representación de la información facial en 3D. De izquierda a derecha: mallado, mapa de profundidad, mallado+textura.

Cualquiera de estas representaciones, u otras más complejas que pudieran aparecer, son empleadas para su uso posterior en un sistema de clasificación que tenga como objetivo la detección automática de emociones.

3.3 Reconocimiento de emociones en expresiones faciales.

Como ya se comentó apartados anteriores, P. Ekman (1982) fue uno de los pioneros del reconocimiento de emociones mediante el análisis de expresiones faciales, encontrando evidencias que soportan la universalidad de este tipo de expresiones, definiendo las seis grandes emociones básicas previamente definidas. Los sistemas de reconocimiento de emociones pueden verse como complemento de los métodos de reconocimiento de ca-

ras, ignorándose en este caso la personalidad de la persona y centrándose en la expresión de su rostro. La mayor parte del análisis de emociones basado en expresiones faciales se realiza a partir de imágenes estáticas. Sin embargo, esto no suele ser suficiente. La razón fundamental radica en la naturaleza dinámica de las emociones faciales, que pueden ser obtenidas a partir de una secuencia de imágenes. Los sistemas dinámicos han producido resultados prometedores. Estos sistemas se dividen en tres clases fundamentales: aproximaciones basadas en el flujo óptico, rastreo de características y aproximaciones basadas en el alineamiento del modelo. La aproximación basada en el flujo óptico usa campos de movimiento densos calculados en áreas específicas de la cara tales como la boca y los ojos. Intenta relacionar los vectores de movimiento con las emociones faciales usando plantillas de movimiento extraídas sobre un conjunto de campos de movimiento de entrenamiento (Tsapatsoulis, 1999; Otsuka, 1997). En la segunda aproximación la estimación de la emoción se obtiene a partir de un conjunto pequeño de características relevantes en la escena. El análisis se realiza en dos etapas: en primer lugar se procesa el frame del video para la detección de las características necesarias, (los ojos, la nariz, la boca...), posteriormente se analiza el movimiento de dichos elementos (Lien et al, 1998; Kulas y Kanade, 1981). La tercera aproximación alinea un modelo 3-D de la cara para estimar tanto el movimiento del objeto como la orientación (ver Essa y Pentland 1993,1995 y Essa et al, 1995).

4.- TÉCNICAS DE ANÁLISIS.

En este apartado se abordarán los aspectos más relevantes de las principales técnicas empleadas para el reconocimiento automático de emociones en audio y en video. Se presentarán ejemplos prácticos de la aplicación de estas técnicas en casos reales.

4.1 Técnicas estadísticas para el reconocimiento de emociones.

Métodos Supervisados de Clasificación

Podemos entender la clasificación como el proceso de asignar objetos a un conjunto prefijado de categorías o clases. En el caso que nos atañe, estas categorías vienen determinadas por las emociones que queremos detectar a partir de una serie de características medidas sobre los individuos. Cuando conocemos a priori el conjunto de clases (emociones), el proceso de clasificación recibe el nombre de clasificación supervisada. Si el conjunto de clases nos es desconocido, el proceso recibe el nombre de clasificación no supervisada. Este método será considerado posteriormente.

Si existe una función desconocida que va del espacio de características al espacio de emociones, nos referimos a ella como función objetivo. La solución del problema de clasificación será una estimación de la función objetivo, un clasificador. Por lo tanto, el problema de reconocimiento automático de emociones puede ser descrito como sigue: dada una muestra de objetos (caras, discursos, etc…), encontrar una función que asigne cada objeto a una de las emociones predefinidas, de modo que se minimice el error

promedio de clasificación para futuras observaciones. El algoritmo de clasificación se conoce también con los nombres de proceso de aprendizaje, reconocimiento de patrones o discriminación. Existen dos etapas básicas en el diseño de un clasificador: la fase de entrenamiento y la fase de validación. En la fase de entrenamiento empleamos los datos muestras, llamados en este contexto, muestra de entrenamiento. Pueden imponerse restricciones sobre el clasificador, generalmente relativas a hipótesis sobre la distribución de las observaciones. Una vez construido el modelo, dichas hipótesis han de ser contrastadas. Empleamos la muestra de entrenamiento, considerando las restricciones si las hubiera, para construir el clasificador. Una vez disponemos de una regla de clasificación que asigna los objetos a las emociones, pasamos a la fase de validación. En esta fase, el clasificador obtenido en la fase de entrenamiento es empleado para clasificar las observaciones pertenecientes a la muestra de validación. El proceso de aprendizaje aparece resumido en la figura 2.

Tanto en la fase de entrenamiento como en la fase de validación, el clasificador asigna cada observación a una emoción. La suma del número de objetos de la muestra de entrenamiento que no son correctamente clasificados, es decir, aquellas en las que la emoción observada no coincide con la emoción predicha por el clasificador, recibe el nombre de error de entrenamiento. Así mismo, la suma de las observaciones de la muestra de validación que no son correctamente clasificadas recibe el nombre de error de validación. La habilidad de un clasificador para clasificar correctamente observaciones que no pertenecen a la muestra de entrenamiento recibe el nombre de capacidad de generalización. Obviamente, buscaremos clasificadores tales que su capacidad de generalización sea máxima.

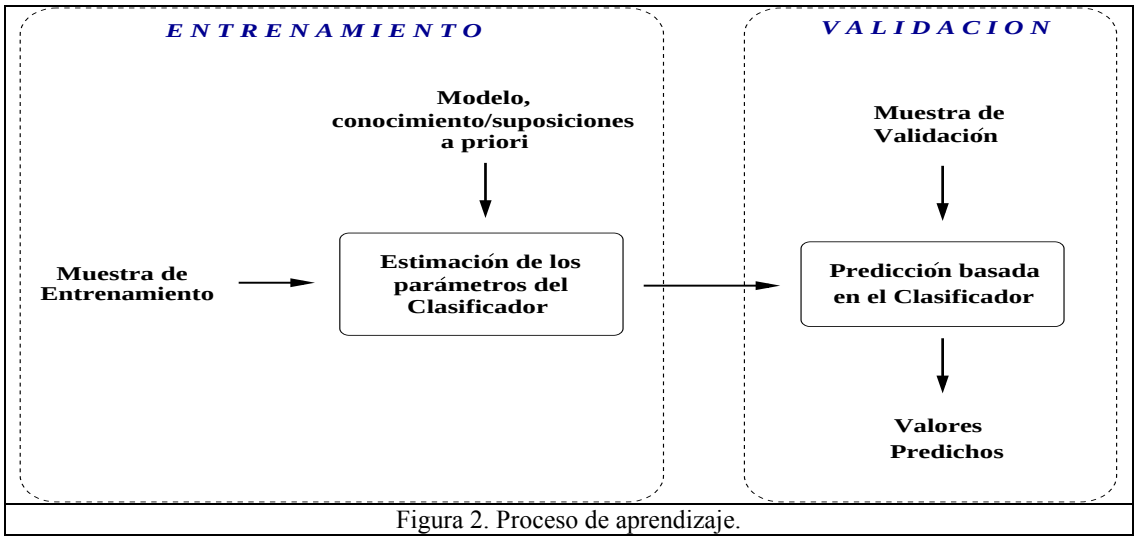

Figura 2. Proceso de aprendizaje.

A continuación pasamos a describir algunas de las técnicas de clasificación supervisada más usadas en el reconocimiento automático de emociones.

Redes Neuronales

El método de aprendizaje conocido como Redes Neuronales fue desarrollado simultáneamente en los ámbitos del análisis estadístico y de la inteligencia artificial. La idea central consiste en extraer combinaciones lineales de los atributos presentes en los objetos obteniendo una serie de características y modelizar las clases como funciones no lineales de dichas características. Karpouzis et al. (1999) y Yacoub et al. (2003) presentan algunos ejemplos de la aplicación de esta técnica de clasificación al problema de reconocimiento de emociones. En el primero de estos artículos, las redes neuronales son adaptadas para identificar y agrupar los músculos de la cara que contribuyen a detectar las emociones. En el segundo artículo las emociones tratan de ser reconocidas a partir de señales obtenidas del discurso. Las redes neuronales obtuvieron mejores resultados de clasificación que las técnicas alternativas con las que son comparadas: máquinas de vectores soporte, árboles de clasificación, k-vecinos más cercanos.

Máquinas de Vector Soporte

Las Máquinas de Vectores Soporte (o SVM de 'Support Vector Machine') son procedimientos de clasificación y regresión basados en la teoría estadística del aprendizaje (Vapnik, 1995). Podemos definir una SVM como una clase específica de algoritmos preparados para el entrenamiento eficaz de una máquina de aprendizaje lineal en un espacio inducido por una función núcleo (o kernel), de acuerdo a unas reglas de generalización empleando técnicas de optimización (ver Moguerza y Muñoz 2006 para una revisión completa). Las dos ideas fundamentales para la construcción de un clasificador SVM son la transformación del espacio de entrada en un espacio de alta dimensión y la localización en dicho espacio de un hiperplano separador óptimo. La transformación inicial se realiza mediante la elección de una función kernel adecuada. La ventaja de trabajar en un espacio de alta dimensión radica en que las clases consideradas serán linealmente separables con alta probabilidad, y por tanto, encontrar un hiperplano separador óptimo será poco costoso desde el punto de vista computacional. Además, dicho hiperplano vendrá determinado por unas pocas observaciones, denominadas, vectores soporte por ser las únicas de las que depende la forma del hiperplano. Una de las principales dificultades en la aplicación de este método radica en la elección adecuada de la función kernel. Es decir, construir la función de transformación del espacio original a un espacio de alta dimensión es un punto crucial para el buen funcionamiento del clasificador. La forma final de la regla de clasificación para un clasificador binario (dos clases, +1 y -1) queda como sigue:

$$f(x) = b + \sum_i \alpha_i K(x, x_i)$$

donde b y α_i son parámetros aprendidos por el clasificador durante el proceso de entrenamiento, $K(x, x_i)$ es el valor de la función kernel para los puntos x y x_i. Si *f(x)* es mayor que un umbral entonces la emoción estimada para un punto x será una +1 y será -

1 en caso contrario. En el problema del reconocimiento de emociones es típico trabajar con más de dos emociones. Supongamos que el número de emociones consideradas es n. Es necesario llevar a cabo una generalización del clasificador binario al caso multiclase. Existen varios algoritmos para esta generalización. En primer lugar, el algoritmo de "uno frente a uno", donde se entrena un clasificador binario por cada par de emociones disponibles. Dispondremos por tanto de 2*n clasificadores, es decir, 2*n valores de la regla de clasificación para cada objeto. El algoritmo "uno frente a todos" supone entrenar tantos clasificadores binarios como número emociones estén disponibles. En este caso disponemos n clasificadores, es decir, n valores de la regla de clasificación para cada objeto. En ambos casos para determinar la emoción que corresponde a cada objeto se realiza una ponderación sobre todas las reglas de clasificación disponibles.

Las SVMs han demostrado ser un método muy efectivo en la clasificación de expresiones faciales espontáneas (Bartlett et al., 2001)). Srivastavaa et al. (2006) han empleado con éxito las SVMs con kernel polinomial para el reconocimiento de personas empleando imágenes de rango de los rostros.

Árboles de Decisión

El propósito de este nuevo método de clasificación es crear una estructura de árbol que represente de forma eficiente las características más relevantes de los objetos considerados. El algoritmo de clasificación se representa como un árbol, donde las ramas son las condiciones establecidas sobre las características de los objetos y las hojas son las emociones consideradas. El proceso de clasificación comienza en la raíz y las ramificaciones del árbol se deciden a partir de las características del objeto más significativas. Dado un objeto, se desplaza a lo largo de las ramas del árbol hasta que finalmente se determina como emoción detectada aquella que define la última hoja. Existen varios algoritmos para construir un árbol de decisión (ID3, C4.5, CART, ID4,..., ver Breiman et al. 1984; Gatnar 1998 y Quinlan 1993). Sin embargo, la idea fundamental en todos ellos es la misma: los ejemplos en cada rama han de ser homogéneos, mientras que el tamaño del árbol ha de ser pequeño (es decir, la descripción ha de ser lo más simple posible). El algoritmo general de árbol de clasificación comienza considerando todas las posibles divisiones del conjunto inicial en subconjuntos. Se estima la calidad de cada una de las divisiones y se elige la mejor de todas ellas (la de menor entropía). El proceso se repite para cada una de las particiones realizadas, hasta que la entropía en cada uno de las hojas creadas es mejor que un valor predeterminado. Sun et al. (2004) han creado una base de datos de expresiones faciales donde los rostros de los sujetos presentan emociones espontáneas. Sobre esta base de datos han empleado con éxito varios algoritmos de clasificación, entre ellos los árboles de decisión.

Redes Bayesianas

Las redes bayesianas son clasificadores empleados para representar distribuciones conjuntas de modo que permitan calcular la probabilidad a posteriori de un conjunto de

clases (emociones) dado un conjunto de características observadas en los objetos, y así clasificar los objetos en la clase más probable. Una red bayesiana se compone de un grafo dirigido en el cual cada nodo está asociado con una característica y con una distribución de probabilidad condicional. El grafo representa la estructura y las distribuciones de probabilidad los parámetros de la red. La idea general consiste en usar una estrategia que pueda buscar de modo eficiente en el espacio de posibles estructuras y extraer aquella que dé mejores resultados de clasificación. Sun et al. (2004) y Gu y Ju (2004) han aplicado las redes bayesianas (entre otros clasificadores) para la detección automática de emociones.

Algoritmos de votación: Bagging, Boosting

El método llamado Bagging ('boostrap aggregating') fue propuesto por Leo Breiman para clasificación y árboles de regresión (Breiman, 1996). Supongamos que disponemos de un modelo, ajustado a nuestra muestra de objetos, tal que obtenemos un clasificador asociado a cada objeto. Repetimos la estimación del clasificador, modificando en cada caso la muestra de entrenamiento. Cada una de estas muestras recibe el nombre de muestra boostrap. El método Bagging consiste en obtener una media de las predicciones sobre un conjunto de muestras boostrap. El Adaboost es el algoritmo de tipo 'boosting' más popular (Freund y Schapire, 1997). El clasificador final se obtiene a partir de una ponderación de clasificadores 'débiles', esto es, con poca capacidad de generalización. Sebe et al. (2004) crearon una base de datos en la que las expresiones fáciles corresponden a estados emocionales de los individuos y aplicaron algoritmos de tipo Boosting y Bagging para el reconocimiento automático de dichas emociones. En ocasiones este tipo de técnicas se emplean como paso previo a otros clasificadores. Por ejemplo, Bartlett et al., (2003) usan, como paso previo al empleo de una SVM, Adaboost (un algoritmo de tipo Boosting) para la extracción de las características de mayor interés dadas las características previamente detectadas.

Modelos Markovianos Ocultos

Los modelos markovianos ocultos (HMM del inglés 'Hidden Markov Model') asumen que el sistema estudiado sigue un proceso de Markov con parámetros desconocidos. La tarea fundamental consiste en determinar los parámetros ocultos a partir de los parámetros observados. La diferencia fundamental respecto a un modelo de Markov habitual consiste en que los estados no son directamente visibles para el observador, pero sí lo son las variables influenciadas por el estado. Cada estado tiene una distribución de probabilidad asociada sobre el conjunto de posibles valores de salida. La secuencia de valores de salida generados a partir de un HMM nos dará cierta información sobre la secuencia de estados. Los tres problemas fundamentales a resolver en el diseño de un modelo HMM son: la evaluación de la probabilidad (o verosimilitud) de una secuencia de observaciones dado un modelo HMM específico; la determinación de la mejor secuencia de estados del modelo; y el ajuste de los parámetros del modelo que mejor se ajusten a los valores observados. DeSilva y Pei Chi (2000) describen el uso de estos modelos en

el reconocimiento automático de emociones. El objetivo de su método es clasificar las 6 emociones básicas a partir de expresiones faciales (video) y emociones en el discurso (audio). Aunque únicamente analizaron dos individuos la conclusión es que la información de video y de audio puede ser combinada usando un sistema basado en reglas para mejorar la tarea de reconocimiento, mejorando los resultados obtenidos por cada una de las representaciones individuales.

Métodos no supervisados

Todas las técnicas anteriormente descritas construyen un modelo, o clasificador, empleando la información relativa a la emoción representada por un conjunto de observaciones de entrenamiento. Posteriormente el clasificador construido se evalúa sobre nuevas observaciones sin tener en cuenta la emoción que éstas representan y se estima la calidad del mismo al comparar la emoción predicha por el clasificador con la emoción que en efecto representaba esa nueva muestra de validación o test. Sin embargo, es posible aplicar técnicas de clasificación en las cuales no se emplea el conocimiento de las emociones en la muestra de entrenamiento. Este tipo de técnicas reciben el nombre de métodos no supervisados. En la clasificación no supervisada diremos que hemos obtenido una buena clasificación cuando los grupos creados sean homogéneos respecto a los individuos que los forman y heterogéneos entre sí. Por ejemplo, Cao et al. (2003) ha empleado uno de estos métodos no supervisados, el análisis de componentes independientes (ICA de 'Independent Component Analysis'), para la detección de emociones en el discurso. Dada la naturaleza del problema de reconocimiento de emociones este tipo de técnicas serán muy útiles cuando las bases de datos disponibles se recojan en entornos no controlados y los individuos reflejen emociones no indicadas a priori por el investigador.

4.2 Combinación de información.

La integración o combinación de la información obtenida a través de audio y de video puede llevarse a cabo de dos modos diferentes. Por un lado es posible combinar los valores obtenidos por los diferentes clasificadores (fusión a nivel de la decisión, Kittler 1998). Por otro lado se pueden combinar las informaciones en el origen (fusión a nivel de características). Martín de Diego et al. (2005), han propuesto diversas técnicas para la combinación de la información como paso previo al entrenamiento de un clasificador. Las ideas fundamentales de las técnicas propuestas son dos: en primer lugar, la estimación efectiva de las diferencias de información entre los diferentes métodos de percepción de las emociones empleados. En segundo lugar, la extensión de la combinación lineal de información a la combinación funcional. La fusión de diversas fuentes de información se presenta como una de las líneas de investigación futuras más prometedoras. Algunos autores (por ejemplo, DeSilva y Pei Chi, 2000) han demostrado que la integración de información complementaria puede mejorar sustancialmente los resultados de clasificación obtenidos al emplear cada una de las fuentes de modo individual. En el ámbito de las SVMs, ciertos esfuerzos han sido realizados para determinar la necesidad

de combinar (Joachims, 2001). En el ámbito del reconocimiento de emociones consideramos que la tarea de integrar información complementaria procedente de diferentes medidas de interés sobre los individuos es una de las futuras tareas a considerar. En general no se ha realizado un esfuerzo exhaustivo por determinar qué aspectos son inherentes a la representación basada en imágenes y cuales lo son a la representación basada en el discurso. ¿Cuáles son los aspectos comunes a ambos modos de reflejar la realidad del individuo y cuáles son los aspectos específicos? Como se ha mencionado previamente, se han realizado algunas aproximaciones en la combinación de expresiones faciales y análisis de voz para el reconocimiento de emociones. Sin embargo, las emociones no sólo se manifiestan en las expresiones del rostro y en el tono de la voz, sino en otros aspectos tales como el movimiento de la cabeza, los brazos, las manos, en los gestos del cuerpo, el movimiento de los ojos… La tarea del reconocimiento automático de emociones debería de ser mejorada si tenemos en cuenta todos estos aspectos. Ben-Yacoub et al. (1999) emplean la combinación de clasificadores binarios conseguidos a partir de la información de la cara y la información del discurso. Conde (2006) emplea representaciones de la cara en dos y tres dimensiones. En este trabajo se realiza una fusión de los clasificadores obtenidos a partir de cada una de estas representaciones individuales.

5.- MATERIAL DISPONIBLE

En esta sección revisaremos parte del material disponible en el ámbito del reconocimiento de emociones. A pesar de que existen numerosas bases de datos de emociones, la mayoría de ellas tienen serias limitaciones. En el ámbito del discurso: la base de datos "Danish Emotional Speech Datasbase", consta de dos hombres y dos mujeres leyendo textos en cinco estados emocionales diferentes: enfado, felicidad, tristeza, sorpresa y un estado neutral (Engberg y Hansen, 2003). Otra base de datos disponible es la llamada GRONINGEN, que incluye alrededor de 20 horas de material en Holandés procedente de 238 personas. En el campo del reconocimiento facial existen numerosas bases de datos que muestran las caras de los individuos bajo diferentes condiciones de iluminación, escala, gestos, orientación, etc, sin embargo, muy pocas bases de datos consideran las emociones. Recientemente el grupo de investigación FRAV (*"Facial Recognition and Artificial Vision Group"*) de la Universidad Rey Juan Carlos, ha creado una base de datos en 3D, la FRAV3D, destinada a reconocimiento facial. La FRAV3D es una base de datos multimodal, ya que tiene información bidimensional o de textura, e información tridimensional. Fue adquirida mediante un escáner láser de Minolta [MIN], modelo VIVID-700, con el software *Polygon Editing Tool v 1.11 (PET)*. La información 3D es proporcionada en dos formatos: un mallado triangular en formato VRML, y una imagen de rango o mapa de profundidad. Las imágenes fueron adquiridas en condiciones controladas, con dos iluminaciones diferentes: la iluminación controlada y no controlada. La iluminación controlada consiste en dos focos halógenos de luz difusa, que mantienen las condiciones de iluminación invariables en el tiempo. La base de datos consta de imágenes de 105 individuos (81 mujeres y 24 hombres), todos ellos de raza caucásica. Se adquirieron 16 capturas por individuo, cada una de ellas proporcionando la corres-

pondiente información de textura y tridimensional: 4 capturas frontales, 8 giros con diferentes ángulos y ejes de giro, 2 gestos y 2 capturas cambiando la iluminación.

El material de video disponible de modo gratuito es aún mucho más limitado que el material de imágenes y bases de datos que incluyan simultáneamente video y discurso son mucho más difíciles de conseguir. Por todo ello, pensamos que el desarrollo de una base de datos apropiada para el estudio de las emociones es una tarea necesaria. El material de dicha base de datos debería de ser audiovisual. Debería de representar un amplio abanico de emociones y no las grandes emociones básicas o aquellas sugeridas por el investigador. Obviamente no debería estar limitado por características socioculturales sino que debería abarcar una amplia gama de razas y culturas. Además, debería de ser natural en su recogida, no siendo las emociones expresiones previstas a priori por el investigador y representadas por un conjunto de actores.

6.- CONCLUSIONES

En este artículo hemos presentado un resumen de las técnicas para el reconocimiento automático de emociones. Hemos definido los principales tipos de emociones primarias, así como la necesidad de usar otro tipo de emociones más específicas. Hemos considerado los dos canales fundamentales para el estudio de las emociones: las expresiones faciales obtenidas a partir de un video y las expresiones léxico-fonéticas obtenidas de un discurso. Las emociones recogidas en uno u otro medio pueden ser diferentes, sin embargo muchos de los trabajos más novedosos y prometedores se centran en la combinación de ambas fuentes de información. Presentamos un resumen de las técnicas de análisis estadístico más ampliamente utilizadas en el ámbito de la detección de emociones. El escaso número de amplias bases de datos disponibles para el estudio de las emociones plantea la necesidad de desarrollar una base de datos apropiada con fines predictivos.

7.- BIBLIOGRAFÍA

BARTLETT, M.S., BRAATHEN, B., LITTLEWORT-FORD, G., HERSHEY, J., FASEL, I., MARKS, T., SMITH, E., Y MOVELLAN, J.R. (2001) *Automatic Analysis of Spontaneous Facial Behavior: A Final Project Report.* Institute for Neural Computation MPLab TR2001.08, University of California, San Diego.

BEN-YACOUB S., ABDELJAOUED, Y., MAYORAZ, E. (1999). Fusion of Face and Speech Data for Person Identity Verification. *IEEE Transactions on Neural Networks*, 10 (5):1065-1074.

BREIMAN, L. FRIEDMAN, J.H., OLSHEN, R.A., STONE, C.J. (1984). *Classification and regression trees*. Belmont, C.A. Wadsworth.

CAMPBELL, N. (2000). Databases of emotional speech. *ISCA Workshop on Speech and Emotion: A conceptual framework for research*.

CAO, Y., FALOUTSOS, P., PIGHIN, F. (2003). Unsupervised Learning for Speech Motion Editing. *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation.*

CONDE, C. (2006). *Verificación facial multimodal: 2D y 3D*. Tesis Doctoral. Universidad Rey Juan Carlos.

CORNELIUS, R.R. (1996). *The Science of Emotion Research and Tradition in the Psychology of Emotion*. Upper Saddle River, NJ: Prentice Hall.

COWIE, R. (2000). Describing the emotional states expressed in speech. *ISCA Workshop on Speech and Emotion: A conceptual framework for research.*

COWIE, R., DOUGLAS-COWIE, E., SAVVIDOU, S., MCMAHON, E., SAWEY, M., SCHRÖDER, M. (2000). FEELTRACE: an instrument for recording perceived emotion in real time. *ISCA Workshop on Speech and Emotion:A conceptual framework for research.*

CROWLEY, J., Y BERARD, F. (1997). Multi-modal tracking of faces for video communications. *Proc. IEEE CVPR*, Puerto Rico, June 1997, pp. 640-645.

COLLOBERT, M., ET AL. (1996). Listen: A system for locating and tracking individual speakers. *Proc. Int. Conf. on Automatic Face and Gesture Recognition, Vermont,* 283-288.

DE SILVA, L.C. Y PEI CHI NG. (2000). Bimodal emotion recognition. *Automatic Face and Gesture Recognition, 2000.* Proceedings. Fourth IEEE International Conference: 332 – 335.

EKMAN, P., FRIESEN, W. (1978). *Facial Action Coding System: Investigator's Guide*. Consulting Psychologists Press.

EKMAN, P. EDITOR. (1982): *Emotion In the Human Face*. Cambridge University Press, New York, NY, 2nd edition,.

EKMAN, P. (1994). *Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique*. Psychological Bulletin 115, 268–287

ENGBERG, I. S. Y HANSEN, A. V. 2003. Danish Emotional Speech Database. Center for PersonKommunikation, Department of Communication Technology, Institute of Electronic Systems, Aalborg University, Denmark.

ESSA, I. Y PENTLAND, A. (1995). *Coding, analysis, interpretation and recognition of facial expressions*. MIT Media Lab., Cambridge, MA, Tech. Rep. 325.
ESSA, I., DARRELL, T. Y PENTLAND, A. (1995). Tracking facial motion. *Proc.Workshop on Motion of Nonrigid and Articulated Objects, 1994,* pp. 36-42.

ESSA, I. Y PENTLAND, A. (1993). *Physically-based modeling for graphics and vision*. Directions in Geometric Computing. Information Geometers.R. Martin, Ed. U.K.

FREUND, Y., Y SCHAPIRE, R. (1997). A decision-theoretic generalization of online leraning and an application to boosting. *Journal of Computer and system Sciences*, 55:119-139.

GRAF, H., COSATTO, E., GIBBON, D., KOCHEISEN, M., Y PETAJAN, E. (1996). Multi-modal system for locating heads and faces. *Proc. Int. Conf. on Automatic Face and Gesture Recognition, Vermont,* pp. 88-93.

GU, H., Y JI, Q. (2004). An automated face reader for fatigue detection. *Automatic Face and Gesture Recognition, 2004.* Proceedings. Sixth IEEE International Conference: 111-116.

IZZO, G. (1998). *PHYSTA Project Report, Multiresolution techniques and emotional speech*. NTUA Image Processing Laboratory, Athens.

JOACHIMS, T., CRISTIANINI, N. Y SHAWE-TAYLOR, J. (2001). Composite Kernels for Hypertext Categorisation. *In Proceedings of the International Conference on Machine Learning, ICML'01*. Morgan Kauffman.

KARPOUZIS, K, VOTSIS, G., MOSCHOVITIS, G., KOLLIAS, S. (1999). Emotion Recognition Using Feature Extraction and 3-D Models. *CSCC'99 Proceedings*: 5371-5376.

KITTLER, J., HATEF, M., DUIN, R.P.W. Y MATAS, J. (1998). On Combining Classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3): 226-239.

LIEN, J., KANADE, T., COHN, J. Y LI, C. (1998). Subtly different facial expression recognition and emotion expression intensity estimation. *Proc. IEEE CVPR*, Santa Barbara, CA, pp. 853-859.

LUCAS, B., Y KANADE, T.. (1981). An iterative image registration technique with an application to stereo vision. *Proc. 7th Intl. Joint Conf. on AI*, 1981.

MARTIN DE DIEGO, I. (2005). *Máquinas de Vectores Soporte: un enfoque basado en la Combinación de Información*. Tesis Doctoral. Universidad Carlos III de Madrid.

MOGUERZA, J.M. Y MUÑOZ, A. (2006). Support Vector Machines with applications. Statistical Science. Aceptado para su publicación.
MURRAY, I. R. , Y ARNOTT, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of literature on human vocal emotion. *Journal of Acoustical Society of America*, 93(2).

OTSUKA, T., Y OHYA, J. (1997). Recognizing multiple persons' facial expressions using HMM based on automatic extraction of frames from image sequences. *Proc. IEEE Int. Conf. on Image Proc.,* (2): 546-549.

PEREIRA, C. (2000). Dimensions of emotional meaning in speech. *ISCA Workshop on Speech and Emotion: A conceptual framework for research*.

QUINLAN, J.R. (1993). *C4.5 programs for machine learning*. Morgan Kaufmann, San Mateo.

SEBE, N. SUN, Y., BAKKER, E., LEW, M.S., COHEN, I., Y HUANG, T.S. (2004). Towards Authentic Emotion Recognition. *International Conference on Systems, Man, and Cybernetics (IEEE SMC'04).*

SRIVASTAVAA, A., LIUB, X., Y HESHERB, C. (2006). Face recognition using optimal linear components of range images. *Image and Vision Computing*, 26:291-299.

SUN, Y., SEBE, N., LEW, M.S. Y GEVERS, T. (2004). Authentic Emotion Detection in Real-time Video. *International Workshop on Human Computer Interaction (HCI'04),* 92-101.

TSAPATSOULIS, N., AVRITHIS, I. Y KOLLIAS, S. (1999). On the use of radon transform for facial expression recognition. *Proc. 5th Intl. Conf. Information Systems Analysis and Synthesis*, Orlando, FL, Jul.

YACOUB, S., SIMSKE, S., LIN, X., BURNS, J. (2003). Recognition of Emotions in Interactive Voice Response Systems. HP Laboratories Palo Alto; HPL-2003-136.

VAPNIK, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Para citar este artículo puede utilizar la siguiente referencia:

MARTÍN DE DIEGO, I.; SERRANO, A.; CONDE, C. Y CABELLO, E. (2006): Técnicas de reconocimiento automático de emociones. En GARCÍA CARRASCO, Joaquín (Coord.) Estudio de los comportamientos emocionales en la red [monográfico en línea]. Revista electrónica Teoría de la Educación: Educación y Cultura en la sociedad de la información. Vol. 7, nº 2. Universidad de Salamanca. [Fecha de consulta: dd/mm/aaaa].
<http://www.usal.es/~teoriaeducacion/rev_numero_07_02/n7_02_martin_serrano_conde_cabello.pdf >
ISSN 1138-9737