IV Jornadas de Reconocimiento Biométrico de Personas

E.T.S. de Ingeniería Informática
Universidad de Valladolid

11 y 12 de Septiembre de 2008

IV Jornadas de Reconocimiento Biométrico de Personas

Editado por:

César González Ferreras
Valentín Cardeñoso Payo
Carlos Vivaracho Pascual

Septiembre 2008.
ISBN: 978-84-691-5008-5
Imprenta: Mata Digital, S.L.
Depósito Legal: VA-864-2008
Impreso en España - Printed in Spain

Comité Organizador

Presidente:	Carlos Vivaracho Pascual	(Universidad de Valladolid)
Secretario:	Juan Manuel Pascual Gaspar	(Universidad de Valladolid)
Miembros:	Valentín Cardeñoso Payo	(Universidad de Valladolid)
	David Escudero Mancebo	(Universidad de Valladolid)
	César González Ferreras	(Universidad de Valladolid)
	Marcos Faundez Zanuy	(Esc. Politécnica de Mataró, UPC Barcelona)
	Julián Fiérrez Aguilar	(Esc. Politécnica Superior, Universidad Autónoma de Madrid)

Índice

Presentación

Ver el nacimiento y crecimiento de algo siempre es emocionante, pero todavía lo es más si uno forma parte de ello. Por eso, cuando en Sevilla desde la ABIE, impulsora de la creación de estas Jornadas, nos propusieron organizar las IV Jornadas de Reconocimiento Biométrico de Personas, decidimos aceptar el reto. No ha sido fácil, pero con la ayuda de todos aquí está el resultado.

El número de contribuciones recibidas ha sido de 19, lo que supone un pequeño incremento con respecto a las anteriores Jornadas. En el proceso de revisión la valoración general de la calidad de los trabajos presentados ha sido bastante buena.

La respuesta de los investigadores ha sido positiva, y creemos que los trabajos que aquí se publican representan a los principales grupos universitarios involucrados en el tema de la biometría en España. Podemos considerar que, en este sentido, estas Jornadas se están consolidando como punto de encuentro del mundo de la investigación.

Sin embargo, y a pesar de los esfuerzos realizados, la respuesta del mundo de la empresa no ha sido la esperada. Consideramos que éste es uno de los retos pendientes para siguientes Jornadas, ya que siempre es importante, y más en un tema tan actual y en expansión como la Biometría, el contacto entre la Investigación y el Desarrollo. Somos conscientes de que no es fácil, pero desde aquí ofrecemos nuestra experiencia a aquellos que acepten el reto de organizar las siguientes.

Esperamos que los asistentes a las Jornadas, además de una participación fructífera en ellas, se hayan llevado un buen recuerdo de la hospitalidad vallisoletana, así como de sus Fiestas y Ferias.

Septiembre 2008

El Comité Organizador
JRBP'08

Una Revisión del Estado del Arte sobre Verificación *Off-Line* de Firmas Manuscritas

José Vélez, Ángel Sánchez

Departamento de Ciencias de la Computación
Universidad Rey Juan Carlos, c/ Tulipán s/n
28933 Móstoles, Spain
{jose.velez, angel.sanchez}@urjc.es

Resumen. Este trabajo describe los avances realizados en la investigación sobre verificación automática *off-line* de firmas manuscritas, tanto a nivel de características discriminantes extraídas de la imagen de una firma como de las técnicas de clasificación utilizadas.

1 Introducción

La verificación de firmas manuscritas es una modalidad biométrica conductual que se ha abordado clásicamente dentro del área forense. Las primeras técnicas utilizadas se remontan a finales del siglo XIX, cuando este tipo de métodos renovó la investigación policial. Desde este punto de vista, las técnicas no han variado sustancialmente desde aquella época, y ni los modernos sistemas de microscopía, ni los progresos en computación han modificado en un grado perceptible sus procedimientos. El desarrollo de algoritmos para el tratamiento automático de firmas es un área relativamente reciente. Se inicia a comienzos de la década de 1970 con los primeros Sistemas de Visión Artificial y desde entonces es un área de investigación muy activa.

En este trabajo se revisan las técnicas de verificación automática *off-line* de firmas manuscritas. Como se puede comprobar, la actividad en esta área (motivada en gran medida por el interés que para la industria tiene) queda patente por la cantidad de trabajos que se publican cada año. Dado el carácter privado de las firmas, no existen bases de datos públicas para la comparación entre resultados de trabajos. Para solventar esta dificultad, y permitir comparar los métodos de verificación propuestos usando los mismos *datasets* de firmas, en el año 2004 se celebró la primera competición internacional de verificación de firmas SVC (*First International Signature Verification Competition*), la cual ha continuado realizándose periódicamente hasta la actualidad. Por otro lado, desde el trabajo de *survey* escrito por Leclerc y Plamondon en 1994 [1] (que a su vez se plantea como continuación de otro realizado por Plamondon y Lorette [2] en 1989), no aparecen en la literatura científica nuevos trabajos recopilatorios extensos y completos sobre las técnicas y clasificadores más actualmente usados para la verificación automática de firmas *off-line*. Nuestro artículo pretende contribuir a cubrir esta laguna.

2 Descripción de la problemática de la verificación de firmas

El problema de la verificación consiste en determinar el grado de similitud entre una firma de test presentada al sistema automático y otra forma firma modelo almacenada, con el fin de establecer si la firma a comprobar es auténtica o una falsificación. Como se enuncia en [3]: "en el problema de verificación de firmas se trata de maximizar las diferencias interpersonales y minimizar las diferencias intrapersonales".

Los falsificadores, que pueden tratar de engañar a un sistema de este tipo, pueden catalogarse en dos grupos: los falsificadores entrenados y los no entrenados [1]. Los primeros, que conocen la firma de la persona que quieren suplantar, se entrenan en reproducir la firma y consiguen unas falsificaciones de gran calidad. Los segundos no sólo no están entrenados, sino que no han visto nunca la firma, por lo que su reproducción no guarda ningún parecido con la auténtica. Curiosamente, y en contra de lo que pudiese pensarse, a este segundo grupo de falsificadores corresponde el 95% del fraude existente en entidades bancarias en el mundo.

El proceso de captura de firmas puede realizarse de dos formas diferentes: *on-line* y *off-line* [1]. En el modo *on-line* la captura se realiza utilizando un dispositivo especial (una lápiz electrónico o una tableta gráfica especial) que recoge información dinámica del escritor durante la firma. Esta información incluye, además del grafismo, datos como presión, velocidad, puntos de inicio, direcciones de los trazos, inclinación, etc. Existen actualmente sistemas que realizan de forma eficiente reconocimiento y verificación utilizando esta información. Por otro lado, el método de captura *off-line* se basa en el escaneo de una firma, una vez realizada sobre un soporte ordinario (papel). En este caso, la información es mucho menor y la resolución espacial y radiométrica a la que se escanea influye en la verificación.

Además, en la formulación *off-line*, puede aparecer el problema de localizar y segmentar la firma en el documento. A su vez, la segmentación dentro de un documento presenta diferentes problemas: desconocimiento de la posición exacta, y existencia de ruido blanco y con estructura, etc.

3 Técnicas de verificación automática de firmas *off-line*

En esta sección se repasan los trabajos más relevantes que han aparecido en el campo de la segmentación y la verificación automática de firmas. Para ello se han realizado numerosas búsquedas en los principales foros especializados, entre los que podemos citar: el *International Workshop on Frontiers of Handwriting Recognition* (IWFHR), la *International Conference on Pattern Recognition* (ICPR), la *International Conference on Document Analysis and Recognition* (ICDAR), así como en las revistas: *International Journal of Pattern Recognition and Artificial Intelligence* (IJPRAI), *Pattern Recognition* y *Pattern Recognition Letters*.

El desarrollo de sistemas informáticos para la verificación *off-line* de firmas se inicia con los trabajos de Nagel y Rosenfeld en 1973 [4][5]. La gran mayoría de los trabajos se refieren a firmas occidentales, aunque también hay trabajos sobre firmas en el mundo árabe [6] y oriental [7][8][9]. Desde entonces, periódicamente, han ido

apareciendo algunos trabajos recopilatorios sobre el estado del arte del tratamiento automático de firmas manuscritas, como: [2] [1] [10] [11] [12] [13] [14] [15][16].

3.1 Tipología de los trabajos sobre verificación *off-line* de firmas

En general, los trabajos se enfocan desde una perspectiva grafonométrica [12], es decir se intenta realizar la verificación realizando diversas medidas sobre la imagen de la firma. Una diferencia con los trabajos anteriores a la aparición del ordenador estriba en el aumento de complejidad de las medidas que el ordenador permite realizar. Los trabajos podemos englobarlos en dos categorías principales:

- Aquéllos que aportan alguna característica discriminante novedosa que puede utilizarse para el problema de la verificación. En estos trabajos se deja en segundo plano el tipo de clasificador utilizado. En los primeros trabajos las características discriminantes se asimilaron del enfoque clásico de la grafoscopía, pero con el tiempo han aparecido características novedosas generalmente dependientes de las posibilidades introducidas por el tratamiento informático de las imágenes y el incremento en capacidad de cómputo de los ordenadores.
- Aquéllos en los que el clasificador usado es la novedad. En estos casos suele justificarse que el clasificador mejora los resultados obtenidos por características discriminantes propuestas por otros autores, o bien que el clasificador en sí mismo posibilita un análisis directo de la firma que no precisa de otras características.

Por supuesto, también encontramos trabajos mixtos en los que se combinan nuevas características con el uso de clasificadores novedosos.

Características discriminantes usadas para la verificación de firmas *off-line*

La extracción de características sobre la imagen de la firma sigue un enfoque clásico de Visión Artificial [17]. En este caso, tras las etapas de captura, preproceso y segmentación se realiza una extracción de de características discriminantes que se utilizan para clasificar el objeto (en este caso para verificar la firma). Diferentes autores suelen clasificar las características discriminantes utilizadas en el proceso de verificación de diferentes maneras. Inicialmente, en un trabajo de 1936, Locard [18] las clasificó en estáticas y pseudo-dinámicas, según utilicen información estática de la imagen o traten de hallar la dinámica del proceso de firmado subyacente. Esta clasificación es retomada por Hou y otros [16] en un trabajo del 2004. En 1991, S. Lee y J.C. Pan [19] citan tres tipos de características: las globales, que se basan en el estudio de cada píxel de la imagen por separado, las estadísticas, basadas en el estudio de las distribuciones de los píxeles de la firma, y las geométrico-topológicas, que describen las formas interiores a una firma. En 2004, Fiérrez y otros [15], clasifican las características discriminantes en dos grupos: características globales (global u *holistic*) que utilizan la imagen de la firma en conjunto y características locales (local o *grid*), que se basan en el estudio de zonas o partes especificas de la firma.

A continuación, se recogen las principales características globales que se repiten en la literatura para el caso de la verificación *off-line*. En general, se hace referencia a los

principales trabajos en las que aparecen, aunque en algunos casos se hace referencia también a trabajos recopilatorios. Las características consideradas son:

- Proporciones de la firma (*aspect ratio*) mediante la medida de la caja que contiene la firma (*bounding box*) [20][21][6].
- Centros de gravedad y otros momentos medidos sobre las proyecciones de la firma sobre los ejes horizontales y verticales [22][21][6].
- Línea base global (*global baseline*) y límites superior e inferior de la firma [23][6].
- Número de bucles, puntos de cruce y puntos extremos en la firma [24][25].
- Medidas sobre diferentes tipos de envolventes de la firma (*envelope*) [22][23].
- Estimaciones del ángulo de la línea base de la firma (*slope)* mediante los ejes de inercia [15][26].
- Estimación del ángulo de inclinación de sus trazos (*slant*) [20][23][6].
- Área de los píxeles activos [23][27][28], a menudo normalizados respecto al área de la envolvente de la firma y utilizando el esqueleto para ser invariantes al grosor del elemento de escritura usado para firmar.
- Número de componentes de la firma [20][27], mediante el análisis de las componentes conexas de la imagen.
- Características basadas en *wavelets* sobre la imagen de la firma [28][23].
- Una última característica global que se utiliza frecuentemente es la propia imagen de la firma, o una imagen escalada de ella a una resolución menor [25] [29] [30].

Por otro lado, entre las características locales que encontramos en la literatura abundan variantes de las globales, aplicadas sobre regiones acotadas de la firma, como ventanas (verticales u horizontales) o celdas [23]. También se encuentran características locales que no derivan de otras globales, entre ellas podemos destacar:

- Densidades por celdas o regiones de la imagen [23][31].
- Medidas locales de orientación de trazos [32][33].
- Ángulo predominante de situación de píxeles por ventanas [32].
- Análisis local de una retícula situada sobre la firma mediante unos artefactos matemáticos llamados Distribuciones de Tamaño Granulométrico (*Granulometric Size Distributions*) [34] que, basados en morfología matemática, permiten obtener una descripción vectorial de cada punto de la retícula.
- Características pseudo-dinámicas como la presión obtenida por equivalencia con la intensidad de la luminosidad del trazo en imágenes en niveles de gris [35], o una reconstrucción de la dinámica del trazado [33].

Clasificadores utilizados en el problema de la verificación de firmas *off-line*

Es un hecho que, al poco tiempo de la aparición de un nuevo tipo de clasificador, algún grupo de investigación lo prueba para el problema de verificación automática de firmas. Así se encuentran trabajos que utilizan las Redes Neuronales, las Máquinas de Vectores de Soporte, los Métodos de Ajuste Elástico, las Modelos Ocultos de Markov o los Clasificadores Difusos, entre otros. En la Tabla 1 se muestran algunos trabajos que utilizan estos clasificadores.

Tabla 1. Algunos tipos de clasificadores utilizados para la verificación *off-line* de firmas y trabajos representativos.

Tipo de clasificador	Trabajos referenciados
Redes Neuronales (NN)	[36][22][25]
Máquinas de Vectores Soporte (SVM)	[8][37]
Ajuste Elástico (*Elastic Matching*)	[38][39][40]
Modelos Ocultos de Markov (HMM)	[41][42][37]
Clasificadores Difusos	[43][6][44]
Redes Bayesianas	[45]

3.2 Descripción de algunos trabajos relevantes en el problema de verificación automática de firmas *off-line*

Aunque han sido muchos los artículos revisados, en los siguientes puntos se describen brevemente algunos de los más significativos tanto por el método que describen como por los problemas que abordan:

- M. Ammar y otros [20] proponen en un trabajo de 1990 un análisis de la firma basado en características globales y en un análisis local que da lugar a una representación en árbol de varios elementos constituyentes (*Global Descriptor String* o GDS). Desafortunadamente, en su trabajo no realizan pruebas experimentales de verificación.
- S. Lee y J.C. Pan [19] proponen una representación de la firma basada en una serie de elementos que simulan el proceso humano de generación de trazos (*strokes*). En el trabajo se exponen 7 reglas heurísticas que se siguen a la hora de construir el trazado de la firma. En este trabajo tampoco se ofrecen resultados experimentales.
- Qi y Hunt [23] comparan un conjunto de características geométricas con el estudio de las características obtenidas al superponer una rejilla (*grid*) a la firma y realizar el análisis los bordes de cada celda de la rejilla para obtener un código binario descriptor. Debido a que el uso directo de la distancia euclídea entre los patrones así obtenidos no da resultados satisfactorios, se utiliza un proceso de ajuste previo entre los patrones de características basado en técnicas de programación dinámica. Los porcentajes de error que postula son muy bajos, pero en los experimentos no se aprecia una clara separación entre las etapas de aprendizaje y de test.
- R. Bajaj y S. Chaudhury [22] construyen un sistema basado en dos tipos de características discriminantes: momentos y envolventes superior e inferior. Utilizan redes de neuronas de tipo *feed-forward* para clasificar. Para una base de datos de 10 individuos (15 firmas por sujeto), usando 5 muestras para el aprendizaje de cada individuo, tienen un FRR del 1% y un FAR del 3% para falsificaciones aleatorias.
- B. Fang y su equipo [46][39] abordan el problema de los falsificadores habilidosos utilizando una aproximación basada en el conocimiento de expertos humanos. Como las falsificaciones son menos suaves y naturales que las firmas genuinas, construyen un índice de suavidad para su estudio. En este trabajo se justifica no poder ofrecer una distinción entre FRR y FAR debido al reducido tamaño de la muestra y sólo se ofrece un 17.4% de error medio en la verificación.

- También en 1999, V. E. Ramesh y otros [21] construyen un sistema que funciona con firmas escaneadas a 72 DPIs. El enfoque se basa en características discriminantes globales, en características de tipo *grid* y en características obtenidas mediante un análisis de *wavelets*. En este trabajo, los autores ensayan diferentes tipos de clasificadores. Utilizando el mejor de ellos, 15 firmas de cada individuo para entrenar y falsificaciones de las firmas para entrenar los rechazos, obtienen un 10% de FRR, un 2% de FAR para las falsificaciones simples y un 30% para las falsificaciones habilidosas.
- Y. Mizukami y otros [47] abordan el problema de usar pocos ejemplares para el entrenamiento (usan uno para aprendizaje y otro para determinar un umbral de rechazo). Su enfoque se basa en la comparación de imágenes utilizando funciones de desplazamiento. Sus resultados (24% de error medio) son muy buenos teniendo en cuenta que trata la problemática de los falsificadores habilidosos. Sus pruebas se realizan sobre firmas japonesas, que al ser tan diferentes a las firmas occidentales hace que sus resultados sean difíciles de comparar con otros.
- K. Huang y H. Yan [27] estudian la firma descomponiéndola en los trazos que aparentemente la componen (mediante fronteras direccionales) y ordenando los trazos en una posible secuencia temporal. Posteriormente, para verificar una firma, estudian la correspondencia entre los modelos de firma que han obtenido mediante un procedimiento que denominan "ajuste por relajación" (*relaxation matching*).
- X. Xiao y G. Leedham [45] proponen el uso de una red bayesiana para tratar el problema de la desaparición de las características discriminantes que ocurre al comparar las envolventes de la firma como método de verificación. Para sus experimentos crean una muestra de 8 individuos con entre 10 y 20 firmas por cada uno. Utilizando el 60% de la muestra para el aprendizaje se obtienen unos resultados para el FRR del 20% y para el FAR (falsificaciones aleatorias) del 14%.
- En 2002 B. Fang y otros [39] abordan la problemática de la dificultad de obtener múltiples firmas por cada individuo. En este caso, intentan utilizar técnicas de ajuste elástico (*elastic matching*) para generar nueva muestra modificando la existente. Además, utilizan ventanas verticales y horizontales para obtener ciertas medidas en las transiciones de píxeles blancos a negros que usan como características discriminantes. Este trabajo encuentra precedente en uno de Oliveira y otros [48], aunque en aquél se centran en la generación de la muestra y no se realiza ningún experimento de verificación. Utilizando 23 firmas de cada individuo para generar 529 muestras de aprendizaje por individuo se consiguen unos resultados del 14% de error medio (de nuevo, justificando que no se puede obtener una distinción entre FRR y FAR).
- En 2005 X. G. You y otros [38] utilizan la distancia entre los puntos localizados en dos modelos elásticos realizados sobre las firmas a comparar, atacando el problema de la imposibilidad de utilizar varias firmas y el de los falsificadores habilidosos a la vez. En este trabajo utilizan sólo 4 muestras por individuo durante la fase de aprendizaje, una para el modelo y 3 para determinar los umbrales de rechazo. Sus resultados fueron de un 18.6% de EER.
- G. Rigoll y A. Kosmala [31], y después E. Justino, Bortolozzi y R. Sabourin [41][50][32], han estudiado el problema utilizando: una segmentación de la firma en celdas, la obtención de una cadena de símbolos y el posterior uso de HMM como clasificador. Las características que se utilizan para cada celda en estos

trabajos son: la densidad de los píxeles negros [31], el ESC (*Extended Shadow Code*) de R. Sabourin [51] y el ángulo predominante [32]. En 2005, Justino y otros [37] comparan, en un trabajo que destaca por el impecable proceso de recolección de muestra y la experimentación, el uso de SVM y HMM para diferentes características extraídas sobre un *grid*. Es importante señalar que éste es uno de los poco trabajos en los que se señala que utilizan una muestra para diseñar el sistema y otra diferente para realizar el test. Además, se realizan diferentes pruebas con un número creciente de ejemplares para el aprendizaje.

- En un trabajo de 2007, A. Piyush y A. N. Rajagopalan [40] verifican firmas mediante la comparación de las proyecciones verticales utilizando una variación de la técnica *Dynamic Time Warpi*ng (DTW) que se basa en programación dinámica, por lo que también se le conoce como *Dynamic Programing Matching*. Esta técnica se usa ampliamente para alinear secuencias de manera óptima, en el sentido de minimizar la distancia entre dos secuencias de características discriminantes de diferente longitud. Los resultados que reportan son de un FRR del 2% y un FAR del 0% para falsificadores aleatorios cuando sólo trata de distinguir entre ambos. Cuando se consideran las falsificaciones habilidosas reportan un FRR del 25%, y un FAR del 0% para los aleatorios y de un 20% para los habilidosos.
- En otro trabajo de 2007, Yu Quiao y otros [52] proponen un enfoque basado en la combinación de modelos *on-line* para el aprendizaje y de verificación *off-line*. En su trabajo se utiliza la técnica de *Conditional Random Fields* (CRF) para encontrar la correspondencia entre la firma *on-line* y la imagen *on-line*. Tras esto se obtiene una trayectoria que, una vez alineada utilizando DTW, se compara con las trayectorias *on-line* almacenadas para ese individuo. En sus experimentos utilizan la base de datos de la competición de verificación de firmas del año 2004 (SVC 2004), la cual consta de firmas inglesas y chinas. Emplean 10 firmas *on-line* de cada individuo para el aprendizaje, otras 10 *on-line* se convierten a *off-line* y se utilizan para el test y también se hace lo mismo con 20 falsificaciones simuladas. Sus resultados arrojan un 7.3% de ERR, aunque no detallan qué porcentaje corresponde a firmas chinas y cuál a firmas inglesas.

Se puede resumir que el FAR obtenido para las falsificaciones aleatorias ronda el 0% y el FRR varía entre el 0% y el 25%. Sin embargo, debe tenerse en cuenta que no se ha encontrado ningún trabajo en el cual sólo se utilice una única muestra durante la fase de aprendizaje. Además, cuando el número de ejemplares utilizados en el aprendizaje se reduce, el error crece de forma importante. En cuanto al FAR frente a falsificadores habilidosas, debe señalarse que normalmente es muy alto (superior al 30%). Cuando este error es bajo se debe a que se entrena con las falsificaciones, cosa que en un sistema real no sería normalmente posible.

Por otro lado, respecto a la metodología observada en estos trabajos, se debe señalar que pocos separan la muestra que se utiliza para diseñar el sistema de la muestra que se utiliza para probarlo. Esto puede restar validez a los resultados ya que los algoritmos que se diseñan podrían estar adaptados a las particularidades de la muestra de test que se utiliza.

Con respecto a los conjuntos de entrenamiento y de test utilizados se debe señalar la completa falta de homogeneidad. Cada autor utiliza una muestra propia que luego no hace pública. Además, la muestra que utilizan casi siempre se concentra en tomas

de una sola sesión para cada individuo. En ninguno de los casos se realizan capturas de muestra a un mismo individuo a lo largo de diferentes años.

En muchos casos, para la verificación de firmas se requiere una etapa de segmentación de las mismas en los documentos que las contienen (p. ej. en cheques bancarios). Se mencionan algunos trabajos que abordan esta problemática: Hobby [53], Lamarche y Plamondon [54], Larrea y otros [55], Madasu y Novell [56], entre otros autores. Por último, se presenta en la Tabla 2 una comparativa de trabajos en cuanto a errores de clasificación y número de ejemplares del aprendizaje.

Tabla 2. Comparativa de trabajos sobre verificación *off-line* que consideran falsificaciones aleatorias.

Autores	**FRR**	**FAR**	**Firmas entren.**	**Descripción método**
Bajaj et al (1997)	1%	3%	5	Características discriminantes
Ramesh et al (1999)	10%	2%	15	Caract. Discrim. globales + *grid* + *wavelets* + 72 DPIs
Xiao et al (2002)	20%	14%	6-12	Redes Bayesianas
Vélez et al (2003)	24%	24%	1	Redes de Compresión
Justino et al (2005a)	25%	27%	5	HMM
Justino et al (2005b)	30%	0%	5	SVM
Piyush et al (2007)	25%	0%	10	*Dynamic Time Warping*
Vélez et al (2007)	12%	12%	1	*Snakes*

4 Conclusión

En este trabajo se ha realizado una descripción del problema de la verificación de firmas *off-line*. Asimismo, se ha presentado un estado del arte actualizado respecto al problema de la verificación de firmas *off-line* en su conjunto.

Referencias

1. Leclerc, F., Plamondon, R.: Automatic signature verification the state of the art 1989-1993, IJPRAI 8 (1994) 643-660
2. Plamondon, R., Lorette, G.: Automatic signature verification and writer detection: the state of the art", Pattern Recognition 22 (1989) 107-131
3. Justino, E., et al.: The Interpersonal and Intrapersonal Variability Influences of Off-Line Signature Verification using HMM, Proc. XV SIBGRAPI (2002)
4. Nagel, R.N.: Computer screening of handwritten signatures: a proposal, Computer Science Centre. University of Maryland, TR-220 (1973)
5. R.N. Nagel, A. Rosenfeld, "Steps towards handwritten signature verification", Proc. 1st Intl. Joint Conf. on Pattern Recognition (1973) 59-65
6. Ismail, M.A., Gad, S.: Off-line Arabic signature recognition and verification, Pattern Recognition 33 (2000) 1727-1740

7. Ammar, M.; Identification of fraudulent Japanese signatures from actual handwritten documents: a case study, Proc. IWFHR (1991) 369-374
8. Lv, H., Wang, W.Y. Wang, C., Zhuo, Q.: Off-line Chinese signature verification based on Support Vector Machines, Pattern Recognition Letters 26 (2005) 2390-2399
9. Yoshimura, I., Yoshimura, M.: Off-line verification of Japanese signature after elimination of background patterns, IJPRAI 8 (1994) 693-708
10. Pirlo, G.: Algorithms for Signature Verification, In: Fundamentals in Handwriting Recognition, Springer (1994) 435-454
11. Impedovo, S., Dimauro, G., Pirlo, G.: Algorithms for Automatic Signature Verification, Handbook of Character Recognition and Document Image Analysis, (1997) 605-621
12. Sabourin, R.: Off-line signature verification: Recent advances and perspectives, Proc. First Brazilian Symp. on Document Image Analysis (1997) 84-98
13. Plamondon, R., Srihari, S.N.: On-line and off-line handwriting recognition: A comprehensive survey, IEEE Trans. PAMI 22 (2000) 63-84
14. Impedovo, S., et al.: Recent Advances in Automatic Signature Verification, Proc. IWFHR (2004) 179-184
15. Fiérrez, J., Ortega, J., González, J.: Reconocimiento de firma escrita (in Spanish), In: Tecnologías biométricas aplicadas a la seguridad, Ra-Ma, (2004) 201-222
16. Hou, W., Ye, X., Wang, K.: A Survey of Off-Line Signature Verification, Proc. Intl. Conf. on Intelligent Mechatronics and Automation (2004) 536-541
17. Gonzalez, R.C., Woods, R.E.: Digital Image Processing (3rd Ed.), Addison-Wesley (2007)
18. Locard, E., Traité de criminalistique, Payot (1932)
19. Lee, S., Pan, J.C.: Offline tracing and representation of signatures, Proc. IEEE Conf. on Computer Vision and Pattern Recognition (1991) 679-680
20. Ammar, M., Yoshida, Y., Fukumura, T.: Structural description and classification of signature images, Pattern Recognition 23 (1990) 697-710
21. Ramesh, V. E., Narasimha Murty, M.: Off-line signature verification using genetically optimized weighted features, Pattern Recognition 32 (1999) 217-233
22. Bajaj, R., Chaudhury, S.: Signature verification using multiple neural classifiers, Pattern Recognition 30 (1997) 1-7
23. Qi, Y., Hunt, B.: Signature verification using global and grid features, Pattern Recognition 27 (1994) 1621-1629
24. Papamarkos, N., Baltzakis, H.: Off-line signature verification using multiple neural network classification structures", Proc. IEEE Intl. Conf. Digital Signal Processing (1997) 727-730
25. Baltzakis, H., Papamarkos, N.: A new signature verification technique based on a two-stage neural network classifier, EAAI 14, (2001) 95-103
26. Kalera, M.K., Srihari, S., Xu, A.: Off-line signature verification and identification using distance statistics, IJPRAI 18 (2004) 1339-1360
27. Huang, K., Yan, H.: Off-line signature verification using structural feature correspondence, Pattern Recognition 35 (2002) 2467-2477
28. Deng, P.S., Liao, H.Y.M., Ho, C.W., Tyan, H.R.: Wavelet based off line handwritten signature verification, Computer Vision and Image Understanding, 76 (1999) 173-190
29. Barua, S.: Neural Networks applied to computer security, Proc. SPIE (1992) 735-742
30. Frías-Martínez, E., Sánchez, A., Vélez, J.F. : Support vector machines versus multi-layer perceptrons for efficient off-line signature recognition", EAAI 19 (2006) 693-704
31. Rigoll, G., Kosmala, A.: Systematic Comparison between On-Line and Off-Line Methods for Signature Verification with Hidden Markov Models, Proc. ICPR (1998) 1755-1757
32. Justino, E., Yacoubi, E., Bortolozzi, F., Sabourin, R.: An off-Line signature verification system using HMM and graphometric features, Proc. Intl. Conf. DAS (2000) 211-222
33. Huang, K., Yan, H.: Off-line signature verification based on geometric feature extraction and neural network classification, Pattern Recognition 30 (1997) 9-17

34. Sabourin, R., Genest, G., Preteux, F.J.: Off-line signature verification by local granulometric size distributions", IEEE Trans. PAMI 19 (1997) 976-988
35. Ammar, M., Yoshida, Y., Fukumura, T.: A New Effective Approach for Off-line Verification of Signatures by Using Pressure Features, Proc. ICPR (1986) 566-569
36. Cardot, H., et al.: An artificial neural network architecture for handwritten signature authentication, Proc. of the SPIE, Applications Artificial Neural Networks II (1993) 633-644
37. Justino, E., Bortolozzi, F., Sabourin, R.: A comparison of SVM and HMM classifiers in the off-line signature verification, Pattern Recognition 26 (2005) 1377-1385
38. You, X.G., Fang, B., He, Z.Y., Tang, Y.Y. Similarity measurement for off-line signature verification, In: Advances In Intelligent Computing, LNCS 3644, Springer (2005) 272-281
39. Fang, B. et al.: Offline signature verification with generated training samples, IEE Proc. Vision, Image and Signal Processing 149 (2002) 85-90
40. Piyush Shanker, A., Rajagopalan, A.N.: Off-line signature verification using DTW, Pattern Recognition Letters 28 (2007) 1407-1414
41. Justino, E., Yacoubi, A. el, Bortolozzi, F., Sabourin, R.: An Off-Line Signature Verification System Using Hidden Markov Model and Cross-Validation, Proc. ICDAR (2000) 859-869
42. Coetzer, J., Herbst, B.M., Preez, J.A. du: Offline Signature Verification Using the Discrete Radon Transform and a Hidden Markov Model, EURASIP Journal on Applied Signal Processing 4 (2003) 559-571
43. Woo, Y.W., Han S., Jang, K.S.: Off-Line Signature Verification based on Directional Gradient Spectrum and a Fuzzy Classifier, LNCS 4319, Springer (2006) 1018-1029
44. Hanmandlu, M., Hafizuddin, M., Yusof, M., Madasu, V.K.: Off-line signature verification and forgery detection using fuzzy modelling, Pattern Recognition 38 (2005) 341-356
45. Xiao, X., Leedham, G.: Signature verification using a modified Bayesian network, Pattern Recognition 35 (2002) 983-995
46. Fang, B., et al.: A smoothness index based approach for off-line signature verification, Proc. ICDAR (1999) 785-787
47. Mizukami, Y., et al.: An off-line signature verification system using an extracted displacement function, Pattern Recognition Letters 23 (2002) 1569-1577
48. Oliveira, R., Kaestner, C., Bortolozzi, F., Sabourini, R.: Generation of signatures by deformations, Proc. BSDIA (1997) 283-298
49. Fang, B., et al.: Off-line signature verification by the tracking of feature and stroke positions", Pattern Recognition 36 (2003) 91-101
50. Justino, E., Bortolozzi, F., Sabourin, R.: Off-line signature verification using HMM for random, simple and skilled forgeries, Proc. ICDAR (2001) 1031-1034
51. Sabourin, R., Cheriet, M., Genest, G.: An extended shadow-code based approach for offline signature verification, Proc. DAS (1993) 1-5
52. Qiao, Y., Liu, J., Tang, X.: Offline Signature Verification Using Online Handwriting Registration", IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), (2007) 1-8
53. Hobby, J.D.: Using Shape and layout information to find signatures, text, and graphics, Computer Vision and Image Understanding 80 (2000) 88-110
54. Lamarche, F., Plamondon, R.: Segmentation and feature extraction of handwritten signature patterns", Proc. ICPR (1984) 756-759
55. Larrea, S., et al.: Eliminación del fondo de un cheque bancario (in Spanish), Proc. Terceras Jornadas de Reconocimiento Biométrico de Personas (2006)
56. Madasu, V.K., Lovell, B.C.: Automatic Extraction of Signatures from Bank Cheques and other Documents, Proc. DICTA (2003) 591-600

Verificación Automática de Firmas Manuscritas

Jesús F. Vargas[1,2], Miguel A. Ferrer[1], Carlos Travieso[1], and Jesus B. Alonso[1]

[1] CeTIC - Universidad de Las Palmas de Gran Canaria, Las Palmas 35017, Spain.
{mferrer, ctravieso, jalonso}@dsc.ulpgc.es
WWW home page: http://www.cetic.eu
[2] Universidad de Antioquia, Dpto de Ingeniería Electrónica, Medellin, Colombia.
jfvargas@udea.edu.co
WWW home page: http://electronica.udea.edu.co/

Resumen Se presenta un análisis de la información presente en una imagen estática de una firma manuscrita a partir de la detección de los puntos de alta presión. A partir de una imagen en escala de grises, se propone una modificación al método tradicional para establecer el umbral de alta presión. Empleando una versión binaria de la imagen original, y la imagen resultante del procedimiento de detección de puntos de alta presión, el problema se traslada al espacio polar, en donde se realiza un segmentación radial y angular del espacio de características para determinar en cada una de las celdas resultantes, la relación entre el número de puntos de alta presión y el número de pixeles de la imagen original. Finalmente, los dos vectores de características resultantes son empleados para entrenar dos modelos de clasificación basados en k-vecinos más cercanos y una red neuronal probabilística. Para la experimentación se empleo una base de datos que contiene muestras de 160 firmantes diferentes, con 24 muestras originales y 24 falsificaciones en cada caso. Se presentan los resultados obtenidos en las pruebas de verificación en términos de FAR, FRR y EER para falsificaciones de tipo *simple*, y una comparación con otros trabajos publicados y que emplean metodologías similares.

1. Introducción

Se ha planteado que la verificación de la identidad es un aspecto crucial dentro de la actual sociedad de información y comunicación. El número de situaciones en que se hace necesario un procedimiento rápido y de bajo costo para la autentificación de documentos, para el acceso o intercambio de información, y más aún para el comercio electrónico crece diariamente. Teniendo en cuenta la importancia que desde un punto de vista económico representan todas estas tareas, se presenta una gran dependencia de la efectividad de los sistemas de seguridad que tienen como objetivo evitar los accesos fraudulentos a dichos sistemas de información.

Para el caso de la verificación personal, es posible considerar dos tipos de medios biométricos: Los Fisiológicos, los cuales se derivan de una medición directa de partes del cuerpo humano; y los de Comportamiento, los cuales se derivan de la mediciones realizadas a partir de una acción ejecutada por un individuo y que permite caracterizarlo de una manera indirecta. Como ejemplos del primer caso se pueden citar la huella dactilar, la cara, la palma de la mano, y la retina entre otras. Dentro del segundo grupo

se pueden encontrar la voz, la firma, y su ritmo de tecleado en el ordenador [1].
Respecto a la firma manuscrita, históricamente ha sido uno de los medios con mayor aceptación dentro de la sociedad, de hecho diariamente son miles las verificaciones personales que se realizan por medio de una firma. Este tipo de procedimiento no representa ninguna incomodidad para la persona involucrada, como si sucede con muchas otras técnicas denominadas invasivas, y en las cuales el uso de ciertos dispositivos necesarios para la captura de información despiertan cierto recelo por parte de las personas (Un láser que recorre su ojo, o tener que hablar enfrente de un micrófono) [2]. El interés sobre esta técnica frente a otras características personales biométricas que podrían derivar en sistemas de verificación quizás más sencillos, es su utilización tradicional y su reconocimiento legal que hace que la firma manuscrita sea aún empleada en numerosas transacciones como por ejemplo en los cheques bancarios [3].
El análisis de la firma de un individuo solo puede ser llevado a cabo cuando dicha persona está (o estuvo) consciente y dispuesto a escribirla en la forma en que usualmente la hace, aunque podría darse el caso de estar intimidada en el momento de realizar la escritura. Caso contrario ocurre con otras técnicas en donde por ejemplo, la huella dactilar puede ser obtenida mientras la persona este inconsciente (p.e. drogada) [4]. Adicionalmente es un sistema de identificación que no provoca rechazos por parte de la persona identificada por lo que su uso puede ser deseable en circunstancias en que se desea mantener buenas relaciones públicas por ejemplo para identificar a clientes. La escritura presenta variaciones significativas de velocidad y movimientos musculares que son propias de cada escritor. El proceso de falsificación de una firma, si se quiere que sea exitoso, involucra un doble proceso en donde el falsificador además de copiar las características del escritor imitado debería ocultar sus propias características.

2. Sistemas de Verificación de Firmas Manuscritas tipo Offline

2.1. Descripción Teórica

La firma manuscrita es el resultado de un complejo proceso que depende de las condiciones físicas y psicológicas del firmante en el instante mismo en que realiza la firma y que pueden pues verse afectadas por variadas circunstancias. Algunas personas simplemente escriben su nombre mientras que otras realizan trazos que vagamente guardan alguna similitud con el suyo. Existen firmas complicadas, y otras sencillas de falsificar [2]. Es bien sabido que dos firmas, aún si han sido realizadas por la misma persona, no son exactamente iguales. Para algunos expertos en análisis de firmas, el hecho de que dos firmas realizadas sobre papel sean exactamente iguales, permite pensar que una de las dos es una falsificación. Las firmas sucesivas de una misma persona presentarían siempre diferencias globales y locales, e incluso pueden diferenciarse en su escala y orientación. Dadas este tipo de variaciones, es común encontrarse con que un experto en firmas tenga éxito detectando falsificaciones pero que falle cuando intente verificar una firma auténtica [5].
Lo anterior convierte a la verificación de firmas manuscritas (VFM) en un difícil problema de discriminación, ya que la firma será altamente variable y por tanto su verificación no trivial, aún para los expertos humanos. Es por esto que la verificación de firmas manuscritas atrae la atención de un importante número de investigadores que

intentan generar técnicas biométricas para la verificación personal [6],[7]

2.2. Fundamentos de VFM

Falsificaciones de una firma Como se menciono anteriormente, un sistema de verificación de firmas clasifica una firma como genuina o falsa. Dependiendo de la calidad y la dedicación con que ha sido elaborada la falsificación de una firma, se habla de tres tipos de falsificación:

- Aleatoria: Cuando el falsificador no conoce la firma original, y en el mejor de los casos tan solo conoce el nombre de la persona a quien intentará falsificar su firma.
- Poco Elaborada o simple: Cuando el falsificador tiene la oportunidad de observar brevemente la firma original, e intenta realizar inmediatamente una falsificación de la misma.
- Muy Elaborada: Es el caso en el que el falsificar puede observar la firma original, y además tiene la oportunidad de practicar su falsificación tanto como él crea necesario. Teniendo en cuenta lo anterior, es claro que los resultados de los sistema de verificación suelen presentar una confiabilidad más baja cuando empleamos falsificaciones del tipo muy elaboradas.

Tipos de Error Al igual que en cualquier aplicación de reconocimiento de patrones, para un sistema de verificación de firmas manuscritas se tienen en cuenta dos tipos de error.

- Tasa de Falsos Rechazos: Generalmente notado como FRR (False Rejection Ratio), representa la relación entre la cantidad de firmas que han sido erróneamente rechazadas por el sistema, y el número total de firmas evaluadas. Se denomina también error tipo I.
- Tasa de Falsas Aceptaciones: Generalmente notado como FAR (False Acceptance Ratio), representa la relación entre la cantidad de firmas que han sido erróneamente aceptadas por el sistema, y el número total de firmas evaluadas. Se denomina también error tipo II. Como objetivo del sistema, ambos porcentajes deberán ser tan bajos como sea posible, pero es importante decir que suele ser castigado con una susceptibilidad mayor el tipo de error II, es decir, es más crítico para el sistema verificar como verdadera una firma que en realidad es falsa, que el caso contrario.

Complementariamente, existe una estadística utilizada para mostrar el rendimiento biométrico de una aplicación; por lo general, durante la tarea de verificación. La Tasa de igual error EER (Equal Error Rate) es la ubicación en una curva ROC (Característica de funcionamiento del receptor) donde el error tipo I y el error tipo II son iguales. Por lo general, cuánto más bajo sea el valor de la tasa de igual error, mayor será la precisión del sistema biométrico.

Preprocesamiento de Imágenes Considerando un sistema tipo Offline, en donde se dispone de un documento digitalizado, es necesario realizar algunos procedimientos previos con el fin de eliminar elementos no deseados o que no tienen interés para el sistema de verificación de firmas. Teniendo en cuenta que en una aplicación real la firma de una persona está incluida en un documento que puede contener además otros elementos como texto (tipográfico y/o manuscrito), imágenes, sellos y otros similares, es necesario realizar la extracción de la firma. Tal vez unas de las aplicaciones que despierta el mayor interés de los investigadores y que contempla la presencia de estos elementos mencionados, corresponde con el preprocesamiento de imágenes de cheques bancarios.
La presencia de ruido en la imagen que contiene la firma que se quiere analizar, generalmente denominado como de “sal y pimienta", puede ser eliminado aplicando filtros de mediana y operadores morfológicos. Como ya se mencionó, la firma manuscrita presenta alta variabilidad interpersonal, lo que contempla la posibilidad de encontrar firmas de diferente tamaño realizadas por la misma persona; lo anterior afecta fuertemente algunos procedimientos de caracterización de la firma, lo que hace necesario el empleo de procedimientos orientados a realizar una normalización en tamaño de las muestras. Muchos de los procedimientos planteados por los investigadores, consideran una imagen en blanco y negro, razón por la cual se realiza el correspondiente algoritmo de binarización. También es común el procedimiento conocido como Esqueletización, en donde los trazos que conforman la firma, son adelgazados hasta que tienen grosor de 1 pixel.

3. Base de Datos GPDSsignature

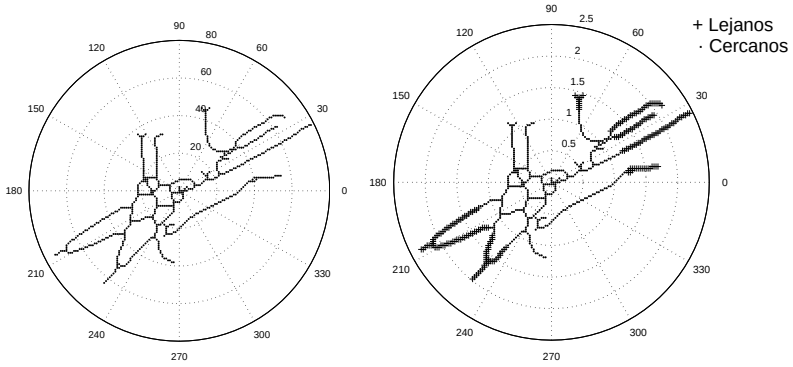
La base de datos GPDSsignature es de tipo Offline. Contiene registros de 160 firmantes. De cada uno de los firmantes se tienen 24 muestras de firmas originales y 24 falsificaciones de tipo *Elaboradas Simples*. Para realizar las falsificaciones, se facilito una imagen de la firma original, y se permitió practicar la falsificación tanto como el falsificador deseó. Las muestras originales facilitadas a los falsificadores fueron diferentes en cada caso.
Para la recolección de las muestras tanto originales como falsificaciones, se emplearon plantillas que contienen cuadriculas de dos tamaños diferentes, con tamaños 5 x 1.8 cm y 4.5 x 2.5 cm respectivamente. Se cuenta con 12 firmas realizadas en cada una de estos tamaños de cuadriculas. Las plantillas han sido escaneadas con una resolución de 300ppp y almacenadas en formato *png* en escala de grises.

4. Cálculo de Características

4.1. Densidad de pixeles en coordenadas Polares

Una vez preprocesada la imagen de la firma manuscrita, la matriz de datos se transforma a coordenadas polares en donde se estima la densidad de pixeles en diferentes

zonas, empleando para ello una malla que divide el espacio polar en secciones de angulo de tamaño θ. La figura 1(a) muestra la conformación de dicha malla en el espacio polar.



(a) Malla formada en coordenadas polares

(b) Elementos cercanos y lejanos

Figura 1. Transformación de imagen al espacio polar

Una vez establecida la malla, se busca el elemento que este ubicado sobre los valores de ángulos 0, 90, 180, o 270 (los cuatro ejes en coordenadas cartesianas) y que tenga el mayor valor de radio. Los valores de radio de todos los elementos de la firma son normalizados de tal forma que el radio del elemento encontrado sea igual a 1. Una vez normalizados, los elementos de la firma son clasificados como *cercanos* ó *lejanos*, según un umbral que corresponde a la mitad del radio mas grande presente en la firma. La figura 1(b) muestra el resultado del procedimiento mencionado.

A continuación se calcula el número de pixeles contenidos en cada una de las celdas que conforman la malla. Esta operación se realiza a través del cálculo de un histograma en donde los limites para el cálculo del mismo corresponden con los valores de angulo que limitan las celdas de la malla.
Como se mencionó anteriormente, se tienen en cuenta dos zonas: la zona mas cercana que corresponde a las celdas que contienen elementos con valor de radio menor que el umbral, y la denominada zona lejana a la que corresponden los elementos ubicados mas allá del umbral.
Finalmente, se obtienen dos vectores conteniendo la densidad normalizada (respecto al número total) de pixels en cada celda y ordenados según los valores de angulo en el sentido de manecillas de reloj.

4.2. Puntos de Alta Presión

Cuando se analiza una imagen en escales de grises que contiene una firma manuscrita escaneada, es posible mencionar que aquellos trazos que han sido realizados ejerciendo una mayor presión sobre el bolígrafo, aparecen representados con los niveles mas oscuros. Teniendo en cuenta lo anterior, los puntos de alta presión corresponden con los pixeles pertenecientes a la firma cuyo valor esta por encima de un umbral determinado. Las características de alta presión fueron planteadas en un principio por Ammar et el. [8], quien planteó el umbral mencionado de la siguiente manera

$$\theta_{hpr} = g_{\text{mín}} + 0{,}75\,(g_{\text{máx}} - g_{\text{mín}}) \tag{1}$$

en donde g_{min} y g_{max} corresponden a los niveles mínimo y máximo de intensidad la escala de grises de la imagen. Este planteamiento ha sido usado también por Huang et al. [9], y Sansone et al.[10], en sus respectivos trabajos. Recientemente Mitra et al.[11], propuso un procedimiento diferente para la elección del umbral; a partir de la información de densidad de niveles de gris en la imagen original, el umbral se selecciona adaptativamente como el punto que corresponde con $1/sqrt(2)$ de la frecuencia pico (valor de gris con mayor número de elementos dentro de la imagen). Lv et al.[12], propone una metodología diferente, en donde se establecen dos umbrales con el fin de retener solamente los pixeles pertenecientes al contorno y a las área de mayor interés. Basado en los resultados experimentales, los dos umbrales son establecidos para los valores de nivel de gris de 85 y 205. El rango establecido de divide posteriormente en 12 segmentos, dentro de los cuales se calcula el porcentaje de pixeles contenido en cada uno ellos.

Modificación propuesta para el cálculo del umbral. A partir de la información del histograma en escala de grises de la imagen, se estiman el valor más frecuente G_{maxHis} y el valor mas bajo de gris presente en la imagen G_{min}

$$\begin{aligned} hisI &= histograma\,(I) \\ G_{\text{máx}\,His} &= \text{máx}\,(hisI) \\ G_{\text{mín}} &= min\,(hisI > mean\,(hisI)) \end{aligned} \tag{2}$$

A partir de la ecuación 2 el umbral de alta presión HPPthresh se define como

$$HPPthresh = G_{\text{mín}} + \gamma\,(G_{\text{máx}\,His} - G_{\text{mín}}) \tag{3}$$

en donde γ es un factor determinado empíricamente. La figura 2 ilustra el calculo del umbral de alta presión propuesto.

Una vez estimado el umbral, los puntos de alta presión presentes en la imagen de la firma manuscrita analizada, se definen como

$$I_{ij}^{HPP} = \begin{cases} 1, & \forall I_{ij} \leq HPPthresh \\ 0, & otro\ \ caso \end{cases} \tag{4}$$

en donde I_{ij}^{HPP} será la matriz imagen resultante que contendrá solamente los puntos de alta presión de la imagen original, con $i = 1, 2, 3, \ldots, M$ y $j = 1, 2, 3, \ldots, N$, siendo

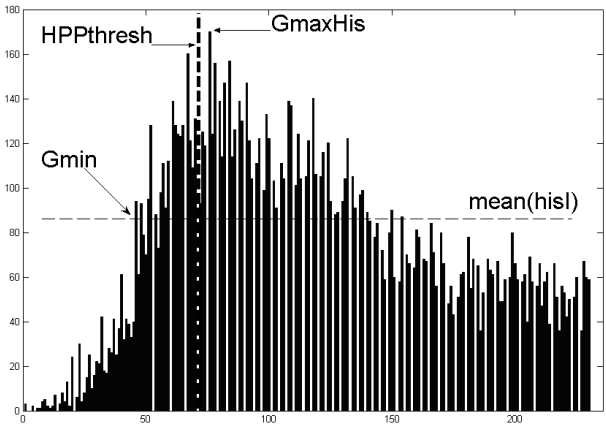


Figura 2. Estimación del umbral de alta presión

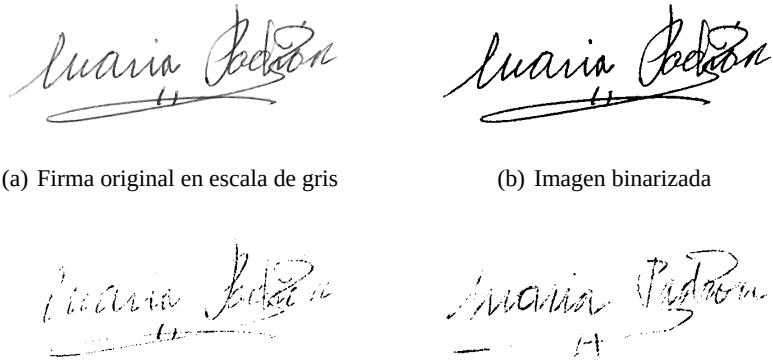
M y N, el número de filas y columnas de la imagen original respectivamente. La figura 3 muestra el proceso de detección de los puntos de alta presión en una muestra genuina y una falsificación de la misma.

4.3. Densidad de Pixeles de Alta Presión en Coordenadas Polares - Metodología propuesta

Para este caso, tanto una versión binarizada de la imagen original, como la imagen resultante en el procedimiento de detección de puntos de alta presión, son transformadas a coordenadas polares, en donde son centradas empleando para ello el cálculo del centro geométrico de la imagen original. Una vez aquí, se segmenta el espacio polar empleando el parámetro θ descrito anteriormente. En cada una de las secciones angulares resultantes, se calcula la relación entre el número de pixeles de la imagen binaria, y el número de puntos de alta presión, esto es

$$HPPPD_k = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij}^{HPP} \in \theta_k}{\sum_{i=1}^{M}\sum_{j=1}^{N} I_{ij}^{bin} \in \theta_k} \tag{5}$$

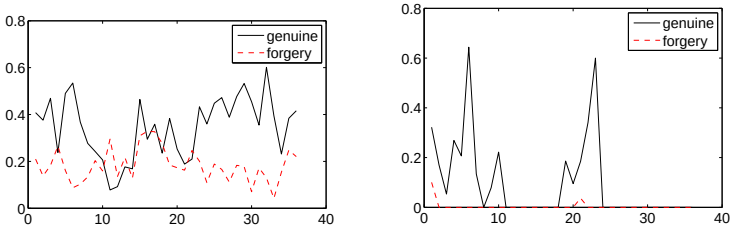
en donde I_{ij}^{bin}, con $i = 1, \ldots, M$ y $j = 1, \ldots, N$, representa la versión binaria de la imagen original I. La figura 4 muestra una comparación entre los vectores de características (de puntos cercanos y lejanos) obtenidos para una muestra original y una falsificación de una firma de la base de datos empleada.



(a) Firma original en escala de gris (b) Imagen binarizada

(c) HPP estimados para original (d) HPP estimados para falsificación

Figura 3. Proceso de detección de puntos de alta presión



(a) HPPPD para puntos cercanos (b) HPPPD para puntos lejanos

Figura 4. Comparación de HPPPD calculada para muestras original y falsificación

5. Clasificación

Una vez que se han estimado los vectores de características, es necesario resolver un problema de clasificación biclase. A continuación se presenta una breve descripción de las metodologías de clasificación empleadas en el presente trabajo.

5.1. K-Vecinos más Próximos

Comúnmente denominado k-nn por sus siglas en inglés (nearest neighbor), es un método que permite clasificar una nueva entrada a partir de las distancias mínimas con patrones de entrenamiento en el espacio de características.
Para una nueva entrada **x**, la regla knn determina el cluster más cercano basado en alguna métrica. Aunque han sido muchas las métricas planteadas para atacar este problema, la más común es la distancia Euclidiana. Después de encontrar los k-vecinos más próximos a **x**, existen dos esquemas diferentes para determinar su etiqueta final, el voto por mayorías y el voto por sumatoria de pesos. En el primer caso, el voto por mayorías,

la nueva muestra es asignada a la clase que aparece con mayor frecuencia en los k-vecinos. Para el segundo caso, el voto por sumatoria de pesos, cada voto es ponderado con la premisa de que los vecinos más cercanos tiene una importancia mayor que lo más lejanos.
Para la clasificación de un patrón, el algoritmo k-nn solo del parámetro k que es un número entero, un conjunto de muestras de entrenamiento para crear las etiquetas de las clases, y la elección de una métrica para medir las distancias. Teniendo en cuenta lo anterior, la implementación del algoritmo es relativamente simple y directa. Para el presente trabajo y basados en los resultados experimentales, el valor de k se estableció en 3.

5.2. Red Neuronal Probabilistica

La red neuronal probabilística, comúnmente denominada PNN por sus siglas en inglés (Probabilistic Neural Network), fue planteada inicialmente por Donal Specht en 1988, y es un algoritmo de entrenamiento de 3 capas, alimentado hacia adelante y de un solo paso usado para la clasificación y el mapeo de datos [13]. A diferencia de otras redes neuronales como la basada en el algoritmo de *Back-propagation*, este tipo de algoritmo esta basado en estimadores no paramétricos de la función densidad de probabilidad basados en kernels. Una de las ventajas de la PNN, es que esta garantiza que las funciones de densidad de probabilidad de las clases sean suaves y continuas.
Las PNN usan estimadores de la función de probabilidad del tipo Parzen, que de manera asintótica se aproximan a la distribución real de los datos, permitiendo que esta sea suave y continua. La PNN emplea funciones de base radial gaussiana esféricas centradas en cada uno de los vectores de entrenamiento. La probabilidad de que un vector nuevo sea asignado a una clase determinada esta dada por

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}\sigma^p M_i} \sum_{j=1}^{M_i} \frac{\exp\left(-(x-x_{ij})^T (x-x_{ij})\right)}{2\sigma^2} \tag{6}$$

en donde i es el número de la clase, j es el número del patrón, x_{ij} es el $j-esimo$ vector de entrenamiento de la clase i, p es la dimensión del vector x, σ es el factor de suavizado (desviación estándar) y $f_i(x)$ es la suma de las gaussiana esféricas centradas en cada uno de los vectores de entrenamiento x_{ij} para la $i-esima$ función de densidad de probabilidad de clase estimada.
La decisión final se toma de acuerdo a la estrategia Bayesiana y será $d(x) = C_i$, si se tiene que

$$f_i(x) > f_k(x) \qquad para \qquad k \neq i \tag{7}$$

teniendo en cuenta que C_i corresponde a la clase i.

6. Protocolo de Pruebas

6.1. Experimentos

Las muestras genuinas y falsificaciones se han dividido a partes iguales en dos grupos escogidos aleatoriamente (12 originales y 12 falsificaciones). El primero de estos

grupos se emplea en la etapa de entrenamiento, mientras el segundo se reserva para la etapa de verificación. Con el fin de obtener resultados estadísticamente representativos y generalizantes, las pruebas se repiten 5 veces en cada caso, realizando la división en grupos de forma aleatoria en cada ocasión. Para medir el desempeño del sistema se han calculado los errores Tipo I y II, y se ha estimado el Equal Error Rate (EER) para cada una de las pruebas realizadas. Los resultados presentados a continuación corresponden a pruebas de verificación teniendo en cuenta falsificaciones del tipo “Simple”.

6.2. Resultados

En la Tabla 1 se presentan los resultados obtenidos en las pruebas realizadas para determinar el valor óptimo del parámetro θ qué permite realizar la división angular en el espacio polar de características, para este caso el clasificador empleado corresponde al modelo K-nn. La tabla 2 presenta los resultados para el caso del clasificador PNN. Como se puede apreciar, el mejor desempeño para los dos clasificadores analizados, se alcanza al realizar una división angular de 90^o.

Segmento Ang.	(%)FAR	(%)FRR	(%)EER	(σ)FAR	(σ)FRR
5	16.05	2.56	9.30	7.81	3.78
10	15.51	2.68	9.09	7.19	3.74
15	15.74	2.57	9.16	7.41	3.55
20	15.48	3.22	9.35	7.42	3.84
30	15.32	3.25	9.29	6.72	3.53
45	14.20	3.30	8.75	6.66	3.94
60	13.70	4.06	8.88	6.95	4.13
90	12.08	3.27	7.67	6.41	3.46

Tabla 1. Resultados para el clasificador KNN

Segmento Ang.	(%)FAR	(%)FRR	(%)EER	(σ)FAR	(σ)FRR
5	27.07	2.92	14.99	2.05	2.50
10	12.68	4.48	8.58	2.09	2.77
15	12.83	4.16	8.49	2.12	2.62
20	12.44	4.87	8.65	2.27	2.44
30	12.09	4.83	8.46	2.26	2.43
45	11.47	4.73	8.10	2.31	2.51
60	10.86	4.46	7.66	2.59	2.54
90	10.64	3.49	7.06	2.45	2.55

Tabla 2. Resultados para el clasificador PNN

Las tablas 3 y 4, nos muestran los resultados obtenidos para diferentes valores del parametro γ, que permite determinar la ubicación del umbral de alta presión. De los

datos observados se deduce que el valor para el cual ambos clasificadores ofrecen un mejor desempeño corresponde a 0.95.

Si realizamos una comparación entre los resultados obtenidos con cada uno de los clasificadores empleados en el presente trabajo (KNN y PNN), es necesario recalcar la mayor estabilidad ofrecida por el modelo PNN teniendo en cuenta los valores más bajos para la desviación estándar medida en las pruebas.

Umbral	(%)FAR	(%)FRR	(%)EER	(σ)FAR	(σ)FRR
0.95	12.62	3.16	7.89	6.81	3.36
0.90	12.58	3.62	8.10	7.01	3.78
0.85	12.73	3.92	8.33	6.75	4.21
0.80	13.13	4.65	8.89	7.04	5.10
0.75	13.73	4.88	9.30	6.76	5.16
0.70	13.74	5.67	9.71	6.80	5.44
0.65	14.11	6.74	10.43	6.67	5.74
0.60	14.98	8.47	11.73	6.75	6.60

Tabla 3. Resultados para el clasificador KNN

Umbral	(%)FAR	(%)FRR	(%)EER	(σ)FAR	(σ)FRR
0.95	10.34	3.66	7.00	2.74	2.57
0.90	11.48	3.54	7.51	2.46	2.66
0.85	12.09	3.71	7.90	2.47	2.82
0.80	12.77	4.29	8.53	2.48	2.49
0.75	13.89	4.54	9.21	2.69	2.90
0.70	15.58	4.89	10.24	2.89	3.06
0.65	18.38	5.56	11.97	2.75	3.94
0.60	20.77	8.94	14.85	4.19	5.32

Tabla 4. Resultados para el clasificador PNN

6.3. Comparación con trabajos similares publicados

Teniendo en cuenta que aún siguen siendo pocas las bases de datos de dominio público, realizar una comparación entre trabajos y metodologías planteadas en el área de verificación de firma manuscritas, sigue siendo una tarea complicada. Con el ánimo de completar el presente estudio, en las Tablas 5 y 6 se presenta una relación de los resultados publicados por diferentes autores, en donde se ha analizado igualmente el uso de características basadas en puntos de alta presión presentes en una firma, así mismo se brindan los datos correspondientes a las bases de datos empleadas en cada uno de estos trabajos.

	# Firmantes	Muestras Originales	Muestras Falsific.
Ammar et al. [8]	20	10	10
Lv et al.[12]	20	25	30
Huang et al. [9]	21	24	24
Sansone et al [10]	49	20	10
Mitra et al. [11]	20	10	10
Ferrer et al. [14]	160	24	24
Propuesto	160	24	24

Tabla 5. Descripción de bases de datos usadas en este y otros trabajos similares

	(%)FAR	(%)FRR	(%)EER
Ammar et al. [8]	6.5	4.00	5.25
Lv et al.[12]	5.30	4.60	5.00
Huang et al. [9]	11.80	11.10	11.45
Sansone et al [10]	4.29	2.04	3.16
Mitra et al. [11]	2.50	4.00	3.25
Ferrer et al. [14]	12.60	14.10	13.35
Propuesto	10.34	3.66	7.00

Tabla 6. Comparación de resultados con trabajos similares

7. Conclusiones

Se ha presentado una nueva metodología para la extracción de los puntos de alta presión basada en características geométricas de la imagen en escala de grises de una firma manuscrita. La metodología hace uso de una transformación de la imagen hacia el espacio polar en donde se determina la distribución de los puntos de alta presión, así como la densidad de los mismos respecto a los puntos que conforman la imagen original en cada una de las regiones angulares generadas para el análisis. La figura 5 presenta el diagrama de flujo del sistema propuesto. Los resultados experimentales muestran un desempeño aceptable del sistema (EER=7 %), siendo incluso mejor que algunos otros trabajos publicados en el área. Los porcentajes de error obtenidos se acercan a los alcanzados por otros planteamientos en donde se hace uso de una combinación de diferentes parámetros.

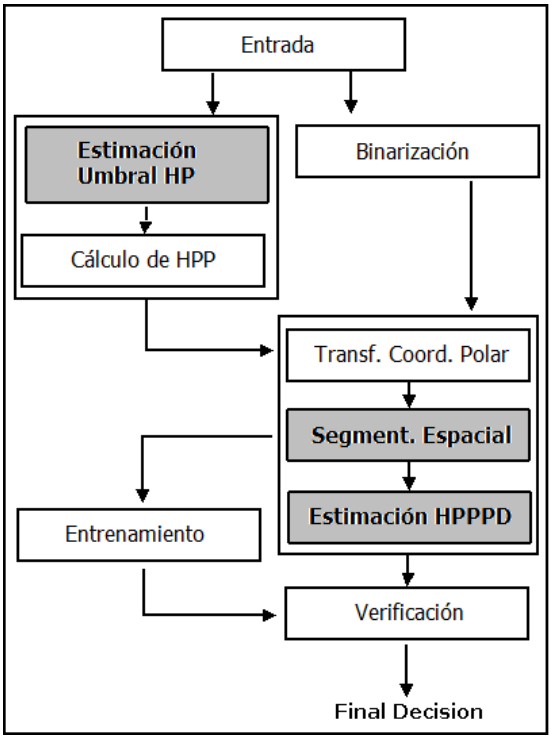


Figura 5. Diagrama de flujo del sistema de verificación de firmas propuesto.

Agradecimientos

Este trabajo se desarrolla en el marco del proyecto de la Comunidad Autónoma de Canarias PI042005/134, así mismo con la colaboración del Gobierno Español a través del proyecto MEC TEC2006-13141-C03/TCM. F. Vargas es beneficiario del programa de becas de alto nivel para América Latina - *Programa Alβan* E05D049748CO.

Referencias

1. S. Liu and M. Silverman. A practical guide to biometric security technology. *IEEE IT Professional*, 3(1):27–32, 2001.
2. R. Plamondon and S.N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(1):63–84, 2000.
3. S. Larrea, M.A. Ferrer, C.M. Travieso, J.F. Vargas, and J.B. Alonso. Eliminación del fondo de un cheque bancario. In *Proceedings of the Terceras Jornadas de Reconocimiento Biométrico de Personas*, pages 83–102, Nov. 2006.
4. K. Franke, J. Ruiz del Solar, and M. Köpen. Soft-biometrics: Soft computing for biometric-applications. Technical report, IPK, 2003.
5. N.M. Herbst and C.N. Liu. Automatic signature verification based on accelerometry. Technical report, IBM J.Res.Dev., 1977.
6. M.C. Fairhurst and E. Kaplani. Perceptual analysis of handwritten signatures for biometric authentication. In *IEE Proceedings Vision, Image and Signal Processing*, pages 389–394, Dec. 2003.
7. J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V.Espinosa, A.Satue, I. Hernaez, J.-J. Igarza, C. Vivaracho, D. Escudero, and Q.-I. Moro. Mcyt baseline corpus: a bimodal biometric database. In *IEE Proceedings of Visual Image Signal Processsing*, Dec. 2003.
8. M. Ammar, Y. Yoshida, and T. Fukumura. A new effective approach for automatic off-line verification of signatures by using pressure features. In *in Proceedings 8th International Conference on Pattern Recognition*, pages 566–569, 1986.
9. K. Huang and H. Yan. Off-line signature verification based on geometric feature extraction and neural network classification. *Pattern Recognition, Elsevier Science*, 30(1):9–17, 1997.
10. C. Sansone and M. Vento. Signature verification: Increasing performance by a multi-stage system. *Pattern analysis & Applications, Springer*, 3:169–181, 2000.
11. A. Mitra, P. Kumar, and C. Ardil. Automatic authentification of handwritten documents via low density pixel measurements. *International Journal of Computational Inteligence*, 2(4):219–223, 2005.
12. H. Lv, W. Wang, C. Wang, and Q. Zhuo. Off-line chinese signature verification based on support vector machine. *Pattern Recognition Letters, Elsevier*, 26:2390–2399, 2005.
13. D. F. Specht. Probabilistic neural networks. *Neural Networks*, 3:109–118, 1990.
14. M. Ferrer, J. Alonso, and C. Travieso. Offline geometric parameters for automatic signature verification using fixed-point arithmetic. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):993–997, 2005.

Verificación de firma off-line usando características de contorno

Almudena Gilpérez, Fernando Alonso-Fernandez, Julian Fierrez, Javier Ortega-Garcia

Biometric Recognition Group - ATVS
Escuela Politecnica Superior - Universidad Autonoma de Madrid
Avda. Francisco Tomas y Valiente, 11 - 28049 Madrid, España
{almudena.gilperez,fernando.alonso,julian.fierrez,javier.ortega}@uam.es
http://atvs.ii.uam.es

Abstract. En este trabajo, presentamos un sistema de verificación de firma off-line basado en características de contorno. Las imágenes se analizan a nivel local, codificándose propiedades direccionales de los trazos, así como de longitud interna entre huecos de los mismos. Los resultados, obtenidos a partir de un subconjunto de firmas de la base de datos MCYT, muestran que las características direccionales funcionan mucho mejor que las que analizan la longitud interna. También observamos que la combinación de las características propuestas no proporciona mejoras adicionales debido a la posible correlación existente entre ellas.

1 Introducción

El creciente interés en la biometría se debe al gran número de aplicaciones donde la correcta identificación de individuos es un hecho crucial [1]. En este documento, tratamos el problema de la verificación automática de personas a partir de imágenes escaneadas de su firma (llamadas firmas *off-line*, en contraposición a *on-line*, donde se registra el propio acto de firma mediante tabletas capaces de capturar la trayectoria y presión del bolígrafo). La firma es uno de los métodos de autenticación más usados actualmente debido a su aceptación en entornos gubernamentales, legales, financieros y comerciales [3, 2]. Cabe señalar que incluso examinadores forenses alcanzan unas tasas de acierto en reconocimiento de sólo el 70%, confirmando por tanto que la firma off-line es un área aún con grandes retos por resolver [4].

En este trabajo, hemos implementado un sistema de verificación de firma off-line basado en características propuestas para reconocer escritores mediante imágenes de documentos manuscritos [5]. Hemos seleccionado y adaptado un conjunto de características para ser usadas con firmas manuscritas, las cuales se basan en el análisis local de imágenes. Las características implementadas trabajan a partir del análisis del contorno de las firmas. Éstas son consideradas como una textura descrita por distribuciones de probabilidad calculadas a partir de su imagen, las cuales capturan la apariencia visual distintiva de cada muestra. Por tanto, la identidad de cada usuario se codifica mediante distribuciones de

probabilidad (PDF) extraídas de las imágenes. El término "característica" se usa para denotar cada PDF, siendo un vector de probabilidades que captura la unicidad de cada firma.

El resto de este documento está organizado en varias partes: una descripción del sistema implementado en la Sección 2; el marco experimental utilizado, la base de datos, protocolo y resultados en la Sección 3 y finalmente, las conclusiones de nuestro trabajo en la Sección 4.

2 Sistema basado en características de contorno

Nuestro sistema de verificación de firma está dividido en tres etapas: *i*) preprocesado de la firma, *ii*) extracción de características, y *iii*) comparación de características. Estas tres partes se describen a continuación.

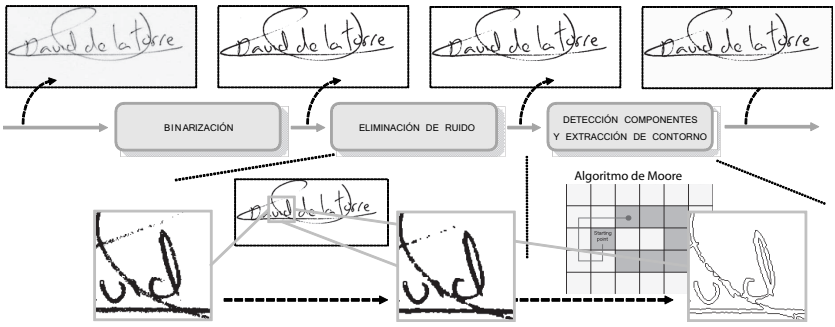


Fig. 1. Fase de preprocesado del sistema.

2.1 Etapa de preprocesado

El objetivo de esta etapa es mejorar las imágenes y adaptarlas a la fase de extracción de características posterior. El preprocesado se divide en cuatro partes, como se muestra en la Figura 1: binarización, eliminación del ruido, detección de componentes y extracción del contorno.

En primer lugar, la imagen escaneada se binariza usando el método de Otsu [6]. Este método funciona correctamente cuando la imagen se caracteriza por un fondo uniforme y objetos de interés similares, como en el caso de las imágenes de las firmas. Además, no necesita la supervisión humana o información previa antes de su ejecución. El siguiente paso es la eliminación de ruido de la imagen binaria, que se realiza a través de operaciones morfológicas, una de apertura seguida de una operación de cierre [7]. Luego, pasaríamos a la detección de componentes, usando conectividad 8. Finalmente, extraemos los contornos internos y externos de los componentes conectados utilizando el algoritmo de Moore [7]. A partir

de un píxel de contorno de un componente conectado, que es establecido como inicio, este algoritmo busca en el sentido de las agujas del reloj un pixel frontera a su alrededor, y repite este proceso hasta que llega al pixel con el que empezamos a analizar el algoritmo. El resultado es una secuencia o vector con los píxeles de las coordenadas de la frontera del componente. Esta representación vectorial es muy eficaz porque permite una rápida extracción de muchas de las características utilizadas más tarde.

	Característica	**Explicación**	**Dimensiones**	**Fuente**
f1	$p(\phi)$	Contour-direction PDF	12	contours
f2	$p(\phi_1, \phi_2)$	Contour-hinge PDF	300	contours
f3h	$p(\phi_1, \phi_3)_h$	Direction co-occurrence PDF, horizontal	144	contorno
f3v	$p(\phi_1, \phi_3)_v$	Direction co-occurrence PDF, vertical	144	contorno
f5h	$p(rl)_h$	Run-length on background PDF, horizontal	60	imagen binaria
f5v	$p(rl)_v$	Run-length on background PDF, vertical	60	imagen binaria

Table 1. Caracteristicas usadas en este trabajo.

2.2 Etapa de extracción de características

Las características usadas en este trabajo se extraen a partir de dos representaciones de la firma obtenidas durante la etapa de preprocesado: la imagen binaria sin ruido, y los contornos de los componentes conectados. Las características usadas en este trabajo están resumidas en la Tabla 1, incluyendo la representación de la firma que usa cada una. La firma se modela como una textura que se describe con funciones de distribución de probabilidad (PDFs). Las funciones de probabilidad de distribución usadas aquí se agrupan en dos categorías diferentes: PDFs de dirección (características: f1, f2, f3h, f3v) y PDFs de longitud (características: f5h, f5v). En la Figura 2, observamos una descripción gráfica de la extracción de las PDFs de dirección. Para ser coherentes con los trabajos donde se proponen estas características [5], seguimos la misma nomenclatura usada en ellos.

Contour-Direction PDF (f1)

Esta distribución direccional se calcula muy rápidamente utilizando el contorno de la firma, con la ventaja adicional de que se elimina la influencia del ancho de los trazos de tinta. Esta característica se extrae considerando la orientación local de fragmentos del contorno. Un fragmento se determina por dos píxeles de contorno(x_k,y_k) y $(x_{k+\epsilon},y_{k+\epsilon})$ donde ϵ es una distancia de separación para poder calcular el ángulo que existe entre el fragmento del contorno y el eje horizontal. Este ángulo se calcula de la siguiente manera:

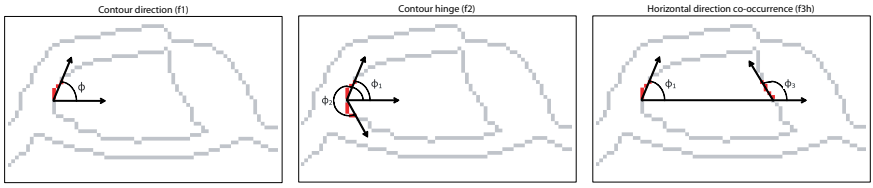


Fig. 2. Descripción gráfica de la extracción de características. De izquierda a derecha: contour direction (f1), contour hinge (f2) y horizontal direction co-occurrence (f3h).

$$\phi = \arctan\left(\frac{y_{k+\epsilon} - y_k}{x_{k+\epsilon} - x_k}\right) \tag{1}$$

A medida que recorremos el contorno, construimos el histograma de los ángulos. Seguidamente, normalizamos este histograma para obtener la PDF f1, que nos informa de cuál es la probabilidad de encontrar en la firma un fragmento de contorno orientado con cada ϕ. El ángulo ϕ reside en los dos primeros cuadrantes, ya que, sin información *on-line*, no sabemos la inclinación con la que el escritor firmó. El histograma abarca un intervalo de 0°-180°, y se divide en $n = 12$ secciones (cajas). Por tanto, cada sección se expande 15°, con la que observamos suficientes detalles como para tener una descripción robusta de cada firma [5]. Además, hemos establecido un $\epsilon = 5$. Estos datos serán utilizados para todas las PDFs de dirección presentadas en este documento.

Contour-Hinge PDF (f2)

Con el fin de captar la curvatura del contorno, así como su orientación, utilizamos la característica "hinge" f2 (o bisagra). La idea principal es considerar dos fragmentos de contorno unidos por un pixel común y calcular la distribución de probabilidad conjunta de las orientaciones ϕ_1 y ϕ_2 de los dos lados. Se obtiene así una función de densidad conjunta, que cuantifica la posibilidad de encontrar dos fragmentos de contorno unidos y formando ángulos ϕ_1 y ϕ_2. Esta PDF se calcula en los (360°) y habrá $2n$ secciones, pero sólo las combinaciones no redundantes se analizan (i.e. $\phi_2 \geq \phi_1$). Para $n = 12$, el vector resultado de esta característica tiene 300 dimensiones [5].

Direction Co-Occurrence PDFs (f3h, f3v)

Basándonos en la misma idea de combinar fragmentos de contorno orientados, utilizamos la "co-ocurrence direction". Para esta característica, usamos la combinación de los ángulos al final de segmentos contenidos en los huecos de los trazos, ver Figura 2. Esta característica se calcula para segmentos horizontales (f3h) y verticales (f3v). También son funciones de densidad conjunta, expandidas en los dos primeros cuadrantes, y divididos en n^2 secciones. Estas características

dan una medida de la redondez de los caracteres y/o de los trazos de la firma.

Run-Length PDFs (f5h, f5v)

Estas características se calculan a partir de la imagen binarizada de la firma utilizando los pixeles correspondientes al fondo. Se capturan las regiones encerradas dentro de las letras y los trazos, además de los espacios vacíos entre ellos, calculando su longitud. Analizamos la distribución de probabilidad horizontal (f5h) y verticalmente (f5v).

2.3 Etapa de comparación de características

Cada cliente del sistema se representa por una PDF que se calcula usando un conjunto de K firmas. Para cada característica, se calcula el histograma de las K firmas y después se normaliza a una distribución de probabilidad.

Para obtener la similitud entre una identidad q y una firma dada i, usaremos la distancia χ^2 [5]:

$$\chi^2_{qi} = \sum_{n=1}^{N} \frac{(p_q[n] - p_i[n])^2}{p_q[n] + p_i[n]} \tag{2}$$

donde p son las entradas del vector de PDFs, y N es la dimensión del vector.

Asimismo, realizamos experimentos que combinan las diferentes características. La distancia final en este caso se calcula como el valor medio de las distancias Hamming debido a las características individuales:

$$H_{qi} = \sum_{n=1}^{N} |p_q[n] - p_i[n]| \tag{3}$$

La distancia χ^2, debido al denominador, da más peso a las regiones de baja probabilidad de la PDF y maximiza la actuación de cada característica individual. Por otro lado, la distancia de Hamming proporciona valores comparables de distancia para las característica individuales [5].

3 Experimentos

3.1 Base de Datos y Protocolo

Hemos utilizado un subconjunto de la base de datos MCYT [8], la cual incluye firmas *on-line* y *off-line* de 330 usuarios de 4 sitios españoles diferentes. También incluye falsificaciones entrenadas de firma. Los falsificantes vieron las firmas de los clientes para poder imitar su forma y se les permitió entrenar varias veces antes de la falsificación. Todos los individuos firmaron con un bolígrafo de tinta en una plantilla de papel sobre una tableta digitalizadora. Por lo tanto, se dispone también de las firmas impresas en papel (firmas off-line). Para los experimentos,

Fig. 3. Ejemplos de firmas de los cuatro tipos encontrados en la base de datos MCYT.

se digitalizaron las muestras de 75 usuarios (con sus respectivas falsificaciones entrenadas) con un escáner a 600 dpi, obteniendo un conjunto de 2250 imágenes de firmas, 15 imágenes de firmas genuinas y 15 imágenes de falsificaciones por usuario (ver Figura 3)[1].

El conjunto de entrenamiento comprende $K = 5$ ó $K = 10$ firmas genuinas (dependiendo del experimento que queramos realizar). El resto de las firmas genuinas se utilizan para evaluación. Para un usuario específico, calculamos los resultados de impostor casual usando las muestras genuinas disponibles de los usuarios restantes. Los resultados de las pruebas con impostores entrenados se calculan utilizando las falsificaciones entrenadas de cada usuario. Como resultado, obtenemos $75 \times 10 = 750$ ó $75 \times 5 = 375$ *scores* de firmas genuinas, $75 \times 15 = 1,125$ *scores* de impostor de falsificaciones entrenadas, y $75 \times 74 \times 10 = 55,500$ ó $75 \times 74 \times 5 = 27,750$ *scores* de impostor de falsificaciones casuales.

En un contexto de verificación, son posibles dos situaciones de error: un impostor es aceptado (Falsa Aceptación, FA) o un usuario correcto es rechazado (Falso Rechazo, FR). Para calcular estos errores, usamos la curva DET (*Detection Error Trade-off*), que representa FA vs. FR. Para obtener una indicación del rendimiento con alineación ideal entre los *scores* de los usuarios, también calculamos el EER (*Equal Error Rate*) usando normalización dependiente de usuario *a posteriori* [9]. La función de normalización es: $s' = s - s_\lambda$, donde s es el *score* calculado mediante la comparación de firmas, s' es el *score* normalizado y s_λ es el umbral de decisión del usuario en el punto de EER obtenido a partir del conjunto de *scores* genuinos y de impostor del usuario λ.

3.2 Resultados

Los resultados del sistema usando normalización dependiente de usuario *a posteriori* se encuentran en la Tabla 2 para características individuales y en la Tabla 3 para combinación de características. Las curvas DET para las características individuales, sin normalización, están pintadas en la Figura 4.

[1] Este conjunto de imágenes está disponible en http://atvs.ii.uam.es

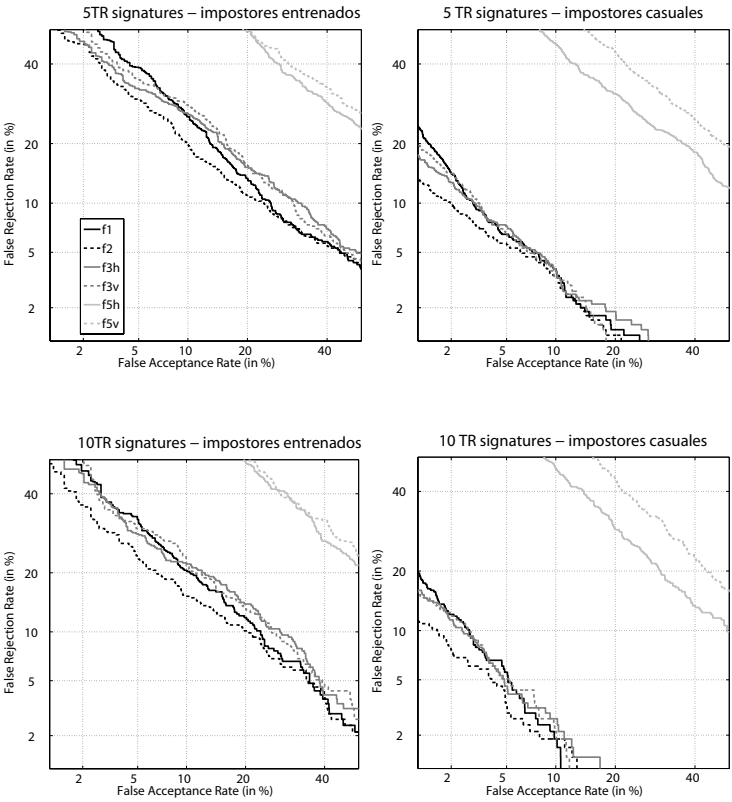


Fig. 4. Rendimiento en verificación sin normalización de scores (umbrales independientes de usuario) para 5 firmas de entrenamiento (5 TR) y 10 firmas (10 TR).

Podemos observar que la mejor característica individual es siempre la "Contour Hinge" PDF f2, independientemente del número de firmas usadas para entrenamiento, y tanto para falsificaciones entrenadas como causales. Esta característica codifica simultáneamente curvatura y orientación del contorno de las firmas. Esto es un dato a tener en cuenta, ya que las otras características que usan también dos ángulos (f3h, f3v) actúan peor que f2. También podemos apuntar que la característica que sólo utiliza un ángulo (f1) muestra un rendimiento comparable con f3h y f3v, incluso superandolos en algunas regiones de las curvas DET. Es importante resaltar el mal resultado obtenido por las PDFs de longitud (f5h y f5v). Esto sugiere que la longitud de las regiones encerradas en las palabras o en los trazos no es una característica distintiva en la verificación de firma off-line, al menos con el preprocesado usado por nuestro sistema.

Un resultado importante es que la combinación de características no produce una mejora, como vemos en la Tabla 3, incluso para combinaciones calculadas a partir de características de diferentes categorías (dirección y longitud). Sólo la combinación de f5h y f5v tiene una mejora significativa. Esto es porque las carac-

terísticas basadas en dirección usan el mismo conjunto de valores de los ángulos, aunque los emparejen de modo distinto. En la Figura 2, podemos observar tres ejemplos que usan el mismo valor para uno de los ángulos. Como resultado, hay correlación entre las características y por tanto su combinación no produce mejoras. Para las características basadas en longitud, su mal rendimiento podría explicar por sí mismo porque no proporcionan beneficios en la fusión.

	IMPOSTORES ENTRENADOS						IMPOSTORES CASUALES					
	PDFs dirección				PDFs longitud		PDFs dirección				PDFs longitud	
	f1	**f2**	**f3h**	**f3v**	**f5h**	**f5v**	**f1**	**f2**	**f3h**	**f3v**	**f5h**	**f5v**
5 TR	12.71	10.18	11.40	12.31	30.33	31.78	3.31	2.18	3.09	3.21	22.18	28.03
10 TR	10.00	6.44	7.78	9.16	28.89	33.78	1.96	1.18	1.40	1.49	20.46	28.58

Table 2. Rendimiento del sistema en términos de EER (en %) para las **características individuales** con normalización de scores dependiente de usuario *a posteriori* para 5 firmas de entrenamiento (5 TR) y 10 firmas (10 TR).

IMPOSTORES ENTRENADOS

	f3=f3h+f3v	**f5=f5h+f5v**	**f1 & f5**	**f2 & f5**	**f3 & f5**	**f1 & f2**	**f1 & f3**	**f2 & f3**
5 TR	12.40	**27.56**	16.69	15.56	13.33	13.11	12.38	11.40
10 TR	8.93	**25.60**	13.64	12.13	9.64	9.87	9.16	8.40

IMPOSTORES CASUALES

	f3=f3h+f3v	**f5=f5h+f5v**	**f1 & f5**	**f2 & f5**	**f3 & f5**	**f1 & f2**	**f1 & f3**	**f2 & f3**
5 TR	3.08	**21.00**	6.40	5.86	4.13	2.87	**2.95**	2.45
10 TR	1.63	**17.86**	4.27	3.73	2.23	1.87	**1.43**	**1.06**

Table 3. Rendimiento del sistema en términos de EER (en %) para las **combinaciones de características** con normalización de scores dependiente de usuario *a posteriori* para 5 firmas de entrenamiento (5 TR) y 10 firmas (10 TR). Se marca en negrita los casos donde se obtiene mejora en el rendimiento respecto a la mejor característica individual implicada.

4 Conclusiones

En este trabajo, presentamos un sistema de verificación de firma *off-line* que utiliza características del contorno de las firmas. La individualidad de cada escritor se codifica usando funciones de densidad de probabilidad (PDFs), agrupadas

en dos categorías: PDFs de dirección y PDFs de longitud. Dichas características trabajan a nivel local, calculando diversas propiedades direccionales de fragmentos del contorno, así como la longitud de las regiones contenidas en las letras y los trazos.

Los experimentos se llevan a cabo usando 2250 imágenes de firma diferentes de 75 individuos, extrados de la base de datos MCYT. El rendimiento en verificación se calcula usando umbrales dependientes e independientes de usuario. Las características basadas en dirección funcionan mucho mejor que las basadas en longitud, siendo el mejor EER de 6.44% y 1.18% para imitadores entrenados y casuales, respectivamente ("contour-hinge" PDF f2, 10 firmas de entrenamiento, normalización *a posteriori*). Destacar también que la combinación de características no produce mejora adicional, posiblemente debido a la correlación entre ellas. Como método de fusión usamos la regla simple de la suma. El uso de otras reglas más complejas [12] es una posible vía de mejora en estudio.

Los resultados de verificación son comparables a otros algoritmos propuestos que se basan en otras características pero usan el mismo marco experimental [10]. Esto nos anima a estudiar la combinación de las mismas usando distintas estrategias de fusión [11].

5 Agradecimientos

Este trabajo ha sido financiado por el proyecto TEC2006-13141-C03-03 del Ministerio de Educación y Ciencia y por la Red de Excelencia Europea BioSecure IST-2002-507634. El autor F. A.-F. agradece a la Consejeria de Educacion de la Comunidad de Madrid y al Fondo Social Europeo por financiar sus estudios de Doctorado. El autor J. F. está siendo fianciado por una Marie Curie Fellowship de la Comisión Europea.

References

1. A. Jain, A. Ross and S. Pankanti, "Biometrics: A Tool for Information Security", *IEEE Trans. on Information Forensics and Security*, 1:125–143, 2006.
2. R. Plamondon and S. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(1):63–84, 2000.
3. M. Fairhurst, "Signature verification revisited: promoting practical exploitation of biometric technology", *Electronics and Communication Engineering Journal*, 9:273–280, December 1997.
4. F. Alonso-Fernandez, M. Fairhurst, J. Fierrez and J. Ortega-Garcia, "Impact of signature legibility and signature type in off-line signature verification", *Proceedings of Biometric Symposium, Biometric Consortium Conference*, Baltimore, Maryland (USA), 1:1-6, September 2007.
5. M. Bulacu and L. Schomaker, "Text-Independent Writer Identification and Verification Using Textural and Allographic Features", *IEEE Trans. on Pattern Analysis and Machine Inteligence*, 29(4):701–717, April 2007.

6. N. Otsu, "A threshold selection method for gray-level histograms", *IEEE Trans. on Systems, Man and Cibernetics*, 9:62–66, December 1979.
7. R. Gonzalez and R. Woods, *Digital Image Processing*, Addison-Wesley, 2002.
8. J. Ortega-Garcia, J. Fierrez-Aguilar, D. Simon, J. Gonzalez, M. Faundez-Zanuy, V. Espinosa, A. Satue, I. Hernaez, J. Igarza, C. Vivaracho, D. Escudero and Q. Moro, "MCYT baseline corpus: a bimodal biometric database", *IEE Proceedings on Vision, Image and Signal Processing*, 150(6):395–401, December 2003.
9. J. Fierrez-Aguilar, J. Ortega-Garcia and J. Gonzalez-Rodriguez, "Target Dependent Score Normalization Techniques and Their Application to Signature Verification", *IEEE Trans. on Systems, Man and Cybernetics-Part C*, 35(3), 2005.
10. J. Fierrez-Aguilar, N. Alonso-Hermira, G. Moreno-Marquez and J. Ortega-Garcia, "An off-line signature verification system based on fusion of local and global information", *Proc. Workshop on Biometric Authentication, BIOAW*, Springer LNCS-3087:295–306, 2004.
11. J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez and J. Bigun, "Discriminative multimodal biometric authentication based on quality measures", *Pattern Recognition*, 38(5):777–779, 2005.
12. A. Ross, P. Flynn and A. Jain, editors, *Handbook of Multibiometrics*, Springer, 2006.
13. F. Alonso-Fernandez, M. Fairhurst, J. Fierrez and J. Ortega-Garcia, "Automatic measures for predicting performance in off-line signature", *Proc. International Conference on Image Processing, ICIP*, 1:369-372, San Antonio TX, USA, September 2007.

A comparative study of local feature sets in position, velocity and acceleration domains for on-line signature verification

J.M. Pascual-Gaspar, V. Cardeñoso-Payo and C.E. Vivaracho-Pascual

ECA-SIMM, Dpto. Informática, Universidad de Valladolid,
Campus Miguel Delibes s/n, 47011 Valladolid, Spain
{jmpascual,valen,cevp}@infor.uva.es

Abstract. Several sets of local features have been proposed for Automatic Signature Verification (ASV) but just a few works address the issue of how to combine isolated local features and the effect of these combinations on the final performance. This work proposes a systematic way to select a combination of local features in three signal domains (position, velocity and acceleration), so that the recognition rate is maximized. The results show drastical improvements in the performance when a proper selection of the feature set is carried out. Best performance was always obatined with feature sets combining the position and velocity domains.

1 Introduction

Feature selection is one of the less documented and systematically characterized issues in the design of ASV systems. Several proposals can be found in the literature which analyze the consistency and effectiveness of isolated signature features [1, 2]. Most of these studies analyze the effectiveness of the individual features, without considering how the system operates when they are combined. Nevertheless, it is well known that the combination of the best individual features not always leads to the best overall feature set [3]. In this work a comparative study of different feature sets of on-line signatures has been performed on three signal domains (position, velocity and acceleration) using a DTW based ASV system.

The rest of this paper is organized as follows: section 2 describes the experimental framework; a detailed description of the process used to select the best feature set is explained in section 3; in section 4 the experimental results will be presented and discussed; finally, section 5 summarizes work's achievements and suggest directions for future improvements.

2 Experimental procedure

A framework based on the experimental methodology designed for the first international signature verification competition (SVC2004 [4]) has been applied to carry out the experiments presented in this work. A description of this experimental framework will be

presented next, aligned with the four different modules which make up any biometric system: 1) sensor and system database, 2) feature extraction, 3) template matching and 4) decision-making.

2.1 Sensor module and system database

There are two possibilities to digitally record a handwritten signature: the *off-line* mode, where the digitalization is done after signature has been captured and stored, without taking into account its generation process; and the *on-line* mode, in which digital data collection is carried out as the signature is captured, including temporal information and thus preserving signature dynamics. For the rest of this paper, we focus on the on-line acquisition mode.

Signatures can be obtained through a variety of devices such as digital pens, personal digital assistants (PDAs) or Tablet-PCs. Pen tablets are perhaps the acquisition devices that allow higher spatial and temporal resolution at an affordable price. Because these kind of devices are still widespread and provide a virtual reference for future alternatives, all the signature databases which have been used in this work were acquired using pen tablets.

Additionally, the four databases we have used to test and compare our selection procedures have been widely used in many state-of-the art ASV systems: MCYT-100 [5], SVC2004 [4], BIOMET [6] and MyIDea [7]. Table 2.1 summarizes the main figures of these databases, as used in the present work. Since there were no forgeries available for some of the users in these databases, those users were not included in our study. Thus, the interested reader might find a difference in the values of figures in table 2.1 and the original references, where she will find further information. An additional point is that, to our knowledge, the study presented here is the first to face the evaluation of a single ASV system over several publicly available on-line signature databases.

Database	Number of users	Genuine signatures	Forgery signatures	Total signatures
MCYT-100	100	25	25	5000
SVC2004	40	20	20	1600
BIOMET	84	15	17	2201
MyIDea	69	18	36	3465

Table 1. Comparative description of the databases.

Although the final system was ultimately evaluated against the four full-size databases, the selection of the best feature sets was carried out using just the first half of MCYT-100 database (MCYT-A). The remaining 50 users (MCYT-B), beside the other three complete databases, were used to validate the system using the best feature sets.

2.2 Features extraction and preprocessing

Hand movements are recorded when signing on a pen tablet. Position and signing gesture information are electronically transduced from the tablet and a special pen to a sequence of temporal samples, at a typical fixed rate of 100Hz. The pen tablet raw features can be classified into two types: a) *positional* (x, y), a 2D point in the path followed by the pen along the signing process; b) *ergonomic* (p, a, i), resulting from the interaction of the hand with the pen. They include the pressure p exerted on the tablet and two angles for the orientation of the pen against the tablet surface (azimuth a and inclination i).

The basic features set $F = (x, y, p, a, i)$ was expanded by including its first and second derivatives $\Delta F = (dx, dy, dp, da, di)$, $\Delta\Delta F = (ddx, ddy, ddp, dda, ddi)$ so the final feature vector consisted of up to 15 components, 5 for each signal domain (position, velocity and acceleration).

Two types of normalization of the features were carried out. First, a geometrical translation was applied to positional features, so that the origin of the coordinate system lays at the signature geometric centre. A z-norm statistical scaling was then applied to get a cero mean and unit variance of each feature: x, y, p, a, i, their first and second derivatives. For every signature, all the vectors having a null pressure component were automatically removed from the temporal sequence of vectors, in all the experiments carried out in this work.

2.3 Template matching

Two common alternatives to determine the similarity between time series associated to different signatures are the *reference-based* and *model-based* approaches [8]. Here, the similarity was calculated using Dynamic Time Warping (DTW) to get the optimal nonlinear alignment of two temporal sequences and a euclidean distance to measure the dissimilarity between two instantaneous feature vectors. DTW is a simple but effective technique widely used in on-line signature verification [9, 10] which exploits dynamic programming techniques to provide a convenient distance measure between temporal series of different duration.

To cope with the inherent intra-user variability of the signing process, a given user should be represented by a number of genuine signature exemplars. Since there is no agreement yet on the optimal number of reference signatures per user, we used five as recommended by Fierrez [8]. This number is low enough as to provide useful results for practical scenarios.

To score the similarity between a test signature and the five reference signatures of a given user, an arithmetic mean was used to combine the DTW aligned distances between the test and each reference. Ten trials were conducted for each experiment and all the scores were used to get the EER, so that the effects of the election of the five reference signatures could be minimized through the random selection of them [6].

2.4 Decision-making and error evaluation

To evaluate the verification error in an ASV application, a decision threshold has to be fixed which compares with output scores linked to each user of the system. In real

applications, this threshold is set before evaluating the system (*a-priori threshold*). The error is then given by the pair formed by the False Rejections Rate and False Acceptance Rate (FRR, FAR). However, to compare different systems is usual to compute the Equal Error Rate (EER). This requires using an *a-posteriori threshold*, which ensure FRR equals FAR. On the other hand, decision threshold can be shared for all users (*universal threshold*) or be specific for each one (*individual threshold*). User-dependent thresholds provide better recognition rates than user independent thresholds [11], although they are not so common in other biometric fields.

When testing ASV systems, two categories of forgeries are usually considered: random and skilled. They are normally presented as two reference separate tests in order to show the performance of the system under different security restrictions. In this paper, we have used user-dependent thresholds and EERs have been obtained both for random and skilled forgeries, since all the databases included both kinds of forgeries, although skilled forgeries were not really professional forgeries but were carried out by other users of the database, under limited and controlled training conditions.

3 Selection of the optimal feature set

Feature selection is a key problem in many pattern recognition problems and there are several alternatives exist to solve it. In our case, we have only 15 initial standard features and the total number of possible combinations goes up to ($2^{15} - 1 = 32767$). Such a high number of combinations encourages discarding any exhaustive searching procedure. Instead, we propose a intertwining of the classical Forward Selection (FS) and Backward Elimination (BE) heuristics [12]. In forward selection, most promising features are progressively incorporated into larger subsets, whereas in backward elimination one starts with the set of all variables and progressively eliminates the least promising ones. We propose an iterative procedure combining these two methods. Using an *a posteriori* quality criterion based on ASV system performance, the best feature is selected and the worst one is discarded at each step of an iterative procedure which starts with an empty feature set and finishes with the best proposal:

1. [Init] Set N to the maximum number of initial features (9).
2. [*FS step*] From a set of N features, select the one with lowest EER.
3. [*BE step*] From the set of N-1 features, discard the one which is not present in the combination with lowest EER.
4. [Iterate] N is set to N-2 and steps 2 and 3 are repeated until there are no more features available.

Although the total number of features to be tested was of 15, preliminary experiments showed that, as predicted by some other authors [2], the angle parameters (a, i) produced extremely low performance ASV systems in all domains (value, delta, delta-delta) when compared to the other three (x, y, p). After discarding this angles, the initial search space cardinal was reduced to 511 sets, which was finally downsized to just 44 sets after applying the FS-BE selection procedure.

4 Results

Table 4 shows the optimal feature sets obtained after the selection process was applied to database MCYT-A, using random and skilled forgeries.

	FS step		**BE step**		**FS step**		**BE step**	
# Iteration	**+feature**	**%EER**	**−feature**	**%EER**	**+feature**	**%EER**	**−feature**	**%EER**
0	–	–	All	0.70	–	–	All	1.89
1	+y	0.73	-ddp	0.53	+dy	3.46	-ddp	1.58
2	+dx	0.34	-ddy	0.46	+dx	1.56	-ddx	1.45
3	+x	0.22	-ddx	0.32	+ddy	1.57	-x	**1.40**
4	+dy	**0.20**	-dp	0.26	+dp	1.64	-p	1.61
	Random forgeries				**Skilled forgeries**			

Table 2. EER(%) of the bests features sets at each step of the selection process.

The results show an overall excellent performance of the DTW verification system, already reported in previous works [13]. Even when there is a variation of EER depending on the feature set, the typical values are clearly below the ones obtained when using standard features (see table 4 below).

It is also noticeable that the optimal feature set depends on the kind of forgery. The combination of geometric coordinates in the position-velocity domain $F_{rd}^{opt} = (x, y, dx, dy)$ provides the best performance (EER = 0.20%) for random forgeries. When skilled forger attacks are an issue, the feature set $F_{sk}^{opt} = (y, p, dx, dy, dp, ddy)$ provides the lowest EER against this type of scenario (1.40%). Random forgeries show a clear inter-user discrepancy of the x, y (position) and dx, dy (velocity) values, on the average. For skilled forgeries, the addition of pressure in the position and velocity domains plus the y-acceleration provides extra information which allows better separation genuine from forged samples. Intermediate alternatives between these optimal features sets could be used for mixed forgeries scenarios.

The previous results were obtained using the same feature sets for all users. When a different optimum feature set is selected for each of the 50 available users, the averaged EER falls to 0.07% for random forgeries and 0.59% for skilled forgeries. Although this is a very promising result (one order of magnitude below the errors reported by state of the art ASV systems), this user-dependent selection of the optimal feature set has been carried out *a-posteriori*, so it would still be an open problem how to use *a-priori* information to drive the selection of a user-dependent feature set.

Finally, results using some common standard and optimal features set over the remaining databases are shown in Table 4. It should be pointed out that the same trends as reported for MCYT-A selection database are found over the rest of databases (except BIOMET with skilled forgeries). Also, the results obtained using the optimal set selected along the lines described in our work even improve the ones of the winner of the competition when applied to SVC2004 database ($EER_{rd} = 0.00\%, EER_{sk} = 3.38\%$ vs. $EER_{rd} = 3.02\%, EER_{sk} = 6.90\%$).

EER(%) with random forgeries

Features set	**MCYT-A**	**MCYT-B**	**SVC2004**	**BIOMET**	**MyIDea**
F^{std}	6.65	5.64	0.78	7.58	**2.45**
ΔF^{std}	5.92	**2.87**	0.60	**3.29**	5.07
$\Delta\Delta F^{std}$	22.93	17.70	11.20	23.91	16.85
F^{std} + Δ	**5.47**	3.25	**0.40**	4.47	3.52
F^{std} + Δ + $\Delta\Delta$	7.50	4.82	1.01	6.58	6.14
F^{opt}_{rd}	**0.20**	**0.38**	**0.00**	**0.33**	**0.92**
F^{opt}_{sk}	0.35	0.46	0.32	0.96	2.39

EER(%) with skilled forgeries

Features set	**MCYT-A**	**MCYT-B**	**SVC2004**	**BIOMET**	**MyIDea**
F^{std}	7.14	6.53	4.70	5.41	**2.89**
ΔF^{std}	5.78	**3.31**	5.92	**3.13**	3.69
$\Delta\Delta F^{std}$	10.67	6.63	15.81	8.62	5.65
F^{std} + Δ	**5.65**	4.21	**4.15**	3.69	3.25
F^{std} + Δ + $\Delta\Delta$	6.16	4.23	6.14	4.43	4.10
F^{opt}_{rd}	1.73	1.16	3.70	**1.25**	2.94
F^{opt}_{sk}	**1.40**	**1.06**	**3.38**	1.48	**2.72**

Table 3. Standard vs optimal sets in both kind of tests.

5 Conclusions and outlook

In this paper, it has been proved that a careful selection of the optimal feature set can drastically improve the performance of ASV systems.

In particular, when optimal features sets were used instead of the tablet raw features set ($F^{std} = (x, y, p, a, i)$), the average EER obtained over the four databases can be reduced from 4.62% to 0.37% (a 92% error reduction) with random forgeries and from 5.33% to 2.01% (a 62% error reduction) with skilled forgeries.

Regarding the security specifications of the system, a different optimal feature set could be selected. The design of a method to drive *a-priori* selection of optimal user-dependent feature sets opens way to further research and it could provide even better results.

References

1. Plamondon, R., Parizeau, M.: Signature verification from position, velocity and acceleration signals: a comparative study. In: Proc. of the 9th International Conference on Pattern Recognition. Volume I. (November 14-17 1988) 260–265
2. Lei, H., Govindaraju, V.: A comparative study on the consistency of features in on-line signature verification. Pattern Recognition Letters **26**(15) (November 2005) 2483–2489
3. Cover, T.: The best two independent measurements are not the two best. IEEE Trans. Systems, Man, and Cybernetics **4** (January 1974) 116–117
4. Yeung, D., Chang, H., Xiong, Y., George, S., Kashi, R., Matsumoto, T., Rigoll, G.: SVC2004: First international signature verification competition. In: Proc. of the First International Conference on Biometrics Authentication (ICBA 2004). (2004) 16–22
5. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon-Zorita, D., Gonzalez-Rodriguez, J., Hernaez, I., Igarza, J.J., Vivaracho, C., Escudero, D., Moro, Q.I.: MCYT baseline corpus: a bimodal biometric database. IEE Proc. Visual Image Signal Processing **150**(6) (2003) 395–401
6. Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., les Jardins, J.L., Lunter, J., Ni, Y., Petrovska-Delacretaz, D.: BIOMET: A multimodal person authentication database including face, voice, fingerprint, hand and signature modalities. In Kittler, J., Nixon, M., eds.: AVBPA. LNCS 2688, Springer-Verlag Berlin Heidelberg (2003) 845–853
7. Dumas, B., Pugin, C., Hennebert, J., Petrovska-Delacretaz, D., Humm, A., Evequoz, F., Ingold, R., von Rotz, D.: MyIDea - Multimodal biometrics database, description of acquisition protocols. In: Proceedings of Third COST 275 Workshop (COST 275), Hatfield (UK) (October 27 - 28 2005) 59–62
8. Fierrez, J., Ortega-Garcia, J.: 12: On-line signature verification. In: Handbook of Biometrics. Springer (2008) 189–209
9. Martens, R., Claesen, L.: On-line signature verification by dynamic time-warping. In: Proceedings of the 13th International Conference on Pattern Recognition. Volume 3. (1996) 38–42
10. Jain, A.K., Griess, F.D., Connell, S.D.: On-line signature verification. Pattern Recognition **35**(12) (December 2002) 2963–2972
11. Fierrez-Aguilar, J., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Target dependent score normalization techniques and their application to signature verification. IEEE Transactions on Systems, Man and Cybernetics, Part C **35**(3) (August 2005) 418–425
12. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(1) (2000) 4–37
13. Kholmatov, A., Yanikoglu, B.: Identity authentication using improved online signature verification method. Pattern Recognition Letters **26**(15) (November 2005) 2400–2408

Estudio de la Aceptación y la Respuesta del Usuario ante la Biometría y sus Diferentes Modalidades

Aitor Mendaza Ormaza, Belén Fernández Saavedra, Raúl Alonso Moreno, Iván Rubio Polo

Universidad Carlos III de Madrid – Grupo Universitario de Tecnologías de Identificación (GUTI) – Dpto. Tecnología Electrónica.
Av. Universidad 30 28911 – Leganés (Madrid)
{amendaza, mbfernan, ramoreno, irubio}@ing.uc3m.es

Resumen. Actualmente la implantación de la identificación biométrica en la vida cotidiana se encuentra en pleno desarrollo. Aun así y pese a una mayor difusión de la misma y su empleo en aplicaciones que requieren seguridad, hoy en día existe un rechazo por parte del usuario debido a ideas preconcebidas y a falsos mitos. Parte del origen de dichas ideas erróneas se debe a la visión que se le da a la biometría en las películas y series de ciencia ficción. Este rechazo, a su vez, se manifiesta también en forma de miedo ya sea por cuestiones de seguridad física, aspectos legales o por temor a perder el anonimato. En el presente artículo, los autores realizan un estudio sobre la aceptación y la respuesta del usuario ante esta tecnología analizando el comportamiento y la reacción de las personas frente a algunas de las principales modalidades biométricas y sus dispositivos. Dicho estudio se ha llevado a cabo durante una labor de divulgación acerca de la biometría y de la base científico-tecnológica sobre la que se encuentra desarrollada en la IX Feria Madrid es Ciencia.

Palabras Clave: Biometría, Aceptación, Respuesta del usuario.

1 Introducción

La biometría, y sobre todo la identificación biométrica informática, es una tecnología que se encuentra hoy en día en pleno desarrollo, tanto en el ámbito de la investigación como en el de la implantación en la vida cotidiana. Esta rápida evolución se debe a la creciente preocupación actual por el tema de la seguridad y a la vinculación que tiene esta técnica para garantizar la misma.

Hoy en día es muy común la inclusión de sensores biométricos en muchos de los aparatos electrónicos que se manejan habitualmente como por ejemplo, el empleo de sensores de barrido de huella dactilar en los ordenadores portátiles como alternativa y/o complemento a la utilización de la contraseña. Esta aplicación en concreto no ha recibido un rechazo significativo por parte del usuario dado que el uso de la modalidad biométrica de reconocimiento mediante huella dactilar está ampliamente aceptado, debido a la eficacia demostrada y a su madurez.

Sin embargo, fuera de la huella dactilar existe un amplio desconocimiento por parte del público en general de las distintas modalidades biométricas, así como una serie de falsos mitos divulgados por las distintas películas y series de ciencia ficción. Todo esto hace que se creen en los usuarios diferentes temores que les llevan a oponerse a la utilización de estás técnicas como métodos para proporcionar seguridad y que pueden llevar al fracaso comercial de este tipo de sistemas.

Es por ello que autores como A. Mansfield, J. L. Wayman en [1] o estándares internacionales como la norma ISO/IEC 19795-1 en [2] expresan la importancia de analizar dentro de las evaluaciones de los sistemas biométricos, a parte de los ya conocidos parámetros de rendimiento como son las tasas de falsa aceptación y falso rechazo, otro tipo de parámetros entre los que se encuentra el factor humano. Dentro de esos parámetros se incluyen el grado de aceptación del sistema, así como el comportamiento del usuario durante la interacción con el mismo.

Por tanto, se hace necesario evaluar y conocer la respuesta de los usuarios en relación con las diferentes modalidades biométricas, frente a los diversos dispositivos que se emplean en dichas modalidades, y de cara al sistema completo y a la aplicación concreta en la que se esté utilizando.

A pesar de esta necesidad, existen pocos trabajos previos relacionados con el tema. Un estudio global de la usabilidad fue realizado por D. T. Toledano, R. F, Pozo, A. H. Trapote y L. H. Gómez [3]. En él se analizaron tres técnicas biométricas (huella dactilar, voz y firma manuscrita) dentro del contexto de efectuar una verificación por Internet. En su estudio utilizaron a usuarios con edades comprendidas entre los 22 y los 24 años y procedentes de escuelas técnicas. Durante su estudio emplearon dos sesiones, una primera de 45 minutos y otra de 20 minutos realizando cuestionarios a dichos usuarios al final de cada una de ellas, en los que se recogía la opinión del usuario.

En el presente artículo se presenta un estudio genérico sobre la aceptación y la respuesta del usuario frente a la biometría, teniendo en cuenta un público general, donde usuarios de todas las edades y diversos grados de conocimiento tuvieron la oportunidad de usar e interactuar directamente con sistemas biométricos de identificación y verificación. Este estudio ha sido realizado durante el desarrollo de la IX Feria de Madrid es Ciencia, en la que se mostraron varios ejemplos de sistemas biométricos de tres de las modalidades existentes en el mercado: huella dactilar, iris ocular y vascular, y algunos de sus dispositivos de adquisición. Durante este evento de divulgación científica, los usuarios conocieron y expresaron sus opiniones, dudas y temores frente a esta nueva tecnología y su utilización como medida de seguridad. Como se pretendía obtener una opinión general y del mayor número de personas posible, no se mostró ninguna aplicación específica de la biometría ni se buscó que los usuarios tuvieran que realizar cuestionarios, para no provocar a priori un rechazo a utilizar los sistemas, además de agilizar la interacción con los mismos.

Este artículo trata de analizar y dar a conocer todas estas impresiones y obtener conclusiones respecto a los diferentes dispositivos de captura, sus respectivas modalidades y a la biometría en general. Para ello, en el próximo apartado se describirán brevemente las diferentes modalidades expuestas en la Feria, las razones de elegir éstas, así como sus respectivos sensores de captura. Posteriormente se detallará el escenario en el que se procedió a la divulgación de las distintas modalidades biométricas explicando su entorno, la interfaz gráfica diseñada para la

interacción con el usuario y el conjunto de usuarios objeto de esta evaluación. Por último se comentarán los resultados obtenidos a partir del estudio realizado y las conclusiones más significativas del mismo.

2 Modalidades biométricas y sus dispositivos

Como paso previo a la evaluación de la respuesta del usuario hubo que elegir las distintas modalidades biométricas mediante las que se pretendía analizar la reacción del público, así como los distintos dispositivos que se iban a emplear para cada una de ellas.

Las principales características que empujaron a tomar estas elecciones fueron que se trataran de modalidades poco o nada intrusivas y que el funcionamiento de los dispositivos para la adquisición de la muestra biométrica fuese sencillo y no supusiera un gran esfuerzo de cara a la utilización por parte de personas no entrenadas o habituadas a su uso. Con estos dos objetivos se pretendía evitar un rechazo prematuro de las diferentes modalidades. A su vez también se tuvo en cuenta los datos existentes en el mercado recogidos por el IBG (*International Biometric Group*) en [4] en relación con el uso y la expansión de las diferentes modalidades.

Finalmente, tal como se ha comentado anteriormente, de las múltiples modalidades biométricas existentes hoy en día, las que se decidió llevar a la Feria para comprobar la reacción del público fueron tres: huella dactilar, iris ocular y la biometría vascular, o más conocida como reconocimiento biométrico mediante el patrón de venas de la mano.

Estas tres técnicas cumplen con los objetivos anteriores ya que son no intrusivas y se disponía de los sensores de captura para cada una de ellas con un funcionamiento simple y un manejo que apenas requiere esfuerzo por parte del usuario. Además, tal y como revela en el estudio de mercado del 2007 publicado por el IBG [4], estas tres modalidades tienen un diferente grado de divulgación, siendo la huella dactilar la más extendida, el iris una de las que más desarrollo ha tenido en los últimos tiempos y la biometría vascular una modalidad más novedosa y apenas conocida. De esta forma, con estas tres modalidades, se podía analizar la influencia del conocimiento previo de la biometría en el grado de aceptación por parte del usuario.

A continuación se procederá a describir brevemente cada una de estas modalidades biométricas junto con algunas de sus aplicaciones actuales. Posteriormente se comentarán los sensores biométricos utilizados para cada una de ellas.

2.1 Modalidades biométricas

- **Huella dactilar**: esta técnica se basa en el reconocimiento de personas mediante los micro-pliegues que posee la epidermis, o capa externa de la piel, en las yemas de los dedos de las manos. Entre todas las características corporales, las huellas dactilares fueron una de las primeras en la historia del ser humano en ser utilizadas para la identificación, tras el estudio de la firma manuscrita y el reconocimiento facial. Desde entonces hasta ahora, sólo la tecnología ha cambiado, siendo ésta una de las

técnicas más utilizadas y conocidas. Esto ha llevado a que posea un soporte legal, lo cuál puede plantear cierto rechazo por parte del usuario debido a su connotación policial y jurídica.

Aplicaciones actuales: la mayoría de gobiernos del mundo usan la huella dactilar como método de reconocimiento, como puede ser el caso de la base de datos del FBI en Estados Unidos [5]. Asimismo, en algunos sectores se han implementado sistemas de identificación, como puede ser el caso del sistema para el control de empleados implementado en [6].

- **Iris Ocular**: el iris es el aro de color que se encuentra en el ojo rodeando a la pupila y que presenta complejos detalles como surcos, hoyos y estrías. Esta técnica de identificación se basa en el análisis de dichos detalles creando un patrón único para cada usuario. Su uso está cada vez más extendido aunque existe una gran confusión entre esta modalidad no invasiva y el escaneo de retina, hecho que hace que sea en muchos casos rechazada.

  Aplicaciones actuales: Debido a la gran precisión de los sistemas de reconocimiento de iris, en 1998 se estrenó en la ciudad inglesa de Swindon el primer cajero automático provisto de una cámara que capta los ojos del usuario para determinar su identidad. A su vez, se han instalado plataformas para el reconocimiento de iris en áreas de transporte como por ejemplo en los aeropuertos de Schiphol (Ámsterdam), Heathrow (Londres), JFK (Nueva York), Frankfurt (Alemania), Vancouver (Canadá) y Atenas (Grecia), donde hay que destacar que la lista de instalaciones que disponen de estos sistemas crece rápidamente.
- **Vascular**: esta reciente modalidad se basa en la identificación de las personas mediante el patrón de venas de la palma de la mano. Su aparición ha tenido lugar gracias a sus ventajas y características únicas, debido a que este patrón se encuentra situado en el interior del cuerpo humano y que para su adquisición es necesario que haya sangre circulando por las venas. Esto hace que no puedan quedar rastros del patrón en ninguna superficie, cosa que no ocurre con la huella dactilar, la cual sí queda marcada en los objetos que se tocan. Por otro lado, también se consigue que se mitiguen los miedos por parte del usuario ante la posible amputación de las extremidades para una falsificación.

  Aplicaciones actuales: Esta modalidad de identificación biométrica está siendo usada actualmente en distintos bancos japoneses (el Banco de Tokyo-Mitsubishi y el Banco Suruga) para la autenticación de clientes en cajeros automáticos [7]. Además, la Universidad de Tokio Hospital está desplegando esta tecnología para restringir el acceso a distintos departamentos [8].

2.2 Dispositivos empleados

A continuación se procede a describir brevemente los dispositivos que se han seleccionado para cada modalidad, así como las razones de su elección.

- **Oki IRISPASS®-M** [9] para la identificación mediante iris. Ésta es una cámara que presenta altas prestaciones y de uso sencillo, motivos por los cuáles se decidió emplear este dispositivo. Posee autoenfoque, por lo que basta con situarse frente a la cámara a una distancia de medio metro aproximadamente y mirar hacia un led para que ésta tome una foto de uno o los dos ojos, según esté configurada. Además dispone de guías tanto sonoras como visuales que le indican al usuario como colocarse y hacia donde mirar. Esta cámara emplea la tecnología desarrollada por *Iridian Technologies* [10].
- **Panasonic BM-ET100US Authenticam** [11] también para el reconocimiento basado en el iris. Esta cámara de Panasonic es una cámara multifuncional de canal dual (dispone de dos objetivos, uno para tomar la foto del iris, y otro para ser usado como webcam). Esta cámara presenta limitaciones respecto a la anterior, ya que su calidad es menor y no dispone de autoenfoque, siendo más incómoda para el usuario. Éste debe de ser capaz de localizar un led mirando al objetivo y situarse a la distancia exacta hasta que el color del led cambie, momento en el cual se toma la foto al encontrarse el ojo enfocado. La razón de incluir otro dispositivo de iris es mostrar al público la versatilidad de esta tecnología, así como la diferencia existente entre una cámara de altas prestaciones como es la cámara de Oki, y una cámara más limitada en funciones pero que por su coste y tamaño puede ser más atractiva para los usuarios.
- **Biometrika FX3000** [12] para el reconocimiento mediante huella dactilar. El sensor FX3000 de Biometrika es un pequeño sensor de tipo óptico con alta resolución (569 dpi) capaz de capturar la imagen de una huella dactilar, procesarla y verificarla. Se decidió usar este dispositivo por la capacidad de realizar verificaciones *Match-On-Board* (MOB) en el propio dispositivo. Así mismo, la interfaz de captura es muy simple, con una superficie de adquisición bastante amplia (25x17.8 mm^2).
- **Fujitsu PalmSecure** [13] para la identificación mediante la geometría de las venas de la palma de la mano. La elección de este sensor vino motivada por el hecho de que al ser una modalidad más novedosa, no existe mucha variedad en el mercado y este modelo es del que se disponía en el laboratorio, siendo ya conocido por los autores de este artículo.

Por último, además de tener en cuenta diferentes modalidades y dispositivos biométricos, también se decidió el modo de funcionamiento de las mismas: verificación e identificación. De esta manera se podría valorar si esto representa algún tipo de impedimento respecto a la aceptación del usuario. Así, las modalidades de iris y huella dactilar realizaban una verificación, mientras que la modalidad de vascular realizaba una identificación.

3 Descripción del escenario de las pruebas

Como ya se ha comentado previamente, el objetivo último del presente estudio es evaluar la respuesta del usuario ante distintas modalidades biométricas. Pero existen muchos factores en el contexto en el que se desarrolla la biometría que pueden influir en el funcionamiento de un sistema biométrico y, por lo tanto, en la percepción del usuario, como son: el entorno de operación de los sistemas, el tipo de usuarios (edad, profesión, etc) o la interfaz gráfica con la que interactúan dichos usuarios. A continuación se comentarán cada uno de ellos, remarcando las características que se han considerado más relevantes de cara a la evaluación que se ha llevado a cabo.

3.1 El entorno de operación

El entorno en el que se llevó a cabo la labor divulgativa fue, como ya se ha mencionado anteriormente, la IX Feria Madrid es Ciencia [14]. Esta feria contaba con un stand entero dedicado a la Universidad Carlos III de Madrid, y en dicho stand se montaron 6 puestos pertenecientes a distintos grupos y departamentos de la Universidad, entre los cuales se encontraba el dedicado a la biometría. En este puesto se montó todo lo necesario para sustentar una aplicación junto con los respectivos sensores, con la intención de divulgar las distintas modalidades y dispositivos biométricos comentados en el apartado anterior, con el fin de evaluar el comportamiento de los asistentes frente a la biometría.

De cara al público, se montó un monitor en el que se mostraba la interfaz con el usuario, además de disponer de un ratón para que él mismo pudiera llevar a cabo la interacción con la aplicación. Asimismo, se colocó la cámara de iris BM-ET100US de Panasonic y los sensores FX3000 de Biometrika y PalmSecure de Fujitsu alrededor del monitor. La cámara IRISPASS®-M de Oki se situó en un soporte a una altura de 1,70 cm al lado del puesto, próximo al monitor. De esta forma, un usuario con situarse en frente del monitor tendría acceso al ratón, para interaccionar con la aplicación, y al resto de los sensores, sin necesidad de desplazarse. Con esta colocación se pretendía facilitar en la medida de lo posible el uso de los distintos dispositivos biométricos, para reducir al mínimo el posible rechazo por parte del usuario por considerar muy complicada la interacción con el sistema.

Por otro lado, el entorno del pabellón era un entorno no controlado. Sobre el stand coincidían dos focos de luz del pabellón, a gran altura. Dado que todos los sensores que se llevaron a la feria eran ópticos, dicha luz (no halógena) no era la más adecuada para el funcionamiento de los sensores, pudiendo ejercer una influencia negativa en la respuesta de los dispositivos durante la exposición. Unido a este efecto, hay que tener en cuenta la situación del puesto, que se encontraba de un pabellón por donde transitaron numerosas personas a lo largo de la duración de la Feria (4 días) y en muchos casos se producía una aglomeración de gente.

3.2 La interfaz gráfica de usuario

En el diseño de la interfaz gráfica de la aplicación también se buscó la simplicidad de cara al usuario, haciéndola tan sencilla e intuitiva como fuese posible. Al igual que en la situación de los sensores, se buscaba no poner más trabas de las necesarias para el

uso de los dispositivos biométricos, con el objetivo de no causar un rechazo por parte del usuario antes incluso de llegar a interactuar con dicha interfaz.

Esta interfaz aunaba las tres modalidades biométricas con los dispositivos elegidos para cada una de ellas de forma auto-explicativa, de tal forma que el usuario era guiado con mensajes sencillos para la utilización de las diversas modalidades y sus correspondientes sensores.

3.3 El conjunto de usuarios

Al igual que se tenía un entorno no controlado, los usuarios que probarían las aplicaciones con los distintos dispositivos biométricos serían de una gran variedad. Al ser un encuentro multitudinario y abierto al público en general, las personas que iban a transitar por el puesto se esperaba que fuesen de todo tipo (edad, actitud, profesión, etc).

Atendiendo a la edad, los asistentes tenían desde los 10 años, ya que se fomentó la asistencia a la feria de numerosos colegios e institutos, hasta personas de edad avanzada, rondando los 60-70 años. En el ámbito de la tecnología, el rango de los asistentes era también bastante amplio. Había gente sin ningún conocimiento previo de biometría ni de su uso, así como gente con conocimientos a nivel básico que se mostraron más interesados, tanto en las técnicas que se explicaban como en el grado de implantación que se estaba alcanzando en las aplicaciones de uso cotidiano.

4 Evaluación de la aceptación y respuesta de los usuarios

Tal como se ha descrito en los objetivos de este artículo, se pretende evaluar la respuesta y el comportamiento del usuario ante las modalidades biométricas anteriormente comentadas, a la vez que se realizaba una labor divulgativa de la biometría y de su tecnología con el fin de despejar las falsas creencias y temores que le han sido atribuidos.

A continuación se van a exponer las distintas opiniones recabadas de los usuarios para los distintos dispositivos y para su correspondiente modalidad:

- **Oki IRISPASS®-M.** A nivel de sensor de captura se pudo comprobar como, debido al entorno no controlado y tal como se ha descrito previamente, el rendimiento de la cámara fue un poco menor que el observado en el laboratorio, principalmente por las condiciones de iluminación existentes en el pabellón. Se observó que la cámara fallaba al localizar los iris en los usuarios que usaban gafas. Fue necesario que los usuarios se quitasen las gafas para que la demostración funcionase correctamente. Por esta situación los usuarios con gafas manifestaron la incomodidad que ello supondría si dicha modalidad se implantase en alguna aplicación comercial cotidiana. Se les explicó que dichos fallos provocados por las gafas se debían principalmente al hecho de encontrarse en un entorno con iluminación no controlada, y que en aplicaciones comerciales se realiza un estudio previo a la implementación para ajustar tanto el hardware como el software y eliminar este tipo de fallos.

Por otro lado y en relación con la cámara, se observó que a la hora de situarse frente a la cámara, la mayoría de ellos se acercaban hasta casi tocar la superficie de la cámara con la cara, girando la cabeza de tal forma que quedase de frente a la cámara un único ojo, por lo que se pudo comprobar que a pesar de las indicaciones que ésta proporciona sigue siendo necesario que el usuario sea guiado.

Respecto a la modalidad biométrica de reconocimiento mediante el iris, se comprobó que no es del todo desconocida por el público y que efectivamente existe una gran confusión entre esta modalidad y el escaneo de la retina. Por esta razón, se esperaba cierto recelo a tomarse fotos con la cámara por el temor de que se le produjera algún daño ocular. Sin embargo, se pudo comprobar como este pensamiento era erróneo y que a pesar de la confusión de modalidades, los usuarios no presentaban ninguna objeción a utilizar este sensor.

- **Panasonic BM-ET100US Authenticam.** En relación con el sensor, al igual que con la anterior cámara de iris, el entorno y en concreto el tipo de iluminación, afectó al funcionamiento del dispositivo. Como se ha comentado previamente, lo que se buscaba con este dispositivo era principalmente mostrar las diferencias existentes entre las distintas soluciones comerciales que existen para una misma modalidad biométrica. Esto despertó bastante la curiosidad de los usuarios, ya que aunque su uso era más complejo debido a la falta de autoenfoque, muchos de los usuarios quisieron probar su funcionamiento y conocer las diferencias entre los dos dispositivos de iris. A partir de este hecho se pudo observar que, a pesar de la motivación de los mismos por utilizar esta cámara, la falta de entrenamiento provocaba muchos fallos a la hora de adquirir la foto.

  La presencia de estos dos dispositivos permitió, a su vez, explicarle al usuario la existencia de diferentes aplicaciones con diversos niveles de seguridad y la relación presente entre los requisitos de seguridad y el coste necesario para cumplirlos.

  Respecto a la modalidad biométrica los resultados de aceptación fueron los mismos que los mencionados para la cámara anterior.

- **Biometrika FX3000.** En lo que se refiere al sensor y al igual que le ocurría a los dispositivos anteriores, por las condiciones no controladas de iluminación en el entorno, se obtuvo una respuesta del sensor ligeramente peor a la obtenida en el laboratorio. Al ser un sensor de tipo óptico, la luz influyó sobre su rendimiento, provocando que en ocasiones las imágenes capturadas por el sensor no tuviesen calidad suficiente para extraer las minucias. Sin embargo, esto no influyó negativamente en la percepción del usuario sobre la tecnología, ya que este tipo de modalidad biométrica se encuentra ampliamente aceptada por el público en general.

  A nivel de la tecnología, la aceptación es clara al tratarse de una modalidad más veterana, si bien los usuarios mostraban cierto grado de sorpresa al conocer que la huella dactilar es diferente para cada uno de sus dedos.

- **Fujitsu PalmSecure.** En relación con el sensor y al contrario de los casos anteriores, la respuesta de este sensor fue muy buena a pesar del entorno de luz no controlado. Como se puede comprobar en el estudio [15] este sensor es muy sensible en determinadas condiciones de luz. Por fortuna, las condiciones de iluminación existentes en la feria no afectaron sensiblemente a su rendimiento.

  Se comprobó que la respuesta de los usuarios al usar este sensor fue muy positiva pese a que su funcionamiento es más complejo para los usuarios no entrenados. En algunas ocasiones, fue necesario guiarles durante la fase de reclutamiento instruyéndolos sobre la correcta colocación de la mano sobre el sensor ya que, a pesar de que el sensor cuenta con un soporte para guiar en la forma de situar la mano [16], los usuarios tendían a colocar la mano en una posición incorrecta.

  También se observó que a pesar de que muchos de los usuarios eran niños, cuyas manos son más pequeñas, esto no presentaba ningún obstáculo para el funcionamiento del sensor, con la salvedad del mencionado anteriormente respecto a la colocación de la mano.

  En lo referente a la modalidad los usuarios mostraron bastante interés, al ser la más desconocida entre la gran mayoría del público asistente. Las principales curiosidades vinieron por conocer cómo el sensor obtenía el patrón de venas de la mano y por la diferencia de patrones entre una mano y otra.

De forma general, decir que respecto al modo de funcionamiento de los sistemas, el usuario se mostró indiferente sin mencionar ninguna predilección por el modo de verificación o identificación. Si bien decir que la mayoría de ellos no conocía la diferencia entre ellos la cuál les fue explicada.

Por último, cabe destacar en este punto que la mayoría de los usuarios manifestaron dos grandes dudas o temores sobre las distintas técnicas biométricas y su implantación en el uso diario. El primero de ellos tiene que ver tanto con la fiabilidad de los sistemas biométricos, como con la eficiencia de los dispositivos. La percepción que tiene el público en general, debido a las mencionadas películas de ciencia ficción, es que es relativamente fácil saltarse un sistema de seguridad basado en técnicas biométricas. Es importante reforzar la sensación de seguridad del usuario para que acabe aceptando estos sistemas como buenos y fiables. La segunda duda que planteaban los distintos usuarios va relacionada con la anterior y tiene que ver con la posibilidad de fraude. Asociado a la posibilidad de fraude, como era de esperar, los usuarios manifestaron un miedo ante posibles ataques físicos para obtener las muestras biométricas. Otra vez se puede notar la influencia de la ciencia ficción, ya que muchos usuarios manifestaron la posibilidad de que les amputasen alguna parte del cuerpo con la intención de usar dicha parte para engañar el sistema biométrico.

A raíz de la problemática planteada, se observa la necesidad de efectuar una labor divulgativa sobre la biometría y la seguridad que ésta proporciona, a la vez que se encamina la investigación de los distintos dispositivos e implementaciones con el objetivo de garantizar que los sistemas biométricos comerciales sean inmunes a este tipo de ataques y proporcionen un funcionamiento fiable y en el que el usuario pueda confiar.

5 Conclusiones

Hasta ahora la importancia de la opinión de los usuarios a la hora de evaluar los sistemas biométricos y las distintas soluciones comerciales existentes en el mercado ha sido prácticamente nula. El presente artículo ha intentado poner de manifiesto este factor para lo cuál, se ha realizado una evaluación de la aceptación y el comportamiento del usuario frente a este tipo de sistemas durante el transcurso de un evento de divulgación científica como fue la IX Feria Madrid es Ciencia. Se escogieron tres modalidades biométricas y junto con sus dispositivos, se elaboró una interfaz permitiera al usuario analizar y conocer la biometría de forma sencilla e intuitiva.

A partir de los resultados obtenidos se ha podido estudiar el comportamiento del usuario frente a los dispositivos, a sus respectivas modalidades y a la biometría en general como técnica para proporcionar seguridad. A su vez, se han comprobado los problemas que plantea el desconocimiento de la misma, así como la importancia de las labores de divulgación de la biometría para la correcta aceptación de estás técnicas como soluciones que garantizan la seguridad.

Bibliografía

[1] A. Mansfield, J. L. Wayman "Best Practices in Testing and Reporting Performance of Biometric Devices" v2.01. 2002.
http://www.cesg.gov.uk/site/ast/biometrics/media/BestPractice.pdf

[2] ISO/IEC International Standard 19795 Biometric Performance Testing and Reporting – Part 1: Principles and Framework, 2005.

[3] D. T. Toledano, R. Fernández Pozo, A. Hernández Trapote and L. Hernández Gómez, "Usability Evaluation of Multi-Modal Biometric Verification Systems", in Interacting With Computers, vol. 18, nº 5, September 2006, pp. 1101-1122.

[4] http://www.biometricgroup.com

[5] http://www.turismo.uma.es/turitec/turitec2002/actas/Microsoft%20Word%20-%207.MARAVAL.pdf

[6] http://www.pgjdf.gob.mx/periciales/especialidades/AFIS.htm

[7] http://www.fujitsu.com/global/casestudies/WWW2_casestudy_BTM.html

[8] http://www.fujitsu.com/global/news/pr/archives/month/2005/20050511-01.html

[9] http://www.oki.com/jp/FSC/iris/en/

[10] http://www.iridiantech.com/products.php?page=5

[11] http://www2.panasonic.com/webapp/wcs/stores/servlet/ModelDetail?storeId=11201&catalogId=13051&itemId=63725&catGroupId=16817&surfModel=BM-ET100US&displayTab=O

[12] http://www.biometrika.it/eng/fx3000.html

[13] http://www.fujitsu.com/downloads/GLOBAL/labs/papers/palmvein.pdf

[14] http://www.madrimasd.org/madridesciencia/

[15] R. Sanchez-Reillo, B. Fernandez-Saavedra, J. Liu-Jimenez and C. Sanchez-Avila, "Vascular Biometric Systems and Their Security Evaluation" 41st Annual IEEE International Carnahan Conference on Security Technology (ICCST 2007), Proceedings, pp. 44–51, Otawa, Canada, 8-11 October, 2007.

[16] http://www.frontech.fujitsu.com/en/forjp/palmsecure/sensor/specifications/

Entorno experimental para el prototipado de aplicaciones biométricas multimodales en dispositivos móviles

Álvaro Hernández-Trapote[1], Rubén Fernández[1], Beatriz López-Mencía[1], Álvaro Sigüenza[1], Luís Hernández[1], Javier Caminero[2], Doroteo Torre-Toledano[3]

[1] GAPS, Señales Sistemas y Radiocomunicaciones. Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040 Madrid, Spain
{alvaro, ruben, beatriz, alvaro.siguenza, luis }@gaps.ssr.upm.es
[2] Telefónica I+D, Emilio Vargas 6. 28043 Madrid, Spain . fjcg@tid.es
[3] ATVS - Biometric Recognition Group, Escuela Politecnica Superior, Universidad Autonoma de Madrid, Calle Francisco Tomas y Valiente 11, 28049 Madrid, Spain. doroteo.torre@uam.es

Abstract. En este artículo hemos presentado un entorno que soporta la creación de prototipos de aplicaciones biométricas en escenarios móviles. Para ello se facilitan diversos componentes que por un lado, acceden a los recursos audiovisuales de una PDA, y por otro conectan con un servidor para proceder al tratamiento de la información biométrica. Como ejemplo de aplicación práctico del entorno de experimentación se ha desarrollado una utilidad con la que se ha recopilado una base de datos biométrica multimodal de voz y caras en la que han participado 72 personas.

Keywords: entorno experimental, biometría, dispositivos móviles, bases de datos biométricas.

1 Introducción

En la actualidad existe una marcada tendencia hacia la miniaturización y la creación de entornos y ambientes ubicuos. Esta tendencia hace que, cada vez más, los dispositivos móviles y el desarrollo de servicios para éstos empiecen a atraer mayor atención dentro del panorama industrial. Debido a este interés tan extendido, las garantías de fiabilidad y seguridad de la información intercambiada entre dispositivos portátiles se han convertido en un aspecto clave para los usuarios [1], especialmente en determinados servicios que requieren cierto grado de confidencialidad (operaciones bancarias, acceso seguro a servicios personales…). Así, actualmente en aplicaciones desarrolladas para teléfonos móviles, PDAs o Pocket PCs se está empezando a incorporar tecnologías biométricas para satisfacer esta necesidad de seguridad [2], [3]; además, las características propias de estos dispositivos (cámara/s incorporada/s, micrófonos, etc.) hacen que la inclusión de algunos modos biométricos no requiera de ningún hardware adicional, lo que constituye una ventaja para el desarrollo de estas tecnologías en dispositivos móviles.

Sin embargo, a pesar de la madurez de las tecnologías biométricas, el desarrollo de este mercado ha sido muy escaso en comparación con las expectativas previamente creadas. Existiendo una clara carencia de entornos experimentales que den soporte tanto a la evaluación de las diferentes tecnologías biométricas, como al análisis de usabilidad y aceptabilidad de interfaces que incorporen dichas tecnologías en dispositivos móviles. Así, informes de organismos oficiales como el Nacional Science and Technology Council (NSTC) fijan como retos para el desarrollo de las tecnologías biométricas esta necesidad de creación de nuevos entornos y arquitecturas abiertas para la experimentación, así como el análisis de interfaces de usuario intuitivos seguido de la especificación de guías de diseño para su futura adopción [4]. Actualmente podemos encontrar interesantes desarrollos en este sentido, que hacen uso de iniciativas como el BioAPI [5], interfaz creado para la estandarización de la comunicación con los sensores biométricos, y que permiten implementar de una manera fácil y flexible aplicaciones biométricas en diversos entornos. Por ejemplo, en [6] se describe una arquitectura Java que hace uso del BioAPI para el acceso a los sensores y con la que se pueden diseñar aplicaciones biométricas que se ejecuten sobre ordenadores de escritorio. No obstante aún son necesarios esfuerzos para la creación de plataformas biométricas aplicables a entornos móviles.

En este artículo presentamos un entorno experimental que proporciona recursos para trabajar con las tecnologías de reconocimiento de voz y cara sobre dispositivos móviles. En la Sección 2 describimos este entorno experimental junto a la arquitectura que lo soporta. Destacándose sus posibilidades para permitir un fácil y rápido prototipado de aplicaciones que pueden utilizarse para diversas tareas como la evaluación de diferentes tecnologías biométricas, estudios de usabilidad de interfaces biométricos, o la captura de bases de datos biométricas siempre sobre escenarios móviles, con posibilidad de combinación de modalidades (multimodalidad), y entornos de trabajo realistas, etc. Como ejemplo de aplicación de nuestro entorno en la Sección 3 describimos el desarrollo de una aplicación para la captura de una base de datos de voz y caras sobre una PDA; adicionalmente, se explica cómo la aplicación se utilizó para la recolección de una base de datos de 72 usuarios. Finalmente en la Sección 4 presentamos las conclusiones de nuestro trabajo y cómo planteamos las líneas futuras de investigación.

2 Entorno experimental

El entorno experimental que hemos desarrollado está implementado sobre una arquitectura distribuida cliente-servidor (ver Figura 1). La decisión de adoptar esta configuración surgió con el objetivo de liberar de carga de procesamiento al dispositivo móvil. En este contexto de un entorno distribuido, se desarrollaron dos controles ActiveX que conforman la base principal de la arquitectura en el lado cliente, y que implementan las funciones de obtención de los rasgos biométricos, y de envío al servidor para su tratamiento. Una vez que la información biométrica ha sido capturada en el lado cliente, se envía al servidor donde se procesa según la utilidad de la aplicación que se haya diseñado. En la Figura 1 se puede observar que la comunicación entre cliente y servidor se lleva a cabo a través de una conexión WIFI.

No se han tomado precauciones especiales de encriptación en la transmisión más allá del uso del estándar WPA, y es que a pesar de ser éste una aspecto crítico en los sistemas de seguridad biométrica, su estudio está lejos del objetivo de este entorno que, como ya hemos indicado anteriormente, se orienta al apoyo de la investigación en tecnologías biométricas y a estudios de usabilidad para interfaces seguros sobre terminales móviles.

Antes de describir en más detalle los elementos que forman la arquitectura, creemos interesante presentar cuáles fueron los criterios de diseño que fijamos para la implementación del entorno de experimentación:

- *Experimentación en dispositivos móviles*, como ya se ha comentado anteriormente, los dispositivos móviles forman parte de un sector que reclama mayor seguridad, y la biometría puede ser una opción muy interesante para cubrir esta necesidad. La oferta de sistemas operativos para dispositivos móviles es variada, pero la presencia dominante en éstos del SO Windows Mobile y las facilidades en cuanto a portabilidad de los componentes ActiveX nos llevaron a decidirnos por implementar objetos COM que ejecutaran las funciones en el cliente. Al mismo tiempo, estos dispositivos nos permitirán llevar a cabo experimentos con usuarios bajo diferentes condiciones experimentales.
- *Posibilidad de incluir elementos interactivos,* en la Sección 1 se comentaba la falta de estudios de usabilidad sobre interfaces de usuario para tecnologías biométricas. En este tipo de estudios es realmente interesante incluir elementos interactivos como avatares (p.ej.: software Flipz [7] para generar vídeos de bustos parlantes en formato Flash) o componentes multimedia (p.ej.: objetos Windows Media Player, ver Sección 3). Una de las mejores plataformas para este cometido son los navegadores, y más aún cuando se prevén utilizar módulos ActiveX.
- *Facilitar la creación de interfaces,* teniendo en cuenta que se utilizarán componentes ActiveX embebidos en un navegador, los interfaces se podrán generar utilizando recursos web, como el HTML para la presentación gráfica, y el lenguaje Javascript para la interacción y acceso a los componentes activos.

A continuación presentaremos en más detalle los elementos que componen la arquitectura.

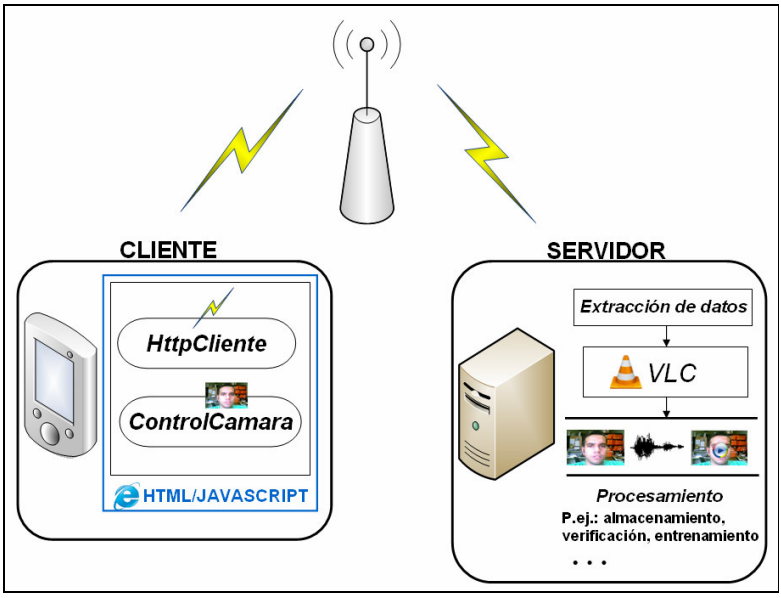


Figura 1. Representación de la arquitectura para el estudio de sistemas de seguridad biométríca sobre dispositivos móviles

2.1 Descripción de la arquitectura en el lado del cliente

Como hemos comentando anteriormente, la base principal del entorno en el lado cliente son dos componentes ActiveX que permitirán acceder a los recursos de la cámara del dispositivo móvil, y enviar las muestras recogidas al servidor. Ambos archivos han sido comprimidos en un archivo *cabinet* (.CAB) para su distribución e instalación. El primero de los componentes ActiveX es '*ClienteCamara'*, que implementa los métodos listados en la Tabla 1. '*ClienteCamara*' está pensado para obtener fotos y vídeos (con audio incorporado) de los usuarios y está basado en OpenCV, un conjunto de librerías de libre distribución para el tratamiento de imágenes [8]. El ActiveX acepta un parámetro (*VideoRenderFile*) mediante el que podemos definir la ruta del archivo temporal en el que se irá guardando la captura de la foto o la grabación del vídeo. Además, para facilitar la gestión de la interacción '*ClienteCamara'* lanza un evento cuando se termina un proceso de captura.

Una vez obtenidos los rasgos biométricos del usuario, éstos pueden ser enviados al servidor para su procesamiento. El encargado de realizar esta operación es el componente ActiveX de nombre '*ClienteHTTP'*. Como se puede intuir de la definición de los métodos de la Tabla 2, su funcionalidad consiste en realizar el envío al servidor, mediante protocolo HTTP y una petición POST, de cadenas de caracteres o archivos (fotos, videos, etc.) previamente capturados en el lado cliente.

Tabla 1. Definición y descripción de los métodos accesibles en el componente ActiveX 'ClienteCamara'

Método	Descripción del método
InitGraph ()	Construye los objetos gráficos internos del componente. Al crear el control ActiveX este método se llama automáticamente, por lo que sólo sería necesario volver a invocarlo en el caso de que detenga el control mediante *ShutdownGraph ()*
SetMessage (message)	Presenta el mensaje *message* en la ventana del objeto ActiveX. Este mensaje no se mostrará por pantalla hasta que no llamemos a *ShowMessage (true)*
TakePicture (file)	Toma una foto y la guarda en el archivo *file*
StartVideoCapture (file)	Inicia la captura de un vídeo e indica en qué fichero, *file*, se almacenará
StopVideoCapture ()	Para la captura del vídeo
ShutDownGraph ()	Detiene los objetos de control de la cámara
Showimage (file)	Muestra la foto del archivo *file* en la pantalla del ActiveX
ShowMessage (bShow)	Habilita la presentación por pantalla del mensaje definido mediante el método *SetMessage (message)*
ShowPreviewWindow (bShow)	Habilita la ventana de previsualización de la cámara
SetPreviewPosition (Left, Top, Width, Height)	Establece el tamaño y la posición de la ventana de previsualización dentro del componente ActiveX

Tabla 2. Definición y descripción de los métodos accesibles en el componente ActiveX 'ClienteHTTP'

Método	Descripción del método
Connect (address , port)	Conecta con el servidor alojado en la dirección *address* por el puerto *port*
Close ()	Cierra la conexión con el servidor
InitializePostArguments ()	Inicializa los argumentos de la llamada POST. Es necesario llamar a este método antes de hacer alguna llamada *AddPostArguments()*
AddPostArguments ()	Añade cadenas de caracteres y archivos a las llamadas POST al servidor

2.2 Descripción de la arquitectura en el lado del servidor

En el lado del servidor se ha utilizado la tecnología Java Servlet. Quizás es necesario recordar que la tecnología Servlet de Java consiste en pequeños programas en lenguaje Java que se activan con una petición desde el cliente. Para manejar e invocar dichos servlets hemos utilizado un contenedor de servlets como Apache Tomcat [9]. En nuestra arquitectura, el servidor alojará un servlet 'genérico' que tiene asignadas básicamente dos tareas:

- En primer lugar nuestro servlet recogerá los mensajes y archivos enviados desde el terminal móvil. Para ello se ha utilizado el paquete *FileUpload* del proyecto

Apache Commons [10]. Este paquete pone a disposición del desarrollador una serie de clases que permiten implementar fácilmente un procedimiento para extraer los objetos que viajen en la petición POST.

- Por otro lado, una vez recogidos las fotos y los vídeos, el servlet podría implementar una serie de métodos para el procesamiento de la información capturada en función del prototipo que se quiera implementar (p.ej.: almacenamiento para la creación de una base de datos, procesos de entrenamiento o verificación para un sistema de autenticación biométrica). Como ejemplo nuestro entorno proporciona un método que haciendo uso del programa de código abierto VLC [11] extrae la información del audio a partir del vídeo capturado en el cliente.

3 Caso práctico: base de datos biométrica multimodal en un entorno móvil

Para demostrar la versatilidad y funcionalidad del entorno experimental implementado describimos en este apartado la adquisición de una base de datos en un entorno real para un sistema de verificación que emplea las tecnologías de reconocimiento de locutor y cara sobre una PDA. Esta base de datos ha sido capturada intentando simular unas condiciones de adquisición lo más realistas posibles, incluyendo:

- Escenarios de captura remotos sin supervisión de un experto.
- Diferentes condiciones de iluminación (artificial y natural) y presencia de ruido en las locuciones de voz, para valorar la robustez de los algoritmos de verificación biométrica frente a dichas fuentes de variabilidad.
- Usuarios sin experiencia previa en el uso de dispositivos biométricos.

3.1 Bases de datos biométricas multimodales en entornos móviles

Uno de los problemas para el desarrollo de las tecnologías biométricas en entornos móviles es la escasa cantidad de recursos y bases de datos capturadas en estas condiciones, importantes tanto para contrastar experimentos, como para utilizarlas en la evaluación de los distintos modos biométricos en función de las condiciones de entorno. Entre estas iniciativas podemos resaltar algunas interesantes como [12], donde intentan simular las condiciones de un entorno móvil capturando la información biométrica mediante una webcam conectada a un portátil. En el caso del proyecto *SecurePhone*, nos encontramos con una base de datos que sí fue capturada mediante una PDA [13] y que contiene 60 usuarios. Sin embargo todavía existe una carencia de este tipo de recursos que se hace aún más evidente si queremos disponer de bases de datos en castellano.

3.2 Recogida de los datos

Nuestra base de datos está formada por 72 participantes (46 hombres y 26 mujeres). Para crear la base de datos, y con el fin de capturar la variabilidad temporal intra-usuario, tal como se propone en [13], hemos dividido el proceso de captura en dos sesiones, dejando un espacio de tiempo entre sesiones de una semana. Además, para cada una de las sesiones anteriores, se capturaban los rasgos biométricos del participante (voz y vídeo) en dos condiciones (o escenarios) diferentes para intentar simular las condiciones reales de funcionamiento del sistema en la PDA.

El *Escenario A* se lleva a cabo bajo condiciones ambientales controladas, esto es, el participante es instado a situarse en lugares bien iluminados donde el foco de luz le viene de frente y de forma homogénea (preferentemente luz exterior natural) y además se solicitaba al usuario que intentara controlar el entorno para protegerlo frente al ruido. En el *Escenario B*, pretendemos que la captura se realice bajo condiciones de iluminación y ruido ambiental realistas, de forma que la calidad de las imágenes y grabaciones va a ser inferior. Ya que en una situación real, el usuario no va a poder siempre acceder al sistema en las condiciones ideales del *Escenario A*, los usuarios también grababan su información biométrica en un entorno donde las fuentes de luz no están controladas, de modo que pueden aparecer sombras en las caras del usuario que dificulten el reconocimiento. Para las grabaciones de voz tampoco se prestará especial atención al ruido de fondo presente en el momento de la grabación.

Para cada escenario (A y B), se capturaron imágenes estáticas de la cara (en vista frontal) y secuencias audiovisuales (vídeos) como explicamos a continuación:

Secuencias audiovisuales:

- Se capturaron 10 vídeos de cada participante.
- Los usuarios eran instados a repetir una secuencia de 5 dígitos por cada vídeo.
- Todos los participantes repetían las mismas 10 secuencias numéricas, de tal forma que, tal como estaban distribuidos los dígitos, asegurábamos que se realizasen al menos 5 repeticiones de cada número. Además, en el diseño de las secuencias se tuvo en cuenta las posiciones de los dígitos dentro de ellas, con el fin de incluir las posibles diferencias de pronunciación y de transiciones que deberían producirse, según sean dichos los dígitos al principio, en una posición intermedia de la grabación o al final de la misma.
- El sistema de captura comprime el vídeo en el perfil simple del codificador de *Windows Media Video 9 Series*, cuyas características son un *bit rate* de 96Kbps y una resolución de 176x144 @ 15Hz (QCIF), degradando la calidad del video y audio capturado, lo que va a afectar a la robustez del proceso de verificación.

Imágenes estáticas:

- 10 capturas.
- 1 imagen facial estática frontal por cada captura.
- La imagen se recoge con formato 24-bit RGB y una resolución de 176x144.

El dispositivo empleado para la captura de la base de datos ha sido una PDA modelo Fujitsu Siemens Pocket LOOX T830.

3.3 Aplicación de captura de la base de datos

Utilizando como punto de partida el entorno distribuido experimental descrito en la Sección 2, se implementó una aplicación Web para la recogida de la base de datos de datos biométricos. Esta herramienta consiste en un interfaz HMTL (Figura 2), que mediante Javascript accede a los métodos y eventos implementados en los controles *ActiveX* detallados anteriormente. En concreto se utilizaron:

- El control *'ClienteCamara'* para acceder a los recursos de la cámara de la PDA y permitir la grabación de los videos y la captura de las imágenes de la cara.
- El control *'ClienteHTTP'*, que permite la comunicación entre la PDA y el servidor. La aplicación proporciona la posibilidad de obtener las fotografías y los vídeos de los usuarios de una forma ordenada y fácil. Así, el interfaz está diseñado de tal manera que el usuario puede elegir la sesión y el escenario en los que está realizando la captura en ese momento.
- Por último, y como ejemplo de componente multimedia para facilitar la interacción del usuario con el interfaz, se incluyó el control ActiveX del reproductor Windows Media Player '*wmp.dll*' para la previsualización de los videos durante la captura de la base de datos. De esta forma, se permite la posibilidad de regrabado si existe algún elemento (ruido, deficientes condiciones de iluminación, presencia de algún agente externo, etc.) que interfiera en el proceso de captura.

Figura 2. Interfaz para la captura de la base de datos

4 Conclusiones

Durante los últimos años se han producido importantes avances en el desarrollo de aplicaciones biométricas para dispositivos móviles; incluso existen algunos productos comerciales que están empezando a estar disponibles. Sin embargo todavía hay algunos aspectos que necesitan ser estudiados para garantizar la viabilidad y el buen funcionamiento de estas aplicaciones en escenarios reales [14]. Para intentar contribuir en esta línea de investigación, hemos desarrollado un entorno experimental que pueda ser un soporte para el desarrollo rápido de prototipos de aplicaciones biométricas sobre entornos móviles. Este entorno pretende ser sencillo y flexible para que pueda ser empleado para plantear estudios o contribuciones en diferentes áreas, como por ejemplo, análisis de usabilidad, pruebas en entornos reales, creación de bases de datos biométricas multimodales, etc.

Como caso práctico del entorno experimental propuesto, hemos elaborado una aplicación de captura de bases de datos biométricas multimodales. Con esta herramienta hemos elaborado una base de datos que consiste en imágenes de caras y vídeos de 72 usuarios en condiciones reales. La información biométrica fue recogida bajo dos escenarios distintos, uno de ellos con un entorno controlado y el otro bajo condiciones más realistas. Esperamos que estos recursos puedan servir para promover el estudio de las tecnologías biométricas en entornos móviles.

Con el entorno experimental que hemos presentado pretendemos implementar un escenario simulado donde un usuario, a través de una PDA, acceda a un área personal utilizando sus rasgos biométricos (p.ej.: caras y voz). Posteriormente, prevemos llevar a cabo análisis de usabilidad que creemos que nos permitirán obtener conocimientos sobre aquellos aspectos de la interacción que puedan estar relacionados con la calidad de servicio percibida por el usuario.

Agradecimientos. Las actividades descritas en este artículo han sido financiadas por el Ministerio Español de Ciencia y Tecnología bajo el proyecto TEC2006-13170-C02-01.

Referencias

1. Koreman, J., Morris, A.C., Wu, D., Jassim, S., Sellahewa, H., Ehlers, J., Chollet, G., Aversano, G., Bredin, H., Garcia-Salicetti, S., Allano, L. Ly Van, B. y Dorizzi, B., "Multi-modal biometric authentication on the SecurePhone PDA", Proc. Of Multi-Modal User Authentication Workshop (MMUA), Toulouse, France. (2006)
2. Oki Press Releases 2006. Vodafone K.K. Selects Oki Electric's Face Sensing Engine for Face Recognition on Mobile Phone. Información disponible en: http://www.oki.com/en/press/2006/z05134e.html
3. OMRON R&D. "OKAO Vision" Face Sensing Technology. Información disponible en: http://www.omron.com/r_d/vision/01.html
4. National Science and Technology Council Subcommittee on Biometrics. "The National Biometrics Challenge". Agosto 2006. Información disponible en http://www.biometrics.gov/Documents/biochallengedoc.pdf
5. BioAPI Consortium, http://www.bioapi.org/
6. Otero-Muras, E, Gonzalez-Agulla, E, Alba-Castro, J.L., Garcıa-Mateo, C., Marquez-Florez, O.W, "An Open Framework For Distributed Biometric Authentication In A Web Environment". Annals of Telecommunications. Vol. 62, No. 1-2. Special issue on multimodal biometrics, (2007)
7. Flipz Software Technology, Disponible en: http://www.flipz.tv/
8. Open Computer Vision Library. Disponible en: http://sourceforge.net/projects/opencvlibrary/
9. The Apache Software Foundation. Apache Tomcat. http://tomcat.apache.org/
10. The Commons FileUpload package. Disponible en: http://commons.apache.org/fileupload/
11. VideoLAN - VLC media player. http://www.videolan.org/
12. Hazen T. J, y Schultz D., "Multi-Modal User Authentication from Video for Mobile or Variable-Environment Applications", Interspeech 2007, pages 1246-1249, (2007)
13. Morris, A.C., Koreman, J., Sellahewa, H. Ehlers, J., Jassim, S., Allano, L. y Garcia-Salicetti, S., "The SecurePhone PDA database, experimental protocol and automatic test procedure for multi-modal user authentication", Tech. Report, Saarland University , Institute of Phonetics, (2006)
14. International Biometric Group, "Survey: Biometrics enter mobile world", Biometric Technology Today – September (2005).

Efectos de la Variabilidad Temporal en el Reconocimiento de Iris

Pedro Tomé González, Fernando Alonso Fernández, Javier Ortega García

Grupo de Reconocimiento Biométrico - ATVS
Escuela Politécnica Superior - Universidad Autónoma de Madrid
Avda. Francisco Tomas y Valiente, 11 - Campus de Cantoblanco
28049 Madrid, España - http://atvs.ii.uam.es
{pedro.tome, fernando.alonso, javier.ortega}@uam.es

Abstract. En este artículo evaluamos los efectos del paso del tiempo entre adquisiciones en reconocimiento de iris. Para los experimentos usamos un sistema de reconocimiento de libre distribución y la base de datos BioSecurID, que contiene 8128 imágenes de iris de 254 individuos adquiridos en cuatro sesiones consecutivas. Los resultados muestran que la separación temporal entre muestras comparadas tienen impacto sobre las tasas de reconocimiento. Se observa una importante degradación de la Tasa de Falso Rechazo, lo que significa que aumenta la variabilidad entre clases. Todas la imágenes de la base de datos han sido adquiridas en similares condiciones controladas, así pues, las diferencias de rendimiento no son causadas por el proceso de adquisición, sino por una reducción en la similitud de las plantillas de iris cuando se adquieren en diferentes instantes.

Key words: Biometría, procesamiento de imágenes, reconocimiento de patrones, reconocimiento de iris, separación temporal.

1 Introducción

La autenticación biométrica ha recibido una considerable atención en los últimos años debido a la creciente demanda del reconocimiento automático de personas. El termino *biometría* se refiere al reconocimiento automático de un individuo basado en su anatomía (por ejemplo, huellas dactilares, cara, iris, geometría de mano, oreja, huella palmar) o en las características de comportamiento (por ejemplo, la firma, forma de andar, la dinámica de tecleo), que no pueden ser robados, extraviados o copiados [1]. Entre todas las técnicas biométricas, el reconocimiento de iris ha sido tradicionalmente considerado como uno de los sistemas de identificación más fiables y precisos [2].

Uno de los inconvenientes de un sistema biométrico es la variabilidad en los datos capturados. Los datos biométricos adquiridos de una persona durante la autenticación pueden ser muy diferentes de los datos que se utilizaron para generar la plantilla durante el registro, lo cual afecta el proceso de comparación [1]. En este artículo, evaluamos los efectos de la separación en el tiempo entre adquisiciones en el reconocimiento de iris. Nuestro objetivo es determinar en qué medida las tasas de reconocimiento se degradan cuando se incrementa el tiempo entre las muestras adquiridas. Los experimentos aquí

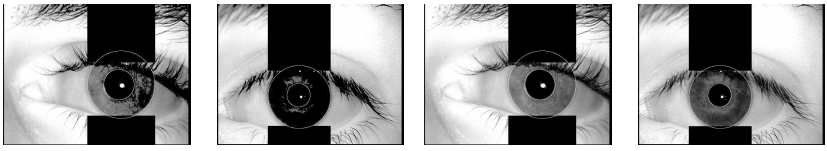
presentados muestran que la Tasa de Falso Rechazo empeora drásticamente. Los experimentos muestran que esta degradación en las tasas de error no es causado por las variaciones en la interacción usuario-sensor, sino por una reducción en la similitud entre las plantillas de iris a lo largo del tiempo.

2 Sistema de reconocimiento

Hemos usado para los experimentos el sistema de reconocimiento[1] desarrollado por Libor Masek [3, 4]. Consiste en la siguiente secuencia de pasos descritos a continuación: segmentación, normalización, codificación y comparación de plantillas.

2.1 Segmentación y Normalización

Para la tarea de segmentación del iris, el sistema utiliza la transformada circular de Hough con el fin de detectar las fronteras de la pupila y del iris. Los límites del iris se modelan como dos círculos concéntricos. El rango de valores de radios de búsqueda se establece manualmente. También se impone un valor máximo a la distancia entre centros de los círculos. En el sistema se anula la utilización del detector de pestañas permitiendo solo la detección de los párpados. Los párpados se modelan como una recta en la parte superior e inferior, y se detectan haciendo uso de la transformada lineal de Hough. La detección de pestañas se basa en una umbralización del histograma. Puede verse un ejemplo en la Figura 1.

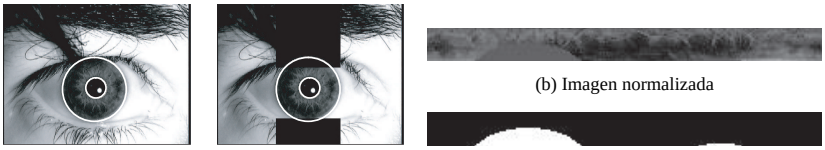


(a) Usando detección de pestañas (b) No usando detección de pestañas

Fig. 1. Ejemplo de eliminación de párpados y pestañas.

Para la normalización de la región del iris, se utiliza una técnica basada en el "*modelo de goma*" desarrollado por Daugman. El centro de la pupila se considera el punto de referencia, trazando los vectores radiales pasan a través de la región de iris. Dado que la pupila puede ser no concéntrica al iris, se reescala la separación radial entre los puntos de muestreo en función del ángulo alrededor del círculo utilizado. La normalización produce un array 2D con dimensión horizontal correspondiente a la resolución angular y dimensión vertical correspondiente a la resolución radial, además de otro array de máscara de ruido de idénticas dimensiones para evitar las reflexiones debidas a las pestañas y párpados detectados en la fase de segmentación. En la Figura 2, se muestra un ejemplo de los pasos de la normalización.

[1] El código puede descargarse libremente desde `www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html`



(a) Imagen original y de ruido

(b) Imagen normalizada

(c) Máscara de ruido

Fig. 2. Ejemplo de los pasos de normalización.

2.2 Codificación de características y Comparación de plantillas

La extracción de características se implementa convolucionando la imagen de iris normalizada con un filtro Log-Gabor 1D. El patrón normalizado 2D del iris se divide en una serie de señales 1D y a continuación, estas señales se convolucionan con el filtro Log-Gabor. Las filas del patrón normalizado 2D se toman como señales de 1D, correspondiéndose cada una a un anillo circular de la región de iris. Esto es así porque se considera que la máxima independencia en iris se produce en la dirección angular [3].

La salida del filtrado se cuantifica en fase en cuatro niveles siguiendo el método de Daugman [5], con cada filtro, produciendo dos bits de datos. La salida de la cuantificación en fase es un código de grises, de modo que al desplazarse de un cuadrante a otro, sólo hay 1 bit de cambio. Esto reducirá al mínimo el número de bits en desacuerdo para patrones del mismo iris ligeramente desalineados, proporcionando mayor precisión en el reconocimiento [3]. El proceso de codificación produce una plantilla binaria con un número de bits de información, y la correspondiente máscara de ruido, que representa las zonas corruptas dentro de los patrones de iris.

Para la comparación entre plantillas (matching), la métrica elegida para el reconocimiento es la distancia de Hamming (HD), ya que son necesarias las comparaciones a nivel de bit. La distancia de Hamming empleada incorpora el enmascaramiento de ruido, de modo que sólo los bits significativos se utilizan en su cálculo. La fórmula viene dada por

$$HD = \frac{\sum_{j=1}^{N} X_j \,(\mathbf{XOR})\, Y_j \,(\mathbf{AND})\, Xn'_j \,(\mathbf{AND})\, Yn'_j}{N - \sum_{k=1}^{N} Xn_k \,(\mathbf{OR})\, Yn_k} \tag{1}$$

donde X_j e Y_j son los dos bits a comparar de cada plantilla, Xn_j e Yn_j son las correspondientes mascaras de ruido para X_j e Y_j, y N es el numero de bits representado por cada plantilla.

Con el fin de dar cuenta de las incoherencias rotacionales, cuando se calcula la distancia de Hamming de dos modelos, una plantilla se desplaza a nivel de bit a izquierda y derecha, calculándose una serie de valores de la distancia de Hamming a partir de sucesivos cambios [5]. Este método corrige desajustes rotacionales en el modelo normalizado del iris causados por diferencias en rotación de imágenes. De entre los valores de distancia calculada se adopta el menor valor calculado.

Fig. 3. LG Iris Access3000.

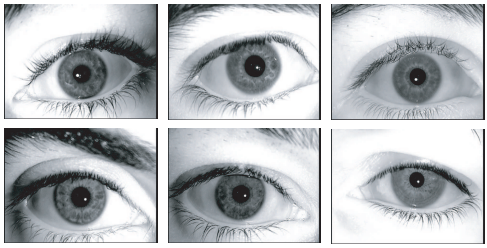


Fig. 4. Ejemplos de iris de la base de datos BioSecurID.

3 Experimentos

3.1 Base de datos y protocolo experimental

Para los experimentos en este documento, usamos las bases de datos BioSec-Baseline [6] y BioSecurID [7]. La base de datos BioSec se compone de 200 usuarios adquiridos en dos sesiones de adquisición, normalmente separadas un periodo de tiempo de entre una y cuatro semanas. La base de datos BioSecurID está formada por 254 usuarios distribuidos en cuatro sesiones de adquisición, normalmente separados de entre una a cuatro semanas entre sesiones consecutivas. Para ambas bases de datos se capturaron un total de cuatro imágenes de iris de cada ojo, cambiando de ojo entre adquisiciones consecutivas. Por tanto, el número total de imágenes del iris es: 200 usuarios $\times$ 2 sesiones $\times$ 2 ojos $\times$ 4 iris = 3200 imágenes de iris para la base de datos Biosec-Baseline, y 254 usuarios $\times$ 4 sesiones $\times$ 2 ojos $\times$ 4 iris = 8128 imágenes de iris para la base de datos BiosecurID. Consideramos cada ojo como un usuario distinto, por lo tanto, tendremos 400 usuarios en la base de datos Biosec y 508 usuarios en la base de datos BiosecurID. La utilización de gafas no se permitió en la adquisición, mientras que el uso de lentes de contacto si estaba permitido.

Ambas bases de datos utilizan el mismo sensor LG Iris Access 3000 (ver Figura 3), el cual genera una imagen de 640 píxeles de ancho y 480 de alto. La adquisición se hizo en un ambiente de oficina, bajo la supervisión de un operador que dio las instrucciones y/o orientaciones necesarias. En cuanto a las condiciones ambientales, se utilizó una iluminación neutral y se colocó al individuo en una silla fija frente al sensor.

La base de datos Biosec se ha utilizado para ajustar los parámetros del sistema de verificación, es decir, la gama de radios de los círculos que modelan los contornos de pupila e iris, así como la distancia máxima permitida entre los centros de ambas circunferencias. Una vez afinado el sistema, se ha utilizado para las pruebas la base de datos BiosecurID. En la Figura 4, se muestran ejemplos de imágenes de iris de BioSecurID. La detección de pestañas no se utiliza en nuestros experimentos. Aunque las pestañas son muy oscuras en comparación con la región circundante, existen otras zonas de la región del iris que son igualmente oscuras. Por lo tanto, el umbral para aislar las pestañas también elimina regiones importantes de iris, tal y como mostramos en la Figura 1, lo que hace que esta técnica sea contraproducente. No obstante, en nuestras bases de datos la oclusión de las pestañas no es muy prominente, por lo que no se utilizó esta técnica para aislar las pestañas.

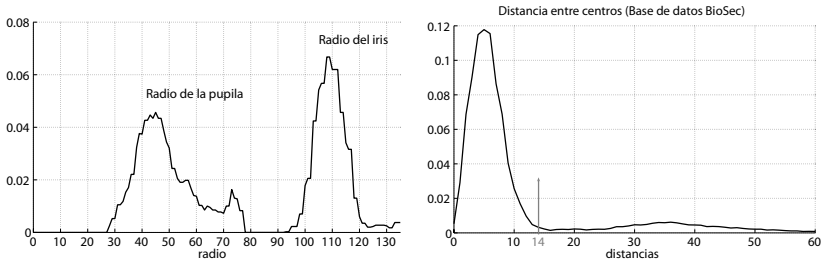


Fig. 5. Distribución de radios de la base de datos BioSec (izquierda) y distribución de distancia entre centros de circunferencias de pupila e iris (derecha).

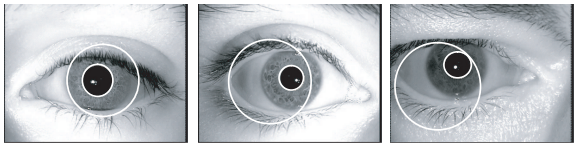


Fig. 6. Ejemplos de imágenes de iris mal segmentadas.

Para la evaluación del rendimiento en verificación, se llevan a cabo los siguientes experimentos en la base de datos BiosecurID: *i*) matchings entre iris de la misma sesiones (*evaluación intra-sesión*); y *ii*) matchings entre iris de la primera sesión contra cada uno de las otras sesiones (*evaluación inter-sesión*). Para la evaluación *intra-sesión*, el protocolo experimental es el siguiente. Se obtienen matchings genuinos (o de usuario) mediante la comparación de cada una de las cuatro imágenes de un usuario con el resto de imágenes del mismo usuario, pero evitando las comparaciones simétricas. Los matchings de impostor se obtienen mediante la comparación de la primera imagen de un usuario con las cuatro imágenes de iris del resto de los usuarios, evitando comparaciones simétricas. Este proceso se repite para cada una de las cuatro sesiones diferentes, por lo que se obtienen cuatro conjuntos de *scores*, uno por cada sesión. Para la evaluación *inter-sesión*, el protocolo experimental es el siguiente. Para un determinado usuario, todas las imágenes de la primera sesión se consideran como las plantillas de registro en el sistema. Se obtienen los matchings genuinos (o de usuario) mediante la comparación de las plantillas correspondientes a las imágenes de la segunda (tercera, cuarta) sesión del mismo usuario. Los matchings de impostor se obtienen mediante la comparación de una plantilla seleccionada al azar de un usuario con una plantilla seleccionada de forma aleatoria del conjunto de imágenes de iris de la segunda (tercera, cuarta) sesión de los demás usuarios. Como resultado, se obtienen tres conjuntos de *scores*.

3.2 Resultados y debate

En la Figura 5 (izquierda) representamos la distribución de radios de las circunferencias que modelan los contornos del iris para la base de datos BioSec. Basándonos en este

Comparaciones Intra-Sesión

	USUARIO	IMPOSTOR
Sesión 1	2036	1213984
Sesión 2	2019	1209327
Sesión 3	2015	1204666
Sesión 4	2007	1201569

Comparaciones Inter-sesión

	USUARIO	IMPOSTOR
Sesión 1 vs. Sesión 2	5310	219981
Sesión 1 vs. Sesión 3	5316	218573
Sesión 1 vs. Sesión 4	5319	221377

Table 1. Número de comparaciones disponibles en los experimentos.

histograma, podemos establecer un rango de 90 a 130 píxeles para la circunferencia externa y de 28 a 78 píxeles para la circunferencia interna. También está representado en la Figura 5 (derecha) la distribución de la distancia entre los centros de las circunferencias de pupila e iris, así como ejemplos de imágenes que en las que falla la extracción cuando superan el valor umbral establecido de valor 14 (Figura 6). Con estas limitaciones, para la base de datos BioSec obtenemos 2631 imágenes de iris correctamente extraídas de las 3200 disponibles, lo que corresponde a una tasa de acierto de aproximadamente el 82.32%. Para BioSecurID, obtenemos 6223 imágenes de iris correctamente extraídas de las 8128 disponibles, lo que corresponde a una tasa de acierto de aproximadamente el 76.56%. Hemos de considerar que los datos obtenidos son coherentes con los de [3], donde una tasa de acierto es de alrededor el 83% para la base de datos CASIA utilizando el mismo sistema. Como resultado de ello, no es posible utilizar todas las imágenes de ojos de la base de datos BiosecurID para la realización de los experimentos. El número de matchings disponibles se resumen en la Tabla 1.

	FRR - INTRA-SESIÓN				FRR - INTER-SESIÓN		
	Sesión 1	**Sesión 2**	**Sesión 3**	**Sesión 4**	**S1 versus S2**	**S1 versus S3**	**S1 versus S4**
FAR=0.01	9.8723	8.469	10.074	11.2606	22.429	25.809	24.262
FAR=0.1	6.9008	6.835	7.593	7.2995	16.7797	18.764	18.227
FAR=1	5.2063	5.0220	5.4839	5.3812	11.4783	13.497	14.0158

Table 2. Resultados para los experimentos de intra-sesión y inter-sesión. Se informa de los FRR en tres puntos específicos de FAR (en %).

Los resultados de verificación los experimentos intra- e inter-sesión están descritos en la Tabla 2. Se describe la Tasa de Falsa Rechazo (FRR) en tres puntos específicos de la Tasa de Falsa Aceptación (FAR). En la Figura 7, también están representadas las distribuciones de error Falso Rechazo (FR) y Falsa Aceptación (FA). Se observa que los índices de error se incrementan considerablemente en los experimentos *inter-sesión* con respecto a los de *intra-sesión* (se observa un aumento de más del doble). Esto revela que separación temporal entre muestras tiene impacto sobre las tasas de reconocimiento. De la Figura 7, cabe señalar que la degradación se produce en la FRR, no en la FAR, lo que significa que el sistema sigue siendo lo suficientemente robusto a los accesos impos-

tores a través del tiempo, pero aumenta la variabilidad intra-clase. Estos resultados se reflejan en la Tabla 3, donde se dan la media y desviación estándar de las distribuciones de *scores* genuinos (o de usuario) e impostor, junto con la representación de dichas distribuciones. Curiosamente, se observa que sólo la media de la distribución de scores genuinos se ve afectada, pero no su desviación estándar. También dan en la Tabla 3 la distancia de Fisher (FD) entre las distribuciones de scores genuinos e impostores, que se define como $FD = (\mu_G - \mu_I)^2/(\sigma_G^2 + \sigma_I^2)$, donde μ_G y σ_G (μ_I y σ_I) son la media y la varianza de la distribución de scores genuinos (distribución de scores de impostor) [8]. Se observa una significante reducción in la FD para los experimentos inter-sesión con respecto a los de intra-sesión.

Observando los experimentos intra-sesión de la Tabla 2, se puede observar que no hay ninguna tendencia en las tasas de error entre las distintas sesiones. Esto también puede ser observado en las distribuciones de scores de la Tabla 3 (arriba). Por otro lado, se observa a partir de los experimentos de inter-sesión que a medida que el tiempo entre muestras se incrementa se produce un aumento de las tasas de error. En particular, son claras las diferencias observadas entre las líneas "s1-s2" de la Figura 7 y las demás. No obstante, parece que una vez que ha pasado un mínimo de tiempo entre las muestras, las tasas de error al parecer no están aumentado. Esto se observa en la Figura 7, donde puede verse una pequeña separación entre las líneas marcadas "s1-s3"y "s1-s4".

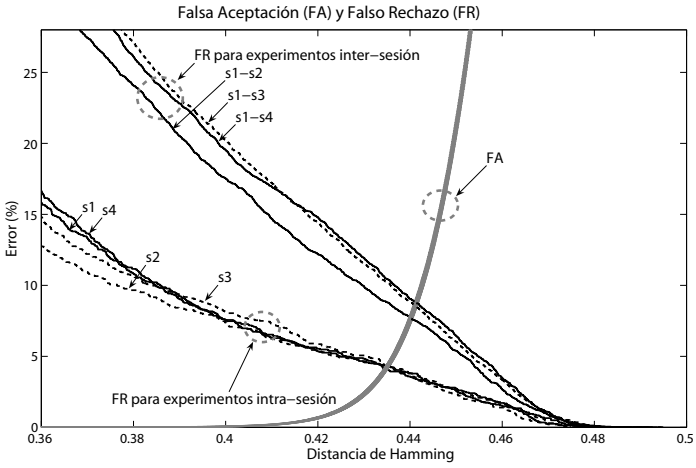


Fig. 7. FA y FR para los experimentos de intra-sesión y inter-sesión.

Ahora estudiamos las diferencias observadas las distribuciones de scores entre los experimentos intra- e inter-clase. Representamos en la Figura 8 la distribución de los dos términos de la distancia de Hamming (Ecuación 1): el denominador (parte izquierda de la Figura 8) representa el número de bits significativos entre las dos plantillas binarias comparadas, mientras que el numerador (parte derecha de Figura 8) es el número de bits distintos entre ellas. Se observa claramente que cuando aumenta la separación temporal, aumenta el número de bits diferentes entre muestras. En otras palabras, la similitud entre

dos plantillas del iris se reduce si se adquieren en diferentes momentos (en nuestros experimentos, la distancia mínima es de una a cuatro semanas). Sin embargo, como se ha observado antes, una vez que ha pasado un mínimo de tiempo entre las muestras, la similitud no se reduce más.

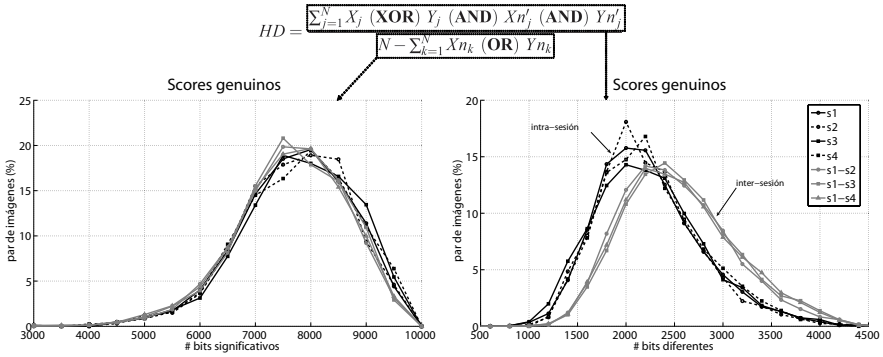
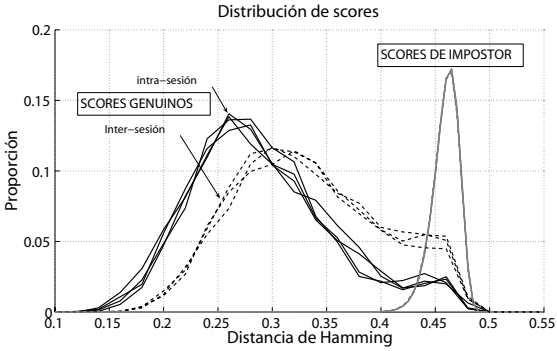


Fig. 8. *Scores* genuinos. Distribución de dos términos de la distancia de Hamming (Ecuación 1) para los experimentos de intra-sesión y inter-sesión experiments.

Las bases de datos utilizadas en nuestros experimentos han sido adquiridas bajo la supervisión de un operador, con iluminación neutral y una pose frontal del contribuyente. Como resultado de ello, no se espera encontrar diferencias significativas entre imágenes de iris de diferentes sesiones debido al proceso de adquisición. Esto es coherente con la Figura 8 (izquierda), donde el número de bits significativos entre las plantillas de iris se mantiene constante con el tiempo. Analizando la distribución de varios parámetros geométricos de la base de datos BiosecurID, tales como (Figura 9): radio de la pupila y el iris, área del iris y tamaño de la máscara de ruido, no se encuentran diferencias notables para las diferentes sesiones, por lo tanto, significa que el proceso de adquisición en sí no tiene efectos sobre las diferencias observadas anteriormente en las tasas de error.

4 Conclusiones

En este artículo se estudian los efectos de la separación temporal entre muestras de adquisición en reconocimiento de iris. En nuestros experimentos se utiliza un sistema de reconocimiento de iris de libre disposición y la base de datos BiosecurID. Esta base de datos contiene imágenes de iris de 254 usuarios adquiridos en cuatro sesiones, separadas de una a cuatro semanas entre sesiones consecutivas, lo que permite evaluar la variabilidad del tiempo. Observamos que el tiempo de separación entre muestras de iris tiene impacto en las tasas de reconocimiento. Se observa un aumento de más del doble en la Tasa de Falso Rechazo. Por el contrario, no se observó ningún efecto sobre la Tasa de Falsa Aceptación. Esto significa que el sistema sigue siendo lo suficientemente



EXPERIMENTOS INTRA-SESIÓN

	μ_G	σ_G	μ_I	σ_I	FD
Sesión 1	0.29	0.067	0.46	0.012	6.11
Sesión 2	0.29	0.064	0.46	0.012	6.63
Sesión 3	0.29	0.068	0.46	0.012	6.07
Sesión 4	0.30	0.066	0.46	0.013	6.05

EXPERIMENTOS INTER-SESIÓN

	μ_G	σ_G	μ_I	σ_I	FD
S1-S2	0.33	0.067	0.46	0.013	3.62
S1-S3	0.34	0.068	0.46	0.013	3.24
S1-S4	0.34	0.068	0.46	0.013	3.24

Table 3. Resultados para los experimentos de intra-sesión e inter-sesión. Distribución de la distancia de Fisher (FD) entre *scores* genuinos y de impostor. Las distribuciones de la media y la desviación estándar de los *scores* genuinos (impostor) se denotan como μ_G y σ_G (μ_I y σ_I) respectivamente.

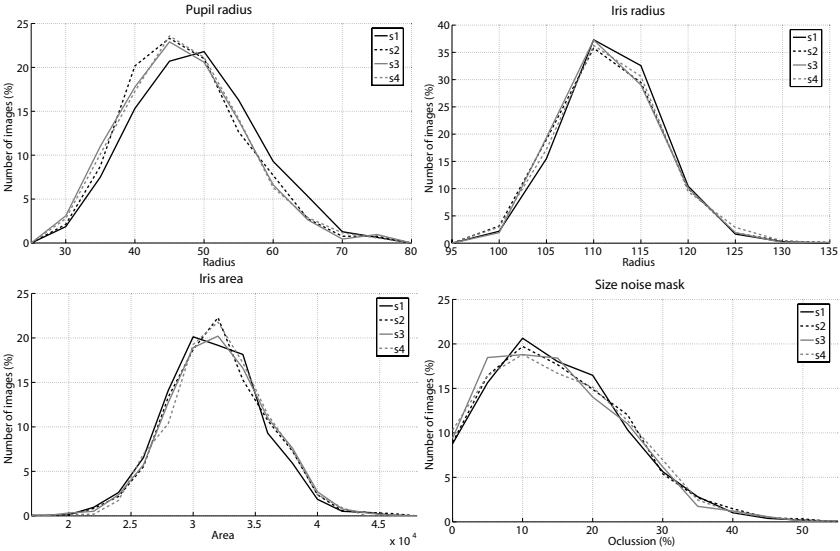


Fig. 9. Distribución de varios parámetros geométricos de las cuatro sesiones (s1 a s4) de la base de datos BiosecurID.

robusto a los accesos impostores a través del tiempo, pero aumenta la variabilidad intra-clase.

La base de datos utilizada en este artículo han sido adquirida bajo condiciones de supervisión, por lo que esta diferencia en el rendimiento a través del tiempo no es causada por el proceso de adquisición (p.e. diferentes interacciones con el sensor). Se demuestra en nuestros experimentos que la similitud entre las plantillas del iris se reduce con el tiempo, a pesar de que son adquiridas bajo similares condiciones controladas.

Las evaluaciones de tecnología existente no han tenido en cuenta los efectos temporales en el reconocimiento de iris [9]. Los resultados obtenidos en este documento nos motivan a la realización de nuevos experimentos utilizando otros algoritmos de reconocimiento y dispositivos de adquisición [10]. También, para hacer frente a este problema, se están estudiando técnicas para actualizar la plantilla y/o el uso de múltiples plantillas [11].

5 Agradecimientos

Este trabajo ha sido financiado por el proyecto MCYT español TEC2006-13141-C03-03. El autor F. A.-F. agradece a la Consejeria de Educacion de la Comunidad de Madrid y al Fondo Social Europeo por financiar sus estudios de Doctorado.

References

1. Jain, A., Ross, A., Pankanti, S.: Biometrics: A tool for information security. IEEE Trans. on Information Forensics and Security **1** (2006) 125–143
2. Jain, A., Bolle, R., Pankanti, S., eds.: Biometrics - Personal Identification in Networked Society. Kluwer Academic Publishers (1999)
3. Masek, L.: Recognition of human iris patterns for biometric identification. MSc thesis, School of Computer Science and Software Engineering, Univ. Western Australia (2003)
4. Masek, L., Kovesi, P.: Matlab source code for a biometric identification system based on iris patterns. School of Computer Science and Software Engineering, Univ. Western Australia (2003)
5. Daugman, J.: How iris recognition works. Proceedings of International Conference on Image Processing **1** (2002) 22–25
6. Fierrez, J., Ortega-Garcia, J., Torre-Toledano, D., Gonzalez-Rodriguez, J.: BioSec baseline corpus: A multimodal biometric database. Pattern Recognition **40**(4) 1389–1392 (2007)
7. Galbally, J., Fierrez, J., Ortega-Garcia, J., Freire, M., Alonso-Fernandez, F., Siguenza, J., Garrido-Salas, J., Anguiano-Rey, E., Gonzalez-de-Rivera, G., Ribalda, R., Faundez-Zanuy, M., Ortega, J., Cardeñoso-Payo, V., Viloria, A., Vivaracho, C., Moro, Q., Igarza, J., Sanchez, J., Hernaez, I., Orrite-Uruñuela, C.: BiosecurID: a Multimodal Biometric Database. Proc. MADRINET Workshop (November 2007) 68–76
8. Marcialis, G., Roli, F.: Fusion of multiple fingerprint matchers by single-layer perceptron with class-separation loss function. Pattern Recognition Letters **26** (2005) 1830–1839
9. Newton, E., Phillips, P.: Meta-analysis of third party evalutions of iris recognition. NISTIR 7440 (2007)
10. K.W. Bowyer, Hollingsworth, K., Flynn, P.: Image understanding for iris biometrics: a survey. Computer Vision and Image Understanding **110** (2008) 281–307
11. Uludag, U., Ross, A., Jain, A.: Biometric template selection and update: a case study in fingerprints. Pattern Recognition **37** (2004) 1533–1542

Speaker Recognition Robustness to Voice Conversion

Mireia Farrús, Daniel Erro, and Javier Hernando

TALP Research Centre, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya, Barcelona
{mfarrus,derro,javier}@gps.tsc.upc.edu

Abstract. Security systems relying on voice identification can be threatened by human voice imitation or synthetic voices. As voice conversion can be seen as a sort of voice imitation, this paper analyses the performance of an automatic speaker identification system by using converted voices in order to know how vulnerable such systems are to this kind of disguise. The experiments are conducted by using intra-gender and cross-gender conversions between two males and two females. The results show that, in general terms, the system is more robust to intra-gender converted voices than to cross-gender ones.

Key words: speaker identification, voice conversion, robustness

1 Introduction

Voice imitation and other types of disguise are potential threats to security systems that use automatic speaker recognition; therefore, several studies have been performed in order to test the vulnerability of speaker recognition systems against imitation by human or synthetic voices.

Automatic voice conversion is the modification of a speaker voice —called *source speaker*— in order to make it being perceived as if another speaker —*target speaker*— had uttered it. Given thus two speakers, the aim of a voice conversion system is to determine a transformation function that *converts* the speech of the source speaker (from which usually a complete database is available) into the speech of the target speaker (from which normally few data are available), replacing the physical characteristics of the voice without altering the message contained in the speech [1, 2].

Several studies have been done to test the vulnerability of speaker recognition systems against voice disguise and imitations by human or synthetic voices. An experiment reported in [3] tried to deceive a state-of-the-art speaker verification system by using different types of artificial voices created with client speech. Other works related to the vulnerability of automatic recognition systems to specifically created synthetic voices can be found in [4] and [5], where the impostor acceptance rate is increased by modifying the voice of an impostor in order to target a specific speaker.

This paper analyses the robustness of an automatic speaker recognition system against converted voices. The conversion system used to get such converted voices comes up from the improvement of a synthesis system based on the harmonic plus stochastic model [6], which uses frames of fixed length, and where a conversion module has been implemented. The performance of the systems has been demonstrated to be notable, even when no training parallel corpus is available. This is partly due to the fact that the system takes advantage of the high flexibility of the harmonic plus stochastic model in order to minimise the errors derived of the signal reconstruction from their already modified parameters [6].

Next, the voice synthesis system and the voice conversion method are introduced, and the voice conversion database used in the experiments is described in section 3. In order to analyse the robustness of an automatic speaker recognition system against converted voices, the system is tested against both original and converted voices (section 4), so that the comparison will allow to see if the performance gets worse by using voice conversion. Finally, conclusions are presented in section 5.

2 Description of the voice conversion system

The aim of voice conversion systems is to modify the voice produced by a source speaker, for it to be perceived by listeners as if it had been uttered by a target speaker. During the training phase, given a speech database recorded from specific source and target speakers, the system has to determine the optimal transformation for converting one voice into the other one. First, the involved speech signals are frame-by-frame analysed, according to a certain speech model that allows signal manipulation. Then, each analysed frame is translated into a fixed number of parameters with good conversion properties. Finally, after finding the correspondence between the acoustic characteristics of the speakers, the transformation function is learnt. During the conversion phase, the system applies such function to convert new input utterances of the source speaker. Fig. 1 shows the general architecture of a voice conversion system.

The speech model chosen for analysis, transformation and reconstruction of signals is the harmonic plus stochastic model (HSM) [6], which provides high quality speech reconstruction and allows the manipulation of both waveform and spectrum in a very flexible way. Moreover, the model is compatible with many voice transformation methods. The harmonic component captures the part of the signal that is similar to a periodic waveform, and it is characterized by the frequencies, the amplitudes and the phases of the harmonically related sinusoids, whereas the stochastic component containing all the non-sinusoidal signal components is modelled by means of LPC filters. During the analysis, all these features are measured at a constant frame rate. During synthesis, the frames are reconstructed and overlapped.

Converting voices directly from the HSM parameters (amplitudes, frequencies, phases and stochastic LPC filters) is extremely complicated. Instead, the problem can be decomposed into three different sub-problems: pitch conversion,

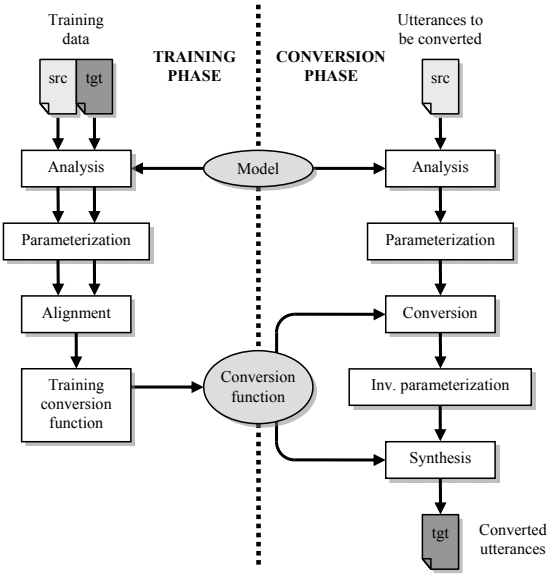


Fig. 1. General architecture of a voice conversion system.

harmonic conversion and stochastic conversion. Since both pitch and stochastic component are represented by very simple parameters (a scalar and an all-pole filter, respectively), the parameterisation task is narrowed to translate the harmonic component into an all-pole filter [7]. Before applying spectral conversion techniques, the harmonic and stochastic all-pole filters are transformed into their associated line spectral frequencies (LSFs) [8], which have very good properties for linear transformations. In order to reconstruct the speech signal from converted LSF vectors, they need to be transformed back into all-pole filters. The stochastic part does not need any extra processing, whereas the harmonic all-pole filter has to be sampled in the frequency domain at multiples of the converted pitch, so that new amplitudes and phases are obtained.

If a parallel training corpus (where the same sentences are uttered by both source and target speakers) is available, the alignment process is simplified and the accuracy of the voice conversion system is increased. In order to train adequate voice conversion functions, a correspondence must be established between the parameter vectors representing the speech frames of the source speaker and those of the target speaker. The method chosen for alignment of source and target frames gives very good results despite its simplicity [9] and consists of the following steps: (i) the boundaries of the phonemes are determined by automatic segmentation based on hidden Markov models, (ii) the phoneme boundaries are used as anchor points to establish a piecewise linear time-warping function for the source-target pairs of parallel sentences, and (iii) each acoustic source vector is paired with the closest target neighbour in the warped time scale.

The method used for spectral envelope conversion is a particular implementation of the GMM-based solution proposed by Stylianou [10] and improved by Kain [11]. It is known that the transformation of the voiced sounds (in which the harmonic component exists) is much more important for voice conversion than the transformation of the unvoiced sounds; therefore, only the voiced frames are transformed, so that only the aligned frame pairs where both members are voiced are considered for training. The spectral conversion method used in this paper consists in applying a GMM-based transformation function to the harmonic LSF vector, and then predicting the stochastic LSF vector from the transformed harmonic one only at voiced frames.

After the alignment and during the training phase, the acoustic mapping between the source speaker and the target speaker is given by a set of frame pairs of the form $\{x_h, x_s\} \leftrightarrow \{y_h, y_s\}$, where the sub-index h denotes the LSF vector of the harmonic component and s denotes the LSF vector of the stochastic component. From now on, and for simplicity, x_h and y_h will be called simply x and y. The paired p-dimensional LSF vectors x and y are concatenated together to form $2p$-dimensional vectors $z = [x^T y^T]^T$. Then, a GMM given by the weights $\{\alpha_i\}$, mean vectors $\{\mu_j\}$ and covariance matrices $\{\Sigma_i\}$ of m different Gaussian components is estimated from the set of vectors $\{z\}$ by means of the expectation-maximization algorithm. Given the relationship between vectors:

$$\mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix}, \qquad \Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}, \tag{1}$$

the probability of a vector x belonging to the ith Gaussian component of the model $p_i(x)$ can be expressed as:

$$p_i(x) = \frac{\alpha_i N(x, \mu_i^x, \Sigma_i^{xx})}{\sum_{j=1}^{m} \alpha_j N(x, \mu_j^x, \Sigma_j^{xx})} \tag{2}$$

where $N(.)$ denotes the Gaussian distribution. Now, the following transformation function can be applied:

$$F(x) = \sum_{i=1}^{m} p_i(x) \left[\mu_i^y + \Sigma_i^{yx} (\Sigma_i^{xx})^{-1} (x - \mu_i^x) \right]. \tag{3}$$

Under the assumption that the stochastic component is highly correlated with the harmonic component in voiced frames, a stochastic envelope prediction function can be learnt using the training speech frames of the target speaker. Once the transformation function for the harmonic component is trained, all the harmonic-stochastic vector pairs of the form $\{y, y_s\}$ and the target speaker's acoustic model given by $\{\alpha_i, \mu_i^y, \Sigma_i^{yy}\}$ can be used for calculating the m vectors $\{\nu_i\}$ and matrices $\{\Gamma_i\}$ that minimise the error of the following prediction function:

$$y_s = \sum_{i=1}^{m} p_i^y(y) [\nu_i + \Gamma_i (\Sigma_i^{yy})^{-1} (y - \mu_i^y)] \tag{4}$$

During the conversion phase, the prediction function is applied to the converted harmonic LSF vector $F(x)$ instead of y.

With regard to pitch level conversion, a basic adaptation between speakers gives good enough results in most of the cases, especially when the speech signals used for test are emotionally neutral. Since $\log(f_0)$ is well represented by a normal distribution, the pitch level is well converted by applying the following transformation based on replacing the mean and variance of the distribution:

$$\log f_0' = \mu^y_{\log f_0} + \frac{\sigma^y_{\log f_0}}{\sigma^x_{\log f_0}} \Big(\log f_0 - \mu^x_{\log f_0} \Big) \; . \tag{5}$$

The full voice conversion system described here is reported to provide very good results in terms of similarity between converted and target voices, although the quality of the converted signals is affected by a certain over-smoothing effect caused by the statistical transformation procedure [2].

3 Voice conversion database

The database used for voice conversion was made available by UPC for the evaluation campaigns of the TC-STAR project [12]. The voice conversion corpora contain around 200 sentences in Spanish and 170 in English —although only the Spanish ones were used in these experiments— uttered by four different professional bilingual speakers, 2 males and 2 females. The average duration of the sentences is 4 seconds, so that about 10-15 minutes of audio were available for each speaker and language. The sentences uttered by the speakers are exactly the same, so that parallel training corpora can be used for training voice conversion functions, In addition, the sentences were recorded as mimic sentences. This means that there were no significant prosodic differences between speakers, since they all were asked to imitate the same prerecorded pattern with neutral speaking style for each of the sentences.

4 Identification Experiments

First of all, the original data set consisting of all four voices described in the previous section was divided in three sets of sentences. The first set was set aside to train the transformation function of the conversion system, and the second and third set of sentences were used to train and test the automatic recognition system, respectively. Each of the four original voices was converted to the rest of the voices. Since there are 12 pairs of source-target voices, a set of 12 converted voices was obtained: four sets corresponding to intra-gender conversions (female to female and male to male conversions), and eight sets corresponding to cross-gender conversion (female to male and male to female conversions). Each set of converted voices consisted of 100 sentences.

The transformation function for the conversion system was trained using 10, 30 and 80 pairs of source-target sentences. Other 10 original sentences were used

to train each of the four speaker models of the recognition system, and 100 more original sentences, together with the converted sentences, were used for testing. The recognition system utilised in the identification experiments was a conventional 32-component GMM system, using short-term feature vectors consisting of 20 MFCC with a frame size of 24 ms and a shift of 8 ms. The corresponding delta and acceleration coefficients were also included.

In order to test the performance of the recognition system, a preliminary experiment was conducted by using only the original voices. Table 1 shows the corresponding identification matrix, where 100 sentences of each original voice were identified from the closed set of four speaker models. Since it was a rather simple experiment that used a low amount of speakers, it gave a high performance, leading to a percent identification of 100% in three of the four voices. Only one of the males (M1) was confused once with the other male (M2), which suggests —given the high performance of the system— that both male voices are characterised by a significant degree of similarity.

Table 1. Identification matrix for two male (M) and two female (F) original voices.

	F1	**F2**	**M1**	**M2**
F1	100	0	0	0
F2	0	100	0	0
M1	0	0	99	1
M2	0	0	0	100

The identification experiments were conducted by testing both intra-gender and cross-gender converted voices. The system tried to identify 100 sentences of each converted voice again from the closed set of four speaker models. Moreover, three sets of converted voices were identified, according to the sentences used in training the transformation function (10, 30 or 80), in order to see how the amount of training data in the conversion phase influenced the performance of the recognition system.

Table 2. *Source* (a), *target* (b) and *other* (c) identifications using 10 sentences in training the transformation function.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	16	93	-

(a) *Source* identification.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	-	46	100
F2	100	-	98	100
M1	100	98	-	100
M2	100	84	7	-

(b) *Target* identification.

Source voice	**Target voice**			
	F1	**F2**	**M1**	**M2**
F1	-	-	54	0
F2	0	-	2	0
M1	0	2	-	0
M2	0	0	0	-

(c) *Other* identification.

Tables 2, 3 and 4 show the identification results corresponding to the number of sentences used to train the transformation function: 10, 30 and 80, respectively. (The converted F1_to_F2 voices by using 10 training sentences were damaged and not available at the time of doing the experiments). In each table, three types of identification are distinguished: (a) **source**: where the converted voice was identified as its corresponding source speaker, (b) **target**: where the converted voice was identified as its corresponding target speaker, and (c) **other**: where the converted voice was identified as a speaker other than the corresponding source and target speakers.

Table 3. *Source* (a), *target* (b) and *other* (c) identifications using 30 sentences in training the transformation function.

Source voice	Target voice F1	F2	M1	M2
F1	-	-	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	9	92	-

(a) *Source* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	99	43	100
F2	100	-	95	100
M1	100	98	-	100
M2	100	91	8	-

(b) *Target* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	1	57	0
F2	0	-	5	0
M1	0	2	-	0
M2	0	0	0	-

(c) *Other* identification.

Table 4. *Source* (a), *target* (b) and *other* (c) identifications using 80 sentences in training the transformation function.

Source voice	Target voice F1	F2	M1	M2
F1	-	0	0	0
F2	0	-	0	0
M1	0	0	-	0
M2	0	5	72	-

(a) *Source* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	100	87	100
F2	100	-	100	100
M1	100	99	-	100
M2	100	95	28	-

(b) *Target* identification.

Source voice	Target voice F1	F2	M1	M2
F1	-	0	13	0
F2	0	-	0	0
M1	0	1	-	0
M2	0	0	0	-

(c) *Other* identification.

The identification results corresponding to 30 training sentences are also plotted in Fig. 2, in which the identification types are also represented by different colours: green, yellow and red for *source*, *target* and *other* identifications, respectively.

Regarding intra-gender identification, the results show that most of the converted voices were identified as their target voices, so that the recognition system failed in identifying the converted voice as the real source voice. Nevertheless, there is one case in which the performance of the system was better —or, in other words, where the voice conversion was not so successful. This is the conversion of the second male to the first male (M2_to_M1). Most of the speakers were identified as the original source voice (M2) instead of as the target voice (M1).

This could probably be explained by the fact that speaker M2 may be highly characterised by his unvoiced segments, and since these are not converted by the system, this unvoiced characteristics still remain in the converted M2_to_M1 voice. However, the identification as the source voice —which will be referred to as *correct identification* by convention— decreases as the amount of conversion training data increases.

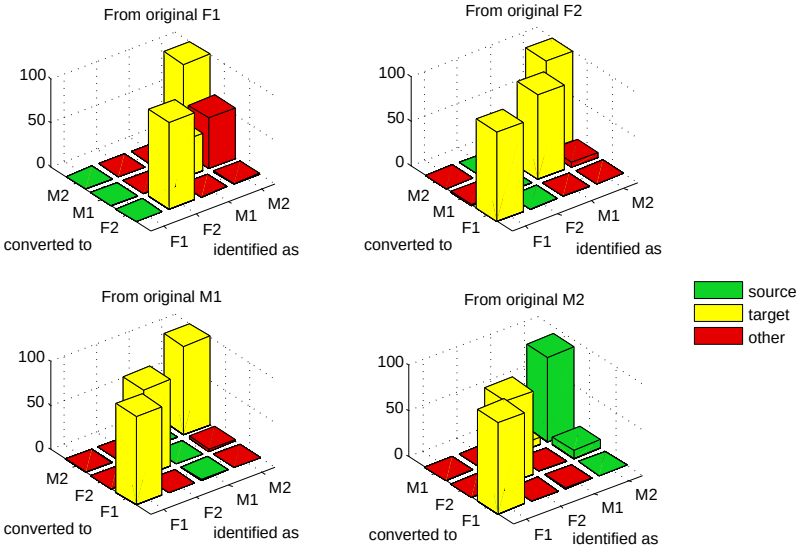


Fig. 2. Identification of each converted voice using 30 sentences in the transformation function. Green, yellow and red bars indicate *source*, *target* and *other* identification, respectively.

It seems thus that the conversion system has difficulties in converting M2 to M1, which could be explained by the fact (seen in Table 1) that both M1 and M2 seem to be similar. However, the reverse phenomenon (M1_to_M2 identified as M1) is not observed in these experiments. Moreover, since the converted F1_to_F2 voice is strangely identified as the male speaker M2 in Table 1, it seems that the recognition system has a slight tendency to identify any speaker as M2.

On the other hand, half of the eight sets of cross-gender converted voices lead to a *miss identification* and *correct conversion* equaling 100%; i.e. not only were the converted speakers not identified as the corresponding source speaker (*miss identification*) but they also were identified as the corresponding target speaker (*correct conversion*).

The other half of the cross-gender conversions were not completely recognised as their corresponding target voices. These are those conversions trying to convert a female speaker to M1 and a male speaker to F2. All the errors are a miss conversion to speaker M2, except in the conversion M2_to_F2, where this

errors can be seen, in fact, as a correct identification of the speaker M2. The worse results are found in the F1_to_M1 conversion, where the tendency of the system to identify speakers as if they were speaker M2 is summed to the hypothetic similarity between M1 and M2 seen in Table 1. In all cases, however, an increase of the correct conversion is observed when the transformation function is trained using 80 sentences.

Summarising, Table 5 shows the types of identification generated by both intra-gender and cross-gender conversions, which are also plotted in Fig. 3. In general terms, intra-gender conversion tends to be identified as its corresponding source speaker in higher degree than cross-gender conversion. On the other hand, cross-gender conversion tends to be more *successful* (speaking in conversion terms) than the intra-gender one, since the percentage of target identification is greater. Nevertheless, cross-gender conversion also leads to a higher percentage of *other* identification; ie. an erroneous conversion in which the converted voice is not identified as either of the source and target speakers.

Table 5. Identification in percent of intra-gender and cross-gender conversions depending on the type of identification generated (*source*, *target* and *other*), where the transformation function has been trained using 30 sentences.

Conversion type	**Source**	**Target**	**Other**
Intra-gender	23.0%	76.7%	0.3%
Cross-gender	1.1%	90.9%	8.0%

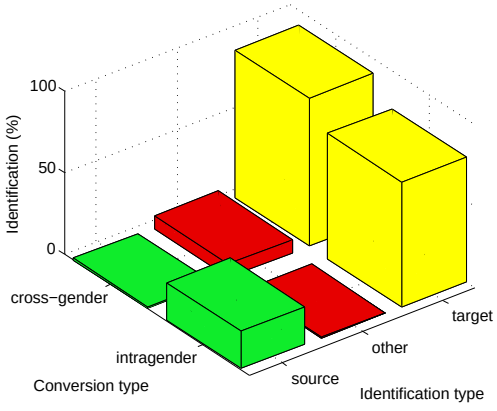


Fig. 3. Identification of intra-gender and cross-gender conversions using 30 training sentences depending on the type of identification generated (*source*, *target* and *other*).

5 Conclusions

In this paper, a set of experiments has been proposed in order to analyse the behaviour of an automatic speaker recognition system against converted voices, using two male and two female voices and several amounts of sentences to train the transformation function. In these experiments, most of the converted voices were identified as their corresponding target speaker; however, they failed sometimes to deceive the system and the source voice was recognised, especially in the intra-gender conversions, which leads to think that the recognition system may be more robust to these kind of conversions than the cross-gender ones. The current results also point out that some voices are more difficult to convert than others, and that the correct identification decreases as the amount of conversion training data increases. Nevertheless, the amount of training data is small enough to interpret the results with extreme caution.

References

1. Duxans, H. Voice Conversion applied to Text-to-Speech systems. PhD Thesis, Universitat Politècnica de Catalunya, Barcelona (2006)
2. Erro, D., Moreno, A.: Sistema de síntesis armónico/estocástico en modo pitch-asíncrono aplicado a conversión de voz. In: Proceedings of the IV Jornadas en Tecnología de Habla. Zaragoza (2006)
3. Lindberg, J., Blomberg, M.: Vulnerability in speaker verification: A study of technical impostor techniques. In: Proceedings of the Eurospeech, pp. 1211–1214. Budapest, Hungary (1999)
4. Masuko, T., Tokuda, K., Tobayashi, T. Imposture using Synthetic Speech Against Speaker Verification Based on Spectrum and Pitch. In: Proceedings of the ICSLP. Beijing, China (2000)
5. Matrouf, D., Bonastre, J.F., Fredouille, C. Effect of speech transformation on impostor acceptance. In: Proceedings of the ICASSP. Tolouse, France (2006)
6. Erro, D., Moreno, A., Bonafonte, A. Flexible Harmonic/Stochastic Speech Synthesis. In: Proceedings of the 6th SSW6. Bonn, Germany (2007)
7. El-Jaroudi,A., Makhoul, J. Discrete All-Pole Modeling. In: IEEE Transactions on Signal Processing (1991)
8. Itakura, F. Line spectrum representation of linear predictive coefficients of speech signals. In: Journal of the Acoustical Society of America, vol. 57 (1975)
9. Duxans, H., Erro, D., Pérez, J., Diego, F., Bonafonte, A., Moreno, A. Voice Conversion of Non-Aligned Data using Unit Selection. In: Proceedings of the TC-STAR Workshop on Speech-to-Speech Translation, Barcelona (2006)
10. Stylianou, Y. Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD Thesis, École Nationale Supérieure des Télécommunications. Paris, France (1996)
11. Kain, A. High resolution voice transformation. PhD Thesis, OGI School of Science and Engineering (2001)
12. Bonafonte, A., Höge, H., Kiss, I., Moreno, A., Ziegenhain, U., van den Heuvel, H., Hain, H.U., Wang, X.S., Garcia, M.N. TC-STAR: Specifications of language resources and evaluation for speech synthesis. In: Proceedings of the LREC. Genoa, Italy (2006)

Nuevos Algoritmos y Ataques a Sistemas de Identificación Biométrica basados en Reconocimiento de Iris

Alberto de Santos Sierra[1], Carmen Sánchez Ávila[2] y Vicente Jara Vera[3]

[1] Grupo de Biometría y Tratamiento Numérico de la Información.
Centro de Domótica Integral (CeDInt)
<alberto@cedint.upm.es>
[2] Grupo de Biometría y Tratamiento Numérico de la Información.
Centro de Domótica Integral (CeDInt)
<csa@mat.upm.es>
[3] Dpto. de Matemática Aplicada a las Tecnologías de la Información.
ETSI de Telecomunicación. Universidad Politécnica de Madrid.
Ciudad Universitaria s/n, 28040 Madrid
<vjara@mat.upm.es>

Resumen En este trabajo se recogen dos enfoques claramente definidos: Primeramente, se propone una mejora a los sistemas actuales de detección de iris, tanto en detección de pupila, como en detección de iris en sí misma. Dichos algoritmos rompen con el esquema clásico de aislamiento de iris, y proponen una nueva idea en este campo. Además, se utilizarán bases de datos actuales para la evaluación de los resultados. Por otro lado, se presenta un esquema de ataque a un sistema de iris en el que a partir del patrón biométrico se reproduce una imagen de iris capaz de confundir a dicho sistema, mediante la utilización de algoritmos genéticos. Este novedoso procedimiento, permitiría a un determinado usuario falsificar su identidad, utilizando simplemente el patrón biométrico de iris de otro usuario. Los resultados de este algoritmo muestran como esto es factible. Además, no existe en la literatura nada similar en cuestión de ataques de este tipo a un sistema de iris.

1. Introducción

El reconocimiento de iris posee diferentes etapas desde el momento en que la imagen es capturada por una cámara, hasta que el sistema es capaz de decidir si el usuario que está accediendo es en verdad quien dice ser, [5], [6], [10]. Estas etapan involucran primeramente un preprocesamiento de la imagen (detección de pupila, iris, párpados, pestañas, ...), extracción de características, procesamiento de las mismas, y comparación. Los algoritmos aquí presentados están más relacionados con el preprocesamiento de la imagen adquirida, concretamente con algoritmos de detección de pupila e iris. Dichos algoritmos, como se verá más adelante con detalle, están basados en morfología matemática, [8]. Por otro lado, los sistemas biométricos presumen de ser capaces de resistir ataques

tales como el acceso de un individuo que haga pasarse por otro usuario. Sin embargo, en sistemas basados en huella ya es posible crear una huella a partir de las minucias obtenidas, o almacenadas en una base de datos. Es decir, conociendo únicamente el patrón biométrico se puede obtener qué huella proporciona dicho patrón, [3]. Inspirado en los resultados obtenidos para huella, este trabajo propone un esquema similar en cuanto al concepto: Obtener una imagen de iris a partir de su patrón biométrico.

Para realizar esto, los algoritmos genéticos son propuestos como una útil herramienta, [1], [2], [7]. Puesto que el tiempo no es una cuestión importante cuando se intenta atacar a un sistema biométrico, sino que prima más la precisión con la que se obtenga el resultado requerido, el uso de los algoritmos genéticos queda por lo tanto justificado.

El documento comenzará por lo tanto tratando los algoritmos de detección de pupila y de iris, finalizando posteriormente con los ataques a estos sistemas biométricos, aportando las oportunas conclusiones y líneas futuras de investigación.

2. Detección de Pupila

El algoritmo de detección de pupila presenta dos nuevos conceptos que no han sido usados con anterioridad. La primera idea trata sobre la eliminación de los brillos en la pupila, algo muy común en imágenes de iris, a partir de la siguiente transformación presentada en la Ecuación 1

$$I'(x,y) = \cos\left(\frac{2\pi}{255} I(x,y)\right) \tag{1}$$

donde $I(x,y)$ es la imagen de iris a la que se quiere eliminar dicho efecto indeseable. Con esta transformación se consigue que aquellos valores cercanos a 255, i.e. colores cercanos al blanco (destellos en la pupila), y aquellos valores cercanos a 0, i.e. colores cercanos al negro (la pupila), posean la misma intensidad. Una vez que esto se consigue, se realizan operaciones morfológicas para eliminar pequeños detalles, y así aislar completamente la pupila, tras una detección de bordes y una umbralización. Es importante resaltar, que el posterior algoritmo de detección de iris basa parte de su fortaleza en una buena detección de pupila, y que además, no es necesario obtener toda la circunferencia que rodee a la pupila, si no únicamente unos pocos puntos.

Los resultados mostrados en la Figura 1 se corresponden con la base de datos Casia V3, [4].

Por otro lado, y continuando con lo relativo a la detección de pupila, se propone un algoritmo para la detección de ésta en bases de datos donde detectar la pupila require un esfuerzo mayor, como es el caso de la base de datos ICE, [9].

En este caso, se utiliza morfología matemática pero aplicada no a la imagen en sí, sino al resultado de dividir la imagen en sus bits correspondientes. Es decir, puesto que cada color está representado con 8 bits, intensidades desde 0 (negro)

Figura 1. (Izq) Imagen Original, (Centro) Resultado de la Transformación con el coseno, (Dch) Resultado de la Segmentación

hasta 255 (blanco), se irán formando capas de imágenes con los bits de cada pixel, desde el más significativo al menos significativo. Con esto se obtendrán 8 capas, cada una de ellas con diferente información sobre la imagen en sí misma. Tomando la capa correspondiente al segundo bit menos significativo, y aplicando posteriormente morfología matemática para hacer más preciso el resultado, se obtienen las imágenes en la Figura 2.

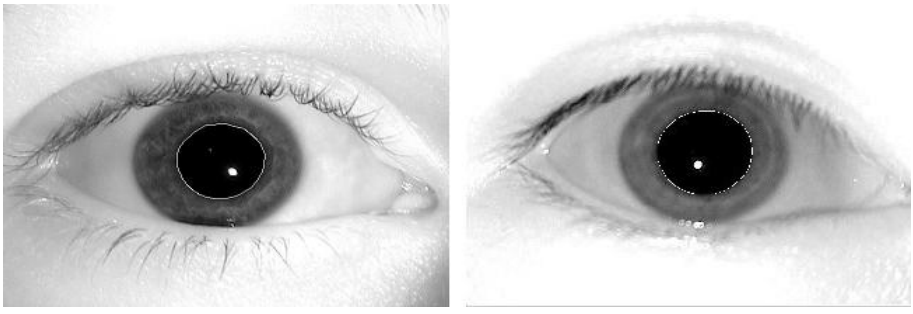


Figura 2. Resultados del procesamiento basado en mapa de bits en la base de datos ICE.

Finalmente, combinando ambos resultados se puede obtener un algoritmo potente de detección de pupila, para imágenes sencillas (versiones de CASIA v1), y para imágenes más complejas, como es la base de datos ICE.

Por último queda decir, que la detección de pupila quedará representada por la Figura 5, donde se puede apreciar que para representar la segmentación de pupila, únicamente hacen falta cinco puntos: dos horizontales, dos verticales y el centro.

3. Detección de Iris

Una vez que se ha detectado tanto los puntos importantes de la pupila como el centro, se procede a extraer el contorno del iris, o en su defecto, información suficiente para poder dilucidar en un posterior control de calidad, cómo de buena es la imagen, cómo de abierto está el ojo, si el ojo está desplazado, etc...

Estos puntos, pueden apreciarse en la Figura 5, donde se proporciona el resultado del algoritmo que se explica a continuación. Una vez obtenido el centro de la pupila (el centro de iris no será calculado, siendo ésta otra de las ventajas de este algoritmo), se centran tres distribuciones diferentes (ver Cuadro 1 y Figura 3), cuya finalidad es simplemente degradar (aclarar) aquellas componentes más alejadas de la pupila, dejando inalteradas aquellas componentes más cercanas al centro de la misma. Posteriormente, se aplica un filtro basado también en operadores morfológicos que se encarga de dejar pasar aquellas componentes más oscuras en una imagen.

Nombre	Expresión Matemática	Parámetros
Gaussiana	$A_\gamma e^{-\sigma_x(x-x_0)^2} e^{-\sigma_y(y-y_0)^2}$	A_γ, σ_x, σ_y, (x_0, y_0)
Coseno I	$A_I cos(x-x_0) cos(y-y_0)$	A_I, (x_0, y_0)
Coseno II	$A_{II} cos^2(x-x_0) cos^2(y-y_0)$	A_{II}, (x_0, y_0)

Cuadro 1. Descripción matemática de las distribuciones

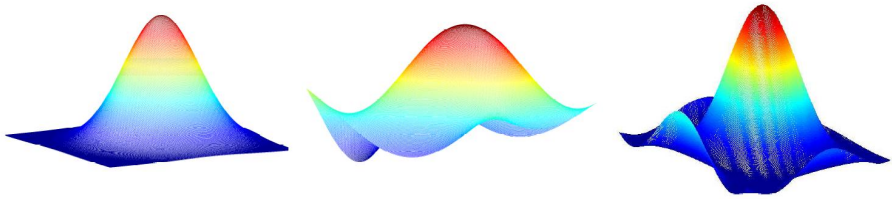


Figura 3. Distribuciones para degradar las imagenes de iris. Su forma hace que las componentes más alejadas a la pupila se vean atenuadas.

La combinación de estas distribuciones junto con el filtro permiten acotar completamente el iris, pero no dar un contorno que ajuste perfectamente el iris. Sin embargo, los puntos aportados por este algoritmo (puntos verticales y horizontales) proporcionan información más que suficiente para poder rechazar una imagen, y pedir otra captura, o por el contrario aceptarla, y continuar con el proceso de extracción de características.

El resultado de los diferentes filtros, puede apreciarse en la Figura 4, donde de izquierda a derecha y de arriba hacia abajo, se aprecia la imagen original, y los resultados provenientes de las distribuciones gaussianas, Coseno I y Coseno II, respectivamente.

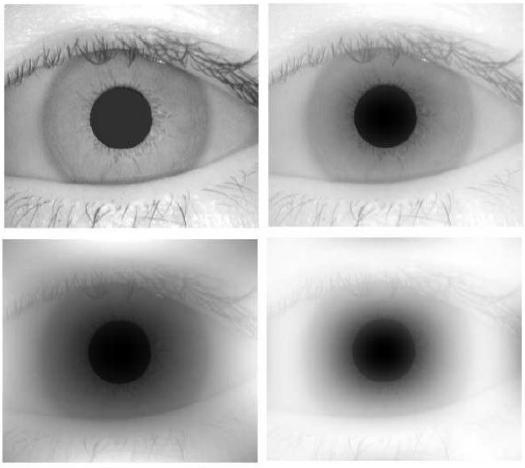


Figura 4. Resultado visual de las diferentes distribuciones

Por último, se presenta en la Figura 5, el resultado final de la segmentación, reuniendo no sólo la detección de iris, sino también la detección de pupila.

Finalmente, los parámetros de la distribución gaussiana (ver Cuadro 1) pueden controlarse y relacionarse de tal forma, que se pueda obtener una segmentación de iris tal y como ha sido siempre concebida, es decir, detectando completamente el iris en la imagen. Mediante una red neuronal que relacione σ_x y σ_y con las medidas obtenidas en una primera segmentación de iris, se obtienen los resultados ofrecidos en la Figura 6. Todos estos resultados se aprovechan del gran contraste de color existente entre el iris (siempre de algún color, a no ser que exista aniridia), y la esclera (parte blanca de los ojos).

Sin embargo, aunque pueda conseguirse esta resolución, con la segmentación ofrecida en la Figura 5 es suficiente, obteniendo resultados muy importantes no sólo en precisión a la hora de segmentar, sino en tiempos de procesamiento pues para realizar la detección de pupila y la detección de iris, únicamente emplea 2.2 segundos de media en realizar toda la segmentación.

4. Ataque a sistemas de iris

La idea de este ataque reside en el siguiente escenario: Sea un usuario $\mathcal{A}$ cuyo patrón biométrico es $\mathcal{T}_\mathcal{A}$. Sea ahora un usuario $\mathcal{B}$ que intenta acceder al sistema haciéndose pasar por $\mathcal{A}$, contando únicamente con $\mathcal{T}_\mathcal{A}$. El problema consiste en

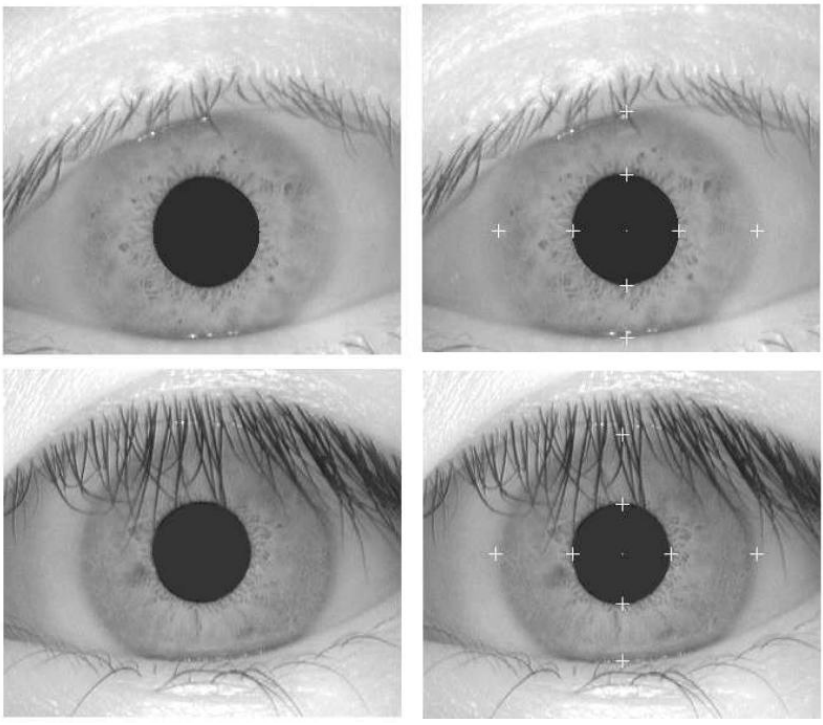


Figura 5. Resultado final de la segmentación

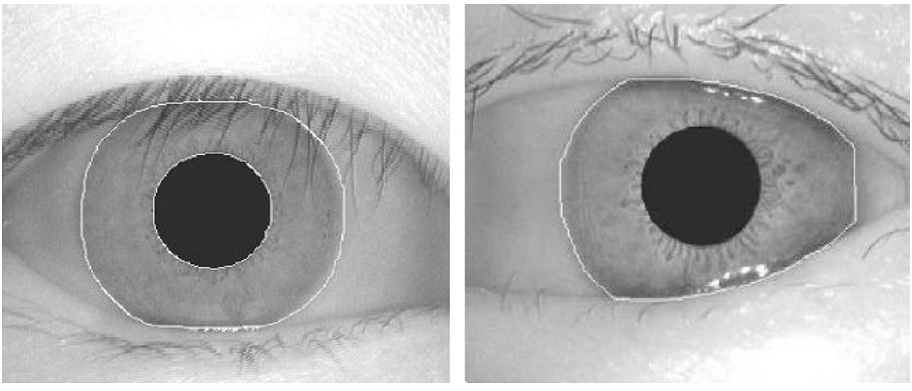


Figura 6. Segmentación clásica mediante este nuevo método

conseguir acceder al sistema a partir del patrón $\mathcal{T}_{\mathcal{A}}$ y con la imagen de iris de $\mathcal{B}$. Una vez que $\mathcal{B}$ ha conseguido una imagen de iris cuyo patrón de iris es $\mathcal{T}_{\mathcal{A}}$, el sistema será incapaz de distinguir entre $\mathcal{A}$ y $\mathcal{B}$. Para solventar dicho problema, se ha implementado un algoritmo genético capaz de transformar la imagen de ojo del usuario $\mathcal{B}$, de tal forma que su patrón, $\mathcal{T}_{\mathcal{B}}$, coincide tanto como se desee con el patrón a falsificar, $\mathcal{T}_{\mathcal{A}}$. Además, el patrón de iris es extraído utilizando una corona circular, y promediando de forma radial los valores de intensidades existentes a lo largo de la corona circular para todos los ángulos, [10].

Una vez presentado el problema, en la Figura 7 se aprecian los dos usuarios $\mathcal{A}$ y $\mathcal{B}$.

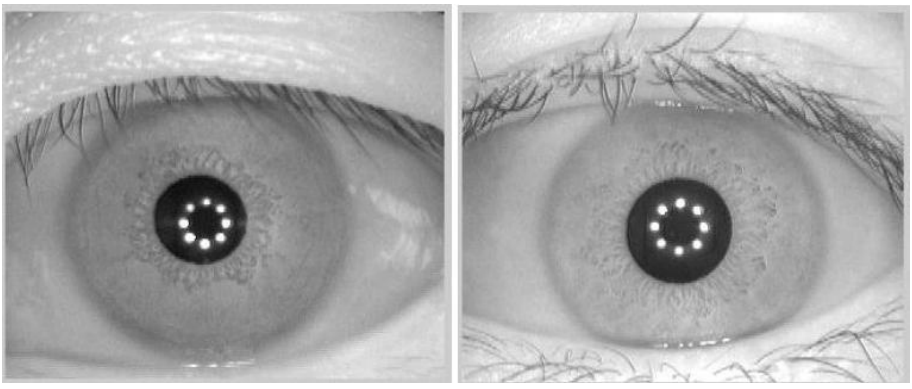


Figura 7. (Izq) Usuario $\mathcal{A}$, (Dch) Usuario $\mathcal{B}$

Además, en la Figura 8, se aprecia primeramente en el lado izquierdo el patrón biométrico de $\mathcal{A}$, es decir, $\mathcal{T}_{\mathcal{A}}$. En el centro, se aprecian diferentes ejecuciones del algoritmo genético, que se adapta perfectamente al patrón biométrico requerido. El algoritmo está diseñado de tal manera que $\mathcal{T}_{\mathcal{B}}$, puede ser tan parecido a $\mathcal{T}_{\mathcal{A}}$, como se desee. Incluso cabe la posibilidad de que sean iguales.

Sin embargo, en contra de parecer una buena idea, hacerlos iguales sería un error. Como puede apreciarse finalmente en la imagen de la derecha, para un mismo usuario existe una gran variación en el patrón biométrico para diferentes muestras. Por lo tanto, el algoritmo preve esta variabilidad, y una vez obtenida la solución, se distorsiona para que no parezca una copia exacta del patrón a falsificar.

Finalmente, en la Figura 9 puede apreciarse el resultado de la falsificación, y cómo la imagen del usuario $\mathcal{B}$ ha sido alterada de tal forma que ahora su patrón, $\mathcal{T}_{\mathcal{B}}$, es lo suficientemente similar a $\mathcal{T}_{\mathcal{A}}$, es decir el patrón de $\mathcal{A}$, como para engañar al sistema de identificación biométrica.

El sistema es incapaz de distinguir un usuario del otro, y por lo tanto, el usuario $\mathcal{B}$ ha conseguido acceder al sistema.

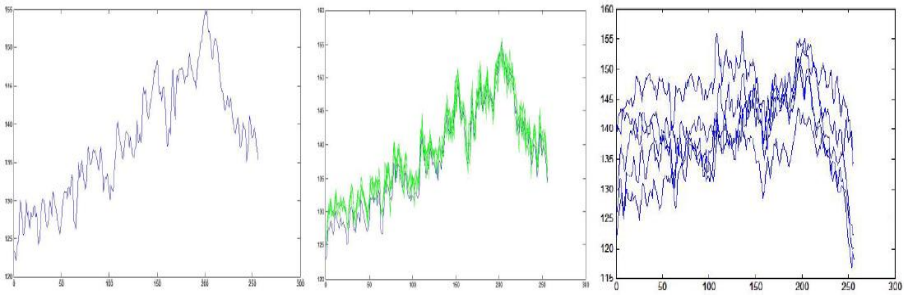


Figura 8. $\mathcal{T}_{\mathcal{A}}$, $\mathcal{T}_{\mathcal{A}}$ y resultados del algoritmo genético, $\mathcal{T}_{\mathcal{A}}$ para diferentes muestras de un mismo usuario $\mathcal{A}$.

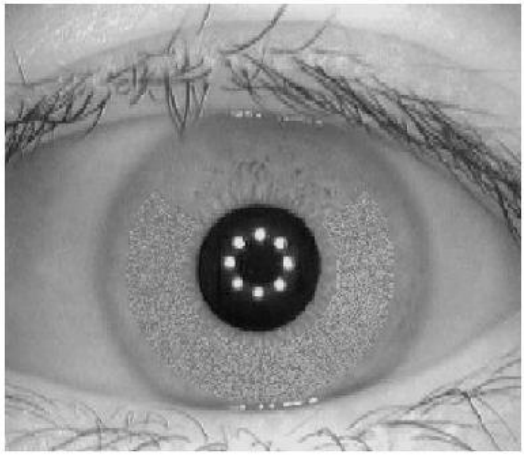


Figura 9. Resultado final del algoritmo: $\mathcal{B}$ cuyo patrón biométrico es $\mathcal{T}_{\mathcal{A}}$.

5. Conclusiones y Trabajo Futuro

En este trabajo se han presentado varios algoritmos tanto para la mejora de la detección de iris, y de pupila, como un posible ataque para que un determinado usuario acceda al sistema sin necesidad de pertenecer a dicho sistema.

Respecto a los algoritmos de segmentación, se propone un nuevo preprocesado basado íntegramente en morfología matemática y distribuciones matemáticas, que hacen más precisa la detección de iris y de pupila. La rapidez con la que estos algoritmos llevan a cabo sus tareas aventajan en gran medida a algunos algoritmos actuales, pues los algoritmos aquí presentados aún no han sido optimizados ni implementados en un dispositivo específico, es decir, fuera de un PC.

Respecto al ataque al sistema biométrico, sería bueno camuflar de alguna manera la modificación hecha al usuario que quiere falsificar la entrada, pues a primera vista se ve como ese iris es un tanto diferente de uno normal, aunque el sistema de reconocimiento sea incapaz de diferenciarlos. Se podrían poner más restricciones al algoritmo genético para que pudiera darle un aspecto más parecido al de un iris humano, o usar autómatas celulares para obtener dicho requisito.

Sin embargo, los avances alcanzados tanto en segmentación como en el ataque al sistema de iris (el primero de este tipo que se ha hecho en la literatura), son bastante buenos y sobre todo prometedores.

6. Agradecimientos

Los autores quieren agradecer al proyecto CENIT Segur@: Seguridad y Confianza en la Sociedad de la Información, financiado por el Ministerio de Industria, Turismo y Comercio.

Referencias

[1] T. Bäck, D. B. Fogel, Z. Michalewicz, Eds. *Evolutionary Computation 1: Basic Algorithms and Operators.*, Institute of Physics Publishing, Bristol, 2000.

[2] T. Bäck, D. B. Fogel, Z. Michalewicz, Eds. *Evolutionary Computation 2: Advanced Algorithms and Operators.*, Institute of Physics Publishing, Bristol, 2000.

[3] R. Capelli, A. Lumini, D. Maio and D. Maltoni, *'Can Fingerprints be reconstructed from ISO Templates?'*, in Proc. *International Conference on Control, Automation, Robotics and Vision (ICARCV2006)*, Singapure, December 2006.

[4] CASIA Iris Image Database. http://www.sinobiometrics.com

[5] J. Daugman, *High Confidence Visual Recognition of Persons by a Test of Statistical Independence*, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No 11, Nov. 1993.

[6] J. Daugman, *'How Iris Recognition Works'*, IEEE Transactions on Circuits and Systems For Video Technology, Vol. 14, n 1, January 2004.

[7] A. E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, Berlin, Germany: Springer, 2003.

[8] R. C. González, R. E. Woods, S. L. Eddins, *Digital Image Processing*, 2^{nd} Edn, Prentice All, 2004.
[9] Iris Challenge Evaluation http://iris.nist.gov/ICE/
[10] A. de Santos Sierra, C. Sánchez Ávila, E. Marchiori, *Iris Recognition: Segmentation enhancement by using Morphological Operators*, Master Final Tesis, June, 2007. Amsterdam.

Ataques directos usando imágenes falsas en verificación de iris

Virginia Ruiz-Albacete, Pedro Tome-Gonzalez, Fernando Alonso-Fernandez, Javier Galbally, Julian Fierrez, and Javier Ortega-Garcia

Biometric Recognition Group - ATVS
Escuela Politecnica Superior - Universidad Autonoma de Madrid
Avda. Francisco Tomas y Valiente, 11 - 28049 Madrid, España
http://atvs.ii.uam.es
{virginia.ruiz, pedro.tome, fernando.alonso, javier.galbally, julian.fierrez, javier.ortega}@uam.es

Resumen En este artículo se estudian las vulnerabilidades de un sistema de reconocimiento de iris frente ataques directos. Para ello se ha creado una base de datos con iris falsos, partiendo de los iris reales de la base de datos BioSec. Usando una impresora comercial, las imágenes de iris han sido impresas y posteriormente capturadas con nuestro sensor de iris. Para los experimentos usamos un sistema de reconocimiento de libre distribución. Basándonos en los resultados obtenidos tras las pruebas en distintos modos de operación, demostramos que el sistema es vulnerable frente ataques directos, resaltando la importancia de desarrollar contramedidas para este tipo de acciones fraudulentas.

Palabras Clave: Biometría, reconocimiento de iris, ataques directos, iris falso

1. Introducción

El importante aumento del número de aplicaciones que requieren una correcta identificación de individuos ha provocado un creciente interés en la biometría. El término *biometría* hace referencia al reconocimiento de forma automática de un individuo, basándose en sus rasgos físicos (por ejemplo, huellas dactilares, iris, geometría de la mano, orejas, huella palmar) o a características en su comportamiento (como la firma, forma de andar y forma de teclear) [1]. Los sistemas biométricos presentan varias ventajas frente a los métodos de seguridad tradicionales basados en algo que sabes (password, PIN), o algo que tienes (tarjeta, llave). En ellos no se necesita conocer una clave (que puede ser olvidada) ni requiere un instrumento (véase una llave que puede ser perdida o robada), sino que realiza un reconocimiento basado en lo que uno es. Entre todas las técnicas biométricas, el reconocimiento de iris ha sido tradicionalmente considerado como uno de los métodos más fiables y precisos de los disponibles [2]. Además, cuenta con la ventaja de ser un rasgo bastante estable a lo largo de la vida de una persona y permite una identificación no invasiva, ya que se trata de un órgano visible de forma externa [3].

Sin embargo, a pesar de todas estas ventajas, los sistemas biométricos presentan algunos inconvenientes, tales como [4]: *i*) la falta de privacidad (por ejemplo, todo el mundo conoce nuestra cara y puede obtener nuestra huella), o *ii*) el hecho de que un rasgo biométrico no pueda ser reemplazado (mientras que una clave olvidada puede ser fácilmente redefinida, una huella dactilar no puede ser regenerada si ha sido robada). Además, los sistemas biométricos son vulnerables a ataques externos, lo que puede reducer su nivel de seguridad. En [5] se identifican 8 puntos posibles de ataque a un sistema de reconocimiento biométrico. Estos puntos de vulnerabilidad son descritos en la Figura 1, y pueden ser agrupados, de forma general, en dos grupos principales:

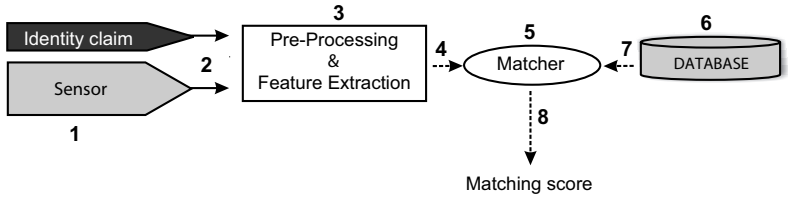


Figura 1. Arquitectura de un sistema automático de verificación biométrica. Los posibles puntos de ataque se han numerado del 1 al 8.

- **Ataques directos**. Este tipo de ataques se basa en el uso de un rasgo biométrico sintético, como por ejemplo un dedo de goma, con el que se intenta acceder al sensor (punto 1 de la Figura 1). Es importante destacar que para este tipo de ataques no es necesario tener ningún conocimiento especial sobre el sistema. De hecho, tiene lugar en el plano analógico, fuera de los límites digitales del sistema, por lo que mecanismos de protección digitales (como la firma digital, el watermarking, etc.) no pueden ser usados.

- **Ataques indirectos**. Este grupo incluye los 7 puntos de ataque restantes identificados en la Figura 1. Los ataques 3 y 5 se pueden llevar a cabo mediante un *troyano* que anule los módulos del sistema. En el ataque 6, se manipula la base de datos del sistema. El resto de ataques (2, 4, 7 y 8) representan posibles puntos débiles en los canales de comunicación del sistema. Al contrario que los ataques directos, en este caso el intruso debe conocer alguna información adicional sobre el funcionamiento interno del sistema, y en muchos casos tener un acceso físico a algún componente de la aplicación. La mayoría de los trabajos referentes a ataques indirectos usan algún tipo de técnica basada en hill-climbing, introducida en [6].

En este trabajo nos centramos en estudiar los ataques directos en sistemas basados en reconocimiento de iris. Para ello hemos creado una base de datos con imágenes sintéticas de iris generadas a partir de los 50 usuarios de la base de datos multimodal BioSec [7]. Este artículo se estructura de la siguiente manera. En la

Sec. 2 se detalla el proceso seguido para la creación de los iris falsos y presentamos la base de datos usada en los experimentos. El protocolo experimental, algunos resultados y comentarios se exponen en la Secc. 3. Finalmente exponemos las conclusiones en la Secc. 4.

IMPRESORA	PAPEL	PRE-PROCESADO [8]
Chorro de tinta	Papel blanco	Equalización de histograma
Láser	Papel reciclado	Filtrado de ruido
	Papel Fotográfico	Apertura/cierre
	Papel de alta calidad	Top hat
	Papel cebolla	
	Cartulina	

Cuadro 1. Pruebas realizadas para la creación de iris falsos.

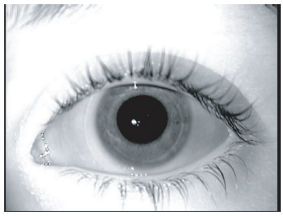
2. Base de Datos de Iris Falsos

Se ha creado para este trabajo una nueva base de datos a partir de imágenes de iris de 50 usuarios extraidos de la base de datos de referencia BioSec [7]. Se ha dividido el proceso en 3 pasos: *i*) Primero se han preprocesado las imágenes originales para obtener una mejor calidad en pasos posteriores, después *ii*) han sido impresas en un papel usando una impresora comercial y por último, *iii*) las imágenes impresas se han presentado al sensor de iris, obteniendo así las imágenes falsas.

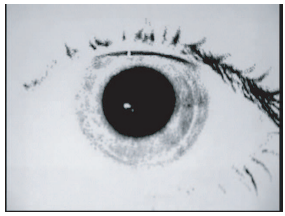
2.1. Método de generación automática de iris

Para lograr una nueva base de datos de forma correcta, es necesario tener en cuenta diversos factores que afectan a la calidad de las imágenes falsas adquiridas. Se han encontrado como principales variables, con una importancia significativa para la calidad de iris: el preprocesado de las imágenes originales, el tipo de impresora y el tipo de papel.

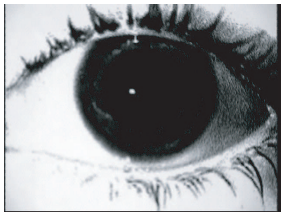
Hemos probado dos impresoras diferentes: una HP Deskjet 970cxi (de chorro de tinta) y una HP LaserJet 4200L (láser). Ambas proporcionan una calidad bastante buena. Por otra parte, hemos observado que la calidad de las imágenes adquiridas depende del tipo de papel usado. En este punto aparece la mayor variedad de opciones. Los tipos de papel probados se indican en la Tabla 1. En nuestros experimentos, el pre-procesado adquiere especial importancia ya que hemos observado que la cámara de iris no captura la mayoría de las imágenes originales impresas que no han sido previamente modificadas. Por ello hemos llevado a cabo diferentes métodos de mejora y refuerzo de las imágenes antes de ser impresas, de forma que sea posible adquirir imágenes falsas de buena calidad. Las opciones probadas se encuentran también resumidas en la Tabla 1. Probando todas las posibilidades con un pequeño grupo de imágenes, la combinación que



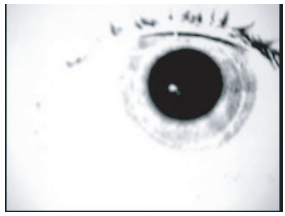
(a) Imagen original - sin mejora



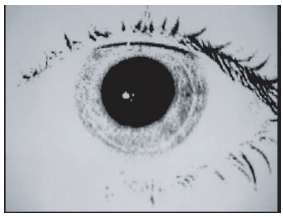
(b) Imagen falsa - sin mejora



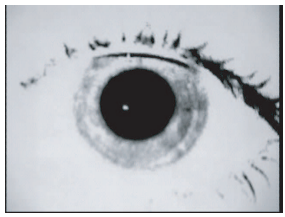
(c) Imagen falsa - Equalización de histograma



(d) Imagen falsa - filtrado de ruido



(e) Imagen falsa - TopHat



(f) Imagen falsa - Apertura+TopHat

Figura 2. Imágenes capturadas con las distintas modificaciones, usando papel de alta calidad e impresora de chorro de tinta.

ha resultado en una mejor segmentación y por tanto en la mejor calidad para la posterior comparación ha sido la impresora a chorro, con papel de alta calidad y un preprocesado de "Apertura-TopHat". En la Figura 2, se muestran ejemplos de diferentes técnicas de preprocesado con este tipo de papel y dicha impresora.

2.2. Base de Datos

La base de datos de iris falsos creada sigue la misma estructura que la base de datos original Biosec. Por tanto, los datos de los experimentos consisten en 50 usuarios $\times$ 2 ojos $\times$ 4 imágenes $\times$ 2 sesiones = 800 imágenes falsas, cada una de las cuales tiene su correspondiente imagen real. La adquisición de las imágenes falsas se ha realizado con la misma cámara que capturó las imágenes de BioSec, una LG IrisAccess EOU3000.

3. Experimentos

3.1. Sistema de reconocimiento

Para nuestros experimentos hemos usado el sistema desarrollado por Libor Masek[1] [9]. Consiste en la secuencia de pasos descritos a continuación: segmentación, normalización, codificación y comparación de plantillas.

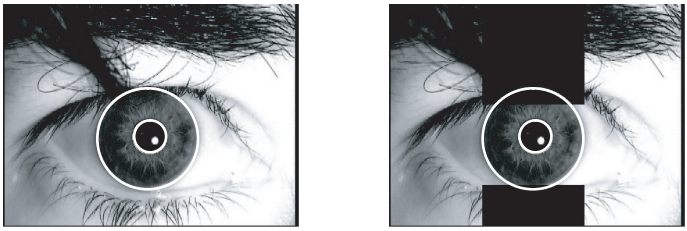
Para la segmentación del iris, el sistema utiliza la transformada circular de Hough para detectar las fronteras de la pupila y del iris. Las fronteras del iris se modelan como dos círculos. El sistema también realiza un paso de detección de párpados. Los párpados son detectados mediante una línea en la parte superior e inferior haciendo uso de la transformada lineal de Hough (ver la Figura 3(a) derecha, en la que la línea de los párpados corresponde al borde del bloque negro). La detección de pestañas se basa en una umbralización del histograma, y está implementado en el código, pero nosotros no lo utilizaremos para nuestros experimentos. A persa de que las pestañas son bastante oscuras comparadas con la región del iris, existen otras zonas del iris con el mismo tono oscuro debido a las condiciones de la imagen. Por ello, un corte basado en este umbral para aislar las pestañas resultaría también en la eliminación de otras partes importantes del iris. Sin embargo, para nuestra base de datos, la oculsión por pestañas no es demasiado prominente.

Para mejorar el funcionamiento de la segmentación, primero pre-estimamos el centroide de la pupila mediante umbralización del histograma, ya que se ha comprobado que la región de la pupila es la de menores niveles de gris de una imagen de iris. Esta pre-estimación nos permite reducir el área de búsqueda de la transformada circular de Hough. Además imponemos tres condiciones a los dos círculos que van a modelar las fronteras de pupila e iris: *i*) a pesar de que estos dos círculos no tienen porqué ser concéntricos, se impone un valor máximo a la distancia entre sus centros; *ii*) no se permite que ninguno de los dos círculos puede tener partes fuera de la imagen de iris; y *iii*) los radios de los círculos no pueden ser similares.

Para la normalización de la región del iris, se utiliza una técnica basada en el "modelo de goma"desarrollado por Daugman [10]. El centro de la pupila se considera el punto de referencia, y basándonos en él, generamos un vector de 2D según el mapeo del radio angular de la región de iris segmentada. En la Figura 3 se muestra un ejemplo de los pasos que se llevan a cabo esta normalización.

La extracción de características se implementa mediante convolución de la imagen de iris normalizada con un filtro Log-Gabor 1D. Las filas del patrón normalizado en 2D se toman como señales en 1D, cada fila correspondiente a un anillo circular de la región del iris. Usa la dirección angular puesto que la independencia máxima tiene lugar en esta dirección. La salida del filtrado se cuantifica en fase en cuatro niveles siguiendo el método de Daugman [10], con cada filtro, produciendo dos bits de datos. La salida de la cuantificación en fase es un código de grises, de modo que al desplazarse de un cuadrante a otro, sólo

[1] El código puede descargarse de forma gratuita en `www.csse.uwa.edu.au/~pk/studentprojects/libor/sourcecode.html`



(a) Imagen original e imagen de ruido

(b) Plantilla normalizada de iris

(c) Máscara de ruido

Figura 3. Ejemplos de los pasos de la normalización.

hay 1 bit cambio. Esto reducirá al mínimo el número de bits en desacuerdo para imágenes del mismo iris ligeramente desalineadas [9]. El proceso de codificación produce una plantilla binaria con un número de bits de información, y la correspondiente máscara de ruido que representa las zonas corruptas (párpados) dentro de los patrones de iris (ver Figura 3 (c)).

Para la comparación entre plantillas (matching), la métrica elegida para el reconocimiento es la distancia de Hamming (HD). La distancia de Hamming empleada incorpora el enmascaramiento de ruido, de modo que sólo los bits sin ruido se utilizan en el cálculo de la misma. La fórmula de la distancia de Hamming modificada viene dada por:

$$HD = \frac{1}{N - \sum_{k=1}^{N} Xn_k(OR)Yn_k} \cdot \sum_{j=1}^{N} X_j(XOR)Y_j(AND)Xn'_j(AND)Yn'_j$$

donde X_j y Y_j son los dos bits a comparar de la plantilla, Xn_j y Yn_j son las correspondientes mascaras de ruido para X_j y Y_j, y N es el numero de bits representado por cada plantilla.

Con el fin de dar cuenta de las incoherencias rotacionales, cuando se calcula la distancia de Hamming de dos modelos, una plantilla se desplaza a nivel de bit a izquierda y derecha, calculándose una serie de valores de la distancia de Hamming a partir de sucesivos cambios [10]. Este método corrige desajustes rotacionales en el modelo normalizado del iris causados por diferencias en rotación de imágenes. De entre los valores de distancia calculada se adopta el menor valor calculado.

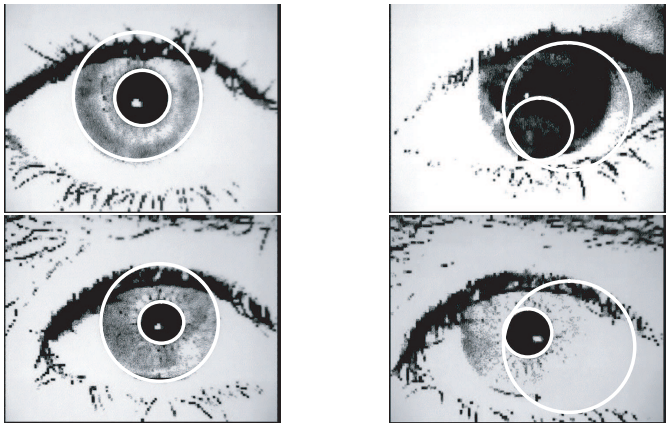
3.2. Protocolo Experimental

Para los experimentos, cada ojo de la base de datos se considera un usuario diferente. De esta forma, tenemos dos sesiones con 4 imágenes cada una de los 100 usuarios (50 participantes $\times$ 2 ojos por participante).

En los experimentos se consideran dos diferentes escenarios de ataque y se comparan al modo de funcionamiento normal:

- **Modo normal de funcionamiento (NOM)**: en este modo, el registro y las pruebas se llevan a cabo con iris reales. Este será el escenario de referencia. En este contexto, la Tasa de Falsa Aceptación (FAR) del sistema se define como el número de veces que un impostor, usando su propio iris, consigue acceder al sistema como un usuario original, por lo que se puede interpretar como la robustez del sistema frente a un ataque sin esfuerzo. Del mismo modo, la Tasa de Falso Rechazo (FRR) denota el número de veces que un usuario original es rechazado por el sistema.
- **Ataque 1**: aquí, tanto el registro como las pruebas se llevan a cabo con iris falsos. En este caso, el atacante se registra en el sistema con un iris falso correspondiente a un usuario genuino y luego intenta entrar a la aplicación usando también un iris falso del mismo usuario. En este escenario, un ataque no exitoso (es decir, el sistema rechaza al atacante) será cuando el impostor no sea capaz de acceder al sistema usando el iris falso. Por tanto, la tasa de éxito del ataque (SR) de este esceracio se puede calcular como: SR = 1−FRR.
- **Ataque 2**: el registro se lleva a cabo con un iris real y las pruebas se realizan con un iris falso. En este caso el usuario genuino se registra con su iris y el atacante intenta acceder a la aplicación con el iris falso correspondiente al usuario legal. Un ataque exitoso tendrá lugar si el sistema confunde un iris falso con su correspondiente genuino, es decir: SR = FAR.

Para calcular el rendimiento del sistema en un modo normal de funcionamiento, el protocolo experimental ha sido el siguiente. Para un usuario dado, todas las imágenes de la primera sesión se consideran como muestras de registro. Las comparaciones genuinas se obtienen comparando las muestras de registro con las imágenes correspondientes de la segunda sesión del mismo usuario. Las comparaciones de impostores se obtienen comparando una imagen de registro con una muestra aleatoria de la segunda sesión de cada uno de los usuarios restantes. De forma similar, para calcular el FRR en el ataque 1, todas las imágenes falsas de la primera sesión de cada usuario son comparadas con las correspondientes imágenes falsas de la segunda sesión. En el segundo ataque, solo se calculan los resultados de los impostores, comparando las 4 muestras originales de registro de cada usuario con sus 4 muestras falsas de la segunda sesión. En nuestros experimentos, no todas las imágenes fueron segmentadas correctamente por el sistema de reconocimiento Por ello no fue posible usar todas las imágenes de ojos para los experimentos de prueba.



(a) Detección correcta de iris (b) Detección incorrecta de iris

Figura 4. Ejemplos de las imágenes falsas correctamente segmentadas (izquierda) y con una detección incorrecta de irir (derecha).

3.3. Resultados

En la Figura 4, se muestran varios ejemplos de imágenes falsas con detección correcta e incorrecta de iris. El número de imágenes correctamente segmentadas para la base de datos original es de 792 (99 % de las 800 disponibles) y 574 para la base de datos falsa (71.75 % de las 800). Es importante resaltar que más del 70 % de las imágenes falsas pasan con éxito las fases de segmentación y normalización. Gracias a las modificaciones incluidas en el paso de segmentación (ver Sección 3.1), hemos mejorado el porcentaje de segmentación del sistema original, el cual se situaba en anteriores experimentos en un 80.56 % y un 38.43 % para la base de datos original y falsa respectivamente. Es importante destacar también que al intentar mejorar el porcentaje de correcta segmentación de las imágenes reales, estamos también mejorando el de las imágenes falsas. En la Tabla 2 se muestra el Porcentaje de éxito (SR) de los ataques directos al sistema de reconocimiento en cuatro puntos de funcionamiento distintos, considerando únicamente la comparación entre imágenes correctamente segmentadas. El umbral de decisión está fijado para alcanzar un FAR=$\{0.1, 1, 2, 5\}$ % en el modo normal de funcionamiento, y después se calcula el porcentaje de éxito de los dos ataques propuestos. Como podemos observar, para cualquier punto de funcionamiento el sistema es vulnerable para ambos ataques (de hecho se observa una tasa de éxito mayor ó igual al 35 %). Esto se hace especialmente evidente según aumentamos la FAR del modo normal de funcionamiento, consiguiendo un éxito de ataque de más de la mitad de las pruebas. También es importante resaltar que el porcentaje de éxito del ataque 1 es similar al del ataque 2. En el ataque 1, un intruso sería registrado correctamente en el sistema usando un iris falso de otra persona y posteriormente se le permitiría el acceso al sistema usando dicha imagen falsa.

NOM	Ataque 1	Ataque 2
FAR - FRR (%)	SR (%)	SR (%)
0.1 - 16.84	33.57	36.89
1 - 12.37	48.02	52.44
2 - 10.78	53.03	56.96
5 - 8.87	61.19	64.56

Cuadro 2. Evaluación del sistema de verificación frente ataques directos. NOM se refiere al modo normal de funcionamiento del sistema y SR al porcentaje de éxito del ataque.

4. Conclusiones

Se ha presentado un estudio de las vulnerabilidades de un sistema basado en reconocimiento de iris. Los ataques se han realizado usando imágenes falsas creadas a partir de imágenes reales de la base de datos de referencia BioSec. Después de imprimir las imágenes con una impresora comercial, estas fueron presentadas al sensor de iris. Se han estudiado diferentes factores que afectan a la calidad de las imágenes falsas adquiridas, incluyendo el pre-procesado de las imágenes originales, el tipo de impresora y el tipo de papel. Hemos elegido la combinación que nos da la mejor calidad y después hemos construido una base de datos falsa con las imágenes de 100 ojos, con 8 imágenes de iris por ojo. La adquisición de las imágenes falsas se ha llevado a cabo con las misma cámara usada en BioSec.

Se han comparado dos escenarios de ataque distintos con el modo normal de funcionamiento del sistema, usando un sistema de reconocimiento de disponibilidad pública. El primer ataque consiste en registrarse y acceder al sistema con un iris falso. El segundo simula el registro con un iris original y el acceso con un iris falso. Los resultados mostraron que el sistema es vulnerable para ambos ataques. También se ha observado que alrededor del 72 % de las imágenes falsas fueron segmentadas correctamente por el sistema. Cuando esto ocurre, al intruso se le garantiza la entrada con una probabilidad bastante alta, alcanzando un porcentaje de éxito en ambos ataques del 50 % o más.

Una posible contramedida para prevenir estos ataques es usar procedimientos de detección de vida. Para el caso de sistemas de reconocimiento de iris, se propone la detección de reflexiones de luz, la detección de movimiento del iris, o la respuesta del iris a modificaciones repentinas de luz [11,12]. Estas líneas de investigación serán seguidas en nuestros trabajos futuros. También se explorará el uso de otro tipo de sensores, como el de soporte manual de OKI usado en la base de datos de Casia[2].

Agradecimientos. Este trabajo ha sido financiado por el proyecto TEC2006-13141-C03-03 del Ministerio de Educación y Ciencia y por la Red de Excelencia Europea

[2] http://www.cbsr.ia.ac.cn/databases.htm

BioSecure IST-2002-507634. El autor F. A.-F. agradece a la Consejeria de Educacion de la Comunidad de Madrid y al Fondo Social Europeo por financiar sus estudios de Doctorado. El autor J. G. está siendo fianciado por una beca FPU del Ministerio de Educación y Ciencia. El autor J. F. está siendo fianciado por una Marie Curie Fellowship de la Comisión Europea.

Referencias

1. Jain, A., Ross, A., Pankanti, S.: Biometrics: A tool for information security. IEEE Trans. on Information Forensics and Security **1** (2006) 125–143
2. Jain, A., Bolle, R., Pankanti, S., eds.: Biometrics - Personal Identification in Networked Society. Kluwer Academic Publishers (1999)
3. Monro, D., Rakshit, S., Zhang, D.: DCT-Based iris recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **29**(4) (April 2007) 586–595
4. Schneier, B.: The uses and abuses of biometrics. Communications of the ACM **48** (1999) 136
5. Ratha, N., Connell, J., Bolle, R.: An analysis of minutiae matching strength. Proc. International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA **Springer LNCS-2091** (2001) 223–228
6. Soutar, C., Gilroy, R., Stoianov, A.: Biometric system performance and security. Proc IEEE Workshop on Automatic Identification Advanced Technologies, AIAT (1999)
7. Fierrez, J., Ortega-Garcia, J., Torre-Toledano, D., Gonzalez-Rodriguez, J.: BioSec baseline corpus: A multimodal biometric database. Pattern Recognition **40**(4) (April 2007) 1389–1392
8. Gonzalez, R., Woods, R.: Digital Image Processing. Addison-Wesley (2002)
9. Masek, L., Kovesi, P.: Matlab source code for a biometric identification system based on iris patterns. The School of Computer Science and Software Engineering, The University of Western Australia (2003)
10. Daugman, J.: How iris recognition works. IEEE Transactions on Circuits and Systems for Video Technology **14** (2004) 21–30
11. Daugman, J.: Anti spoofing liveness detection. available on line at http://www.cl.cam.ac.uk/users/jgd1000/countermeasures.pdf
12. Pacut, A., Czajka, A.: Aliveness detection for iris biometrics. Proc. IEEE Intl. Carnahan Conf. on Security Technology, ICCST (2006) 122–129

Verificación de Locutores mediante Modelos de Mezclas Gaussianas (GMM)

Diego Carrero[1], Luis Puente[1], Belén Ruiz[1] y Mª Jesús Poza[2]

[1] Universidad Carlos III de Madrid. Avda. de la Universidad, 30. Leganés, Madrid (Spain)
dcarrero@di.uc3m.es, lpuente@it.uc3m.es, bruiz@inf.uc3m.es

[2] Universidad Francisco de Vitoria. Ctra. Pozuelo-Majadahonda Km. 1.800. Pozuelo de Alarcón, Madrid (Spain)
mj.poza.prof@ufv.es

Abstract. Los modelos de mezclas Gaussianas son uno de los métodos más utilizados en la verificación de locutores en los últimos años y actualmente es una de las tecnologías más maduras en el campo del reconocimiento de locutores. De la misma forma, la conversión de la señal de voz en parámetros de tipo Mel-Cepstrum ha demostrado ser una de las mejores transformaciones tanto para identificación o verificación de locutores como para tareas de reconocimiento de habla. A lo largo del presente trabajo se estudia, a partir de los coeficientes Mel-Cepstrum y la energía de la señal, el comportamiento de los modelos de mezclas Gaussianas ante diferentes configuraciones del vector de características.

Keywords: Biometría, Verificación de locutores, GMM, Máxima Verosimilitud, Algoritmo E-M.

1 Introducción

En función de la aplicación, el área principal del reconocimiento de locutores se divide en dos tareas específicas: identificación y verificación. En la identificación de locutores, el objetivo es determinar qué persona dentro de un conjunto cerrado se adapta mejor con una muestra de voz. En la verificación de locutores, el objetivo es determinar si una persona es quien dice ser a partir de una muestra de su voz.

En el reconocimiento de voz se pueden diferenciar diferentes tipos de sistemas en función de las limitaciones impuestas tanto en la voz como en el entorno del locutor a la hora de realizar las tareas de entrenamiento y prueba [2]. En un sistema dependiente del texto, el habla utilizada para entrenar y probar el sistema está limitada a ser la misma palabra o frase (o un conjunto muy reducido de expresiones). En cambio, en un sistema independiente del texto la locución empleada para el entrenamiento y las pruebas no está sometida a ninguna restricción.

2 Base Matemática

La base del sistema de verificación bajo estudio es el uso de Modelos de Mezclas Gaussianas (GMM) para la representación de los locutores [2]. La distribución de los vectores de características extraídos del habla de una persona se puede modelar mediante una mezcla de funciones de densidad de probabilidad Gaussianas. Para un vector de características $\underline{x}$, la función de densidad para el locutor s se define como:

$$p(\underline{x}|\lambda_s) = \sum_{k=1}^{K} \omega_s^{(k)} b_s^{(k)}(\underline{x}). \quad (1)$$

La función de densidad anterior es una combinación lineal de K componentes Gaussianas $b_s^{(k)}(\underline{x})$ ponderadas mediante unos pesos de mezcla $\omega_s^{(k)}$. Cada una de las componentes Gaussianas está caracterizada por un vector de medias $\underline{\mu}_s^{(k)}$ y una matriz de covarianzas $\Sigma_s^{(k)}$.

$$b_s^{(k)}(\underline{x}) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_s^{(k)}\right|^{1/2}} exp\left\{-\frac{1}{2}\left(\underline{x}-\underline{\mu}_s^{(k)}\right)^T \left(\Sigma_s^{(k)}\right)^{-1}\left(\underline{x}-\underline{\mu}_s^{(k)}\right)\right\}. \quad (2)$$

Los pesos de mezcla $\omega_s^{(k)}$ además han de satisfacer la siguiente condición:

$$\sum_{k=1}^{K} \omega_s^{(k)} = 1. \quad (3)$$

De forma general, los parámetros de la función de densidad del modelo para el locutor s se denotan como:

$$\lambda_s = \left\{\omega_s^{(k)}, \underline{\mu}_s^{(k)}, \Sigma_s^{(k)}\right\}_{k=1}^{K} \quad (4)$$

Un GMM puede tener diferentes formas en función de la elección de las matrices de covarianza. Por un lado, el modelo puede tener una matriz de covarianza para cada componente Gaussiana (nodal covariance), la misma matriz para todas las componentes (grand covariance) o puede existir una única matriz de covarianza para todos los modelos (global covariance). Y por otro lado, la matriz de covarianza puede ser tanto completa como diagonal [7].

2.1 Estimación de los Parámetros por Máxima Verosimilitud

El objetivo del entrenamiento del modelo λ_s del locutor s es estimar los parámetros del GMM que mejor ajustan la distribución a los vectores de características de entrenamiento. Uno de los métodos más extendidos es la estimación por máxima verosimilitud (ML) [5]. El objetivo de la estimación ML es encontrar los parámetros de un modelo que maximizan la verosimilitud del GMM para los datos de entrenamiento.

Supuesto que se parte de una secuencia de T vectores de entrenamiento tomados de forma independiente:

$$X = \left\{\underline{x}^{(t)}\right\}_{t=1}^{T} \quad (5)$$

La verosimilitud del GMM para el conjunto de vectores se puede calcular a partir de la siguiente expresión:

$$p(X|\lambda_s) = \prod_{k=t}^{T} p\left(\underline{x}^{(t)}|\lambda_s\right). \quad (6)$$

La expresión 6 es una función no lineal de los parámetros del modelo y no es posible su maximización de forma directa. Sin embargo, se puede obtener la estimación de los parámetros del modelo utilizando el algoritmo de Expectación-Maximización (EM). La idea básica del algoritmo EM [3] [5] [6] es, dado un modelo inicial λ, estimar un nuevo modelo λ' de tal forma que $p(X|\lambda') \geq p(X|\lambda)$. Este nuevo modelo estimado se convierte en el modelo inicial en la siguiente iteración y el proceso se repite hasta que se alcanza un determinado umbral de convergencia. En cada iteración del algoritmo EM, se emplean las siguientes expresiones para re-estimar los parámetros del modelo [2].

1. Estimación (Paso E): En el paso de estimación se calcula la probabilidad a posteriori de la característica acústica i mediante la siguiente expresión.

$$p\left(i|\underline{x}^{(t)}, \lambda_s\right) = \frac{\omega_s^{(i)} b^{(i)}\left(\underline{x}^{(t)}\right)}{\sum_{k=1}^{K} \omega_s^{(k)} b^{(k)}\left(\underline{x}^{(t)}\right)}, con\ 1 \leq i \leq K. \quad (7)$$

2. Maximización (Paso M): Una vez calculada la probabilidad a posteriori, se actualizan los parámetros de la distribución conforme a las siguientes expresiones. En el paso M se obtienen los parámetros que maximizan la verosimilitud del modelo dados los datos de entrenamiento.

– Coeficientes de mezcla

$$\omega_s'^{(i)} = \frac{1}{T}\sum_{t=1}^{T} p\left(i|\underline{x}^{(t)}, \lambda_s\right),\ para\ 1 \leq i \leq K. \quad (8)$$

– Vectores de medias

$$\underline{\mu}_s'^{(i)} = \frac{\sum_{t=1}^{T} p\left(i|\underline{x}^{(t)}, \lambda_s\right)\underline{x}^{(t)}}{\sum_{t=1}^{T} p\left(i|\underline{x}^{(t)}, \lambda_s\right)}, \qquad para\ 1 \leq i \leq K. \quad (9)$$

– Matrices de covarianza

$$\Sigma_s'^{(\mathrm{i})} = \frac{\sum_{t=1}^{T} p\left(i|\underline{x}^{(t)}, \lambda_s\right)\left(\underline{x}^{(t)} - \underline{\mu}_s'^{(i)}\right)\left(\underline{x}^{(t)} - \underline{\mu}_s'^{(i)}\right)^{\mathrm{T}}}{\sum_{t=1}^{T} p\left(i|\underline{x}^{(t)}, \lambda_s\right)}, para \begin{cases} 1 \leq t \leq T \\ 1 \leq \mathrm{i} \leq K \end{cases} \quad (10)$$

2.2 Interpretación del Modelo

Existen dos motivaciones para utilizar GMM como representación de la identidad de un locutor [2].

La primera motivación es la posibilidad que ofrecen las componentes individuales del GMM para modelar diferentes conjuntos de características acústicas. El espacio acústico correspondiente a la voz de un hablante puede ser caracterizado por un conjunto de características acústicas que modelan diferentes eventos fonéticos. La forma del espectro de la i-ésima característica acústica puede ser representada

mediante el vector de medias de la función de densidad y las variaciones sobre esta forma espectral media pueden ser representadas mediante la matriz de covarianza.

La segunda motivación para utilizar GMM es la capacidad que posee para representar diferentes tipos de distribuciones. Uno de las principales características de GMM es su capacidad para conseguir buenas aproximaciones para funciones de densidad de probabilidad arbitrarias.

2.3 Verificación de la Identidad del Locutor

A partir de un fragmento de voz, el sistema ha de decidir si la voz de entrada se ajusta al modelo pretendido por el usuario donante o si por el contrario, procede de un usuario impostor. Para un fragmento de voz de un usuario donante y una identidad pretendida, la decisión se ajusta a las siguientes hipótesis:

– H_1: El fragmento de voz pertenece al usuario pretendido.
– H_0: El fragmento de voz no pertenece al usuario pretendido.

La verificación de un locutor consiste en aplicar un test de cociente de verosimilitudes (LRT) [5] a un fragmento de voz para determinar si el usuario donante es aceptado o rechazado. Para un fragmento de voz representado por el conjunto X de los vectores de características extraídos de él y la identidad de un usuario pretendido con el correspondiente modelo λ_C, el LRT viene dado por:

$$\frac{Pr(H_1|X)}{Pr(H_0|X)} \underset{D_0}{\overset{D_1}{\gtrless}} \theta \;\rightarrow\; \frac{Pr(\lambda_C|X)}{Pr(\lambda_{\overline{C}}|X)} \underset{D_0}{\overset{D_1}{\gtrless}} \theta. \qquad \textbf{(11)}$$

La expresión 11 se puede modificar mediante la regla de Bayes y la aplicación de logaritmos para obtener una expresión equivalente:

$$\Lambda(X) = log\big(p(X|\lambda_C)\big) - log\left(p\big(X|\lambda_{\overline{C}}\big)\right). \qquad \textbf{(12)}$$

El término $p(X|\lambda_C)$ corresponde con la verosimilitud del fragmento de voz supuesto que proviene del locutor pretendido y el término $p\big(X|\lambda_{\overline{C}}\big)$ hace referencia a la verosimilitud del mismo fragmento supuesto que no proviene del locutor pretendido. Para determinar la autenticidad del locutor donante se compara el LRT con un umbral θ y en base al resultado del test se acepta o se rechaza la coincidencia de los usuarios donante y pretendido: si $\Lambda(X) > \theta$ se acepta al usuario donante y si $\Lambda(X) < \theta$ se rechaza.

La verosimilitud del fragmento de voz dado el modelo del locutor pretendido se calcula directamente a partir de la siguiente expresión:

$$log\big(p(X|\lambda_C)\big) = \frac{1}{T}\sum_{t=1}^{T} log\left(p\big(\underline{x}^{(t)}|\lambda_C\big)\right). \qquad \textbf{(13)}$$

Por su parte, la verosimilitud del fragmento de voz dado que no pertenece al locutor pretendido se obtiene a partir de una colección de modelos de locutores. Mediante un conjunto de B modelos de locutores de fondo $\{\lambda_1, \ldots, \lambda_B\}$, el segundo término de la parte derecha de la expresión 12 se puede obtener mediante la siguiente expresión:

$$log\left(p(X|\lambda_{\overline{C}})\right) = log\left(\frac{1}{B}\sum_{b=1}^{B} p(X|\lambda_b)\right). \quad (14)$$

Donde $p(X|\lambda_b)$ se calcula tal y como muestra la expresión 13. Esta expresión es la función de densidad de probabilidad conjunta de la expresión supuesto que procede de uno de los locutores de fondo salvo por el escalado $\frac{1}{T}$.

3 Descripción del Sistema de Clasificación

El sistema de clasificación se ha implementado mediante una serie de scripts MATLAB. Para cada locutor, el sistema genera un modelo en la etapa de entrenamiento. Así mismo, existe un modelo en el sistema para modelar a los locutores impostores, es decir, un modelo de mundo. Este modelo se obtiene a partir de las locuciones de diferentes individuos.

Para la obtención de los modelos de los usuarios se ha empleado el algoritmo EM con un límite máximo de 10 iteraciones en la obtención de los parámetros y una diferencia en la verosimilitud del 1% entre iteraciones como criterios de parada. En la inicialización del algoritmo EM se ha realizado un agrupamiento de los datos mediante el algoritmo k-means [11] [12] para obtener los parámetros iniciales del modelo. Para todos los usuarios, incluido el modelo de los locutores de fondo, se han empleado 10 Gaussianas en los modelos y se han utilizado matrices de covarianza diagonales por la reducción del coste computacional.

3.1 Descripción de los Datos Biométricos

En la realización de los experimentos se han empleado datos de voz procedentes de la base de datos BANCA [1]. BANCA está compuesta por un total de 260 locutores divididos en grupos de 52 locutores (26 masculinos y 26 femeninos) por idioma: inglés, francés, alemán, italiano y español. Cada uno de los sexos está grabado en tres ambientes diferentes (controlado, degradado y adverso). En la realización de los experimentos se han empleado las locuciones en inglés.

3.2 Descripción del Módulo Extractor de Características

La extracción de características se ha realizado mediante la herramienta HTK [13] para obtener la energía y los coeficientes Mel-Cepstrum, así como sus velocidades [14]. En la extracción de características se ha eliminado la componente continua de la señal de voz, se ha aplicado un enventanado de Hamming y se ha aplicado un filtro de preénfasis con coeficiente 0.97. Al final del proceso se han generado tramas de 25 ms. de longitud desplazadas cada 10 ms.

Para la verificación de la identidad de los locutores se proponen diferentes configuraciones sobre los vectores de características:

- Configuración 12+12+1+1: La línea base de experimentación se basa en la utilización de 12 coeficientes Mel-Cepstrum y sus correspondientes coeficientes diferenciales junto con la energía y su coeficiente de variación.
- Configuración 12+12+1: La primera variación sobre la línea base consiste en eliminar el diferencial de la energía.
- Configuración 12+12: La segunda variante de la línea base consiste en la clasificación de los usuarios empleando solamente los coeficientes Mel-Cepstrum y sus diferenciales.
- Configuración 12+0: La última propuesta a estudio se asienta en la clasificación de los usuarios mediante los coeficientes Mel-Cepstrum.

4 Metodología de Experimentación

El sistema se ha evaluado de forma independiente para cada sexo en cada uno de los tres entornos. En cada experimento, se han clasificado un total de 208 locuciones: 104 genuinas y 104 falsificadas.

En la fase de entrenamiento se han empleado 4000 vectores de entrenamiento para la obtención de cada modelo de usuario. El modelo de los locutores de fondo se ha obtenido a partir de las muestras de 10 usuarios 9000 de BANCA, también con 4000 vectores.

En la etapa de test el número de vectores empleado para cada usuario es variable y se sitúa entre 700 y 2000 aproximadamente. En esta etapa, cada locutor se presenta ante el sistema un total de cuatro veces: dos como usuario genuino y dos como usuario impostor.

Se proponen tres escenarios diferentes de trabajo sobre los cuales examinar las configuraciones antes citadas:

- Escenario 1: Clasificación de los locutores sin la sustracción de las medias de los coeficientes Mel-Cepstrum.
- Escenario 2: Clasificación de los locutores con la sustracción de las medias de los coeficientes Mel-Cepstrum.
- Escenario 3: Clasificación de los locutores eliminando los silencios de las tramas de voz junto a la sustracción de las medias de los coeficientes Mel-Cepstrum.

5 Resultados Experimentales

En los siguientes epígrafes se recogen los resultados obtenidos en los diferentes escenarios de experimentación. Para cada uno de los escenarios de clasificación se muestra una tabla que recoge el Equal Error Rate (EER) ofrecido por el sistema con las diferentes configuraciones para cada sexo y entorno.

5.1 Escenario 1: Sin la sustracción de las medias

A continuación se ofrecen los resultados de la clasificación de los usuarios sin la sustracción de las medias en los coeficientes Mel-Cepstrum.

Tabla 1. EER para cada uno de los entornos bajo las diferentes configuraciones.

	Configuración			
Entorno	12+12+1+1	12+12+1	12+12	12+0
M-CONTROLADO	3.846%	3.846%	1.923%	1.923%
M-DEGRADADO	7.692%	4.808%	1.923%	2.885%
M-ADVERSO	37.500%	39.423%	22.115%	28.846%
F-CONTROLADO	5.769%	4.808%	4.808%	6.731%
F- DEGRADADO	5.769%	5.769%	2.885%	3.846%
F-ADVERSO	42.308%	39.423%	26.923%	25.000%

Las mejores prestaciones se obtienen para la configuración 12+12 ya que alcanza los EER más bajos para cinco de las seis configuraciones. Por un lado, el empleo del diferencial de energía no sirve para mejorar la clasificación (configuración 12+12+1+1) y su eliminación reduce el EER. Por otro, la utilización de los diferenciales de los Mel-Cepstrum mejora la verificación de los locutores (configuración 12+12). Así mismo, la eliminación de la energía en el proceso de clasificación optimiza las prestaciones del sistema, sobre todo en entornos adversos.

En cuanto a sexos se refiere, el sistema se comporta ligeramente mejor frente a usuarios masculinos. Para éstos consigue un EER medio[1] del 13.061%, frente al 14.503% que ofrece para los locutores femeninos.

5.2 Escenario 2: Con la sustracción de las medias

Los resultados obtenidos al clasificar los usuarios sustrayendo la media de los coeficientes Mel-Cepstrum se recogen en la tabla 2.

Tal y como se desprende de la tabla 2, los mejores resultados se consiguen de nuevo al emplear la configuración 12+12. En este escenario, la configuración 12+12 es la que mejores prestaciones ofrece en cuatro de los seis entornos.

Al igual que ocurre en el escenario 1, la eliminación del coeficiente de energía mejora la clasificación de los usuarios debido a la no sustracción de los silencios en las tramas de voz. Esta mejora se aprecia sobre todo en el entorno adverso, donde se consiguen unas mejoras del 4.807% y del 9.616% en los sexos masculino y femenino respectivamente. El comportamiento del diferencial de energía es similar al escenario 1 y su utilización apenas contribuye a corregir el error. Por su parte, la eliminación de los diferenciales de los Mel-Cepstrum en la configuración 12+0 provoca un empeoramiento en las prestaciones del sistema.

[1] Para cada sexo, el EER medio se ha calculado sobre todas las configuraciones y entornos.

Tabla 2. EER para cada uno de los entornos bajo las diferentes configuraciones.

	Configuración			
Entorno	12+12+1+1	12+12+1	12+12	12+0
M-CONTROLADO	0.962%	1.923%	0.000%	1.923%
M-DEGRADADO	3.846%	3.846%	4.808%	2.885%
M-ADVERSO	9.615%	9.615%	4.808%	5.769%
F-CONTROLADO	2.885%	1.923%	2.885%	2.885%
F- DEGRADADO	4.808%	4.808%	3.846%	6.731%
F-ADVERSO	13.942%	15.385%	5.769%	7.692%

Así mismo, el comportamiento del sistema es mejor hacia los usuarios masculinos. El EER medio del sistema hacia los usuarios masculinos se sitúa en el 4.167% y hacia los usuarios femeninos éste se sitúa en el 6.1308%. En oposición al escenario 1, el EER se ha reducido considerablemente con la sustracción de la media de los coeficientes Mel-Cepstrum.

5.3 Escenario 3: Detección de silencios junto a la sustracción de las medias

Finalmente, la tabla 3 muestra los resultados obtenidos en la clasificación de los locutores bajo el escenario 3.

En este escenario no existe una configuración de parámetros dominante en cuanto a prestaciones se refiere ya que las configuraciones 12+12+1, 12+12 y 12+0 ofrecen los mejores resultados en cuatro de los seis entornos. Con la introducción de la eliminación de silencios, el comportamiento del sistema mejora hacia ambos tipos de locutores. Los EER medios en la clasificación de los usuarios son 3.285% y 3.926% para usuarios masculinos y femeninos. La eliminación de silencios no sólo optimiza la clasificación de los usuarios, también hace que el sistema se comporte de forma análoga hacia los dos tipos de locutores.

Tabla 3. EER para cada uno de los entornos bajo las diferentes configuraciones.

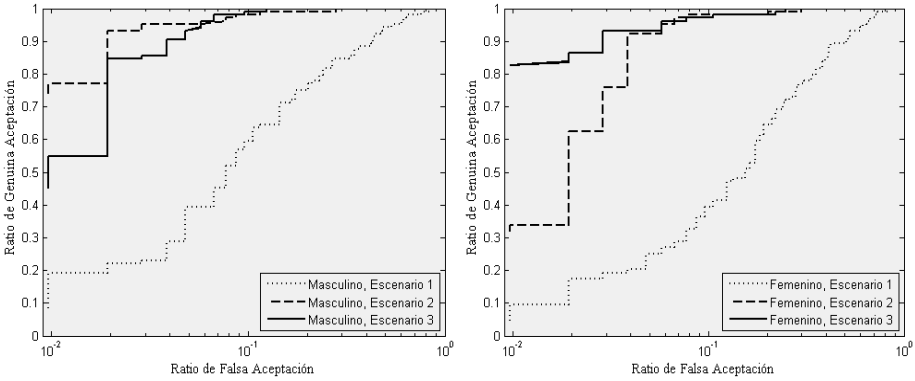
	Configuración			
Entorno	12+12+1+1	12+12+1	12+12	12+0
M-CONTROLADO	0.000%	0.962%	0.000%	1.923%
M-DEGRADADO	2.885%	1.923%	1.923%	1.923%
M-ADVERSO	8.654%	5.769%	5.769%	7.692%
F-CONTROLADO	1.923%	2.885%	2.885%	2.885%
F- DEGRADADO	2.885%	1.923%	4.808%	1.923%
F-ADVERSO	7.692%	6.731%	5.769%	4.808%

La figura 1 muestra la evolución de la curva ROC del clasificador en el entorno adverso y para la configuración 12+12 bajo cada uno de los escenarios. Tanto la sustracción de la media de los coeficientes como la eliminación de silencios mejoran las prestaciones del clasificador. La tabla 4 recoge los valores AUC asociados a las curvas ROC de la anterior figura.

Tabla 4. Valores AUC y EER asociados al clasificador en cada uno de los tres escenarios de clasificación propuestos. Las modificaciones introducidas en el proceso de clasificación mejoran considerablemente las prestaciones del clasificador.

	Escenario 1		Escenario 2		Escenario 3	
Locutor	AUC	EER	AUC	EER	AUC	EER
Masculino	0.85308	22.115%	0.98992	4.808%	0.98479	5.769%
Femenino	0.80015	26.923%	0.97549	5.769%	0.98936	5.769%

Fig. 1. Curvas ROC para la configuración 12+12 en los tres escenarios de experimentación. La figura de la izquierda recoge las curvas ROC para los locutores masculinos y la figura de la derecha lo hace para los femeninos. Se puede apreciar claramente el aumento del Área Bajo la Curva (AUC) desde el escenario 1 hasta el escenario 3.



6 Conclusiones

Una vez presentados los resultados de los experimentos, a continuación se recogen las conclusiones obtenidas.

- La sustracción de la media de los coeficientes Mel-Cepstrum mejora considerablemente el comportamiento del sistema de clasificación.
- La eliminación de silencios en la señal de entrada contribuye a reducir el EER del sistema de clasificación.
- La eliminación del coeficiente de energía en la clasificación mejora las prestaciones del sistema en entornos adversos y degradados. El comportamiento de

la energía se ve afectado por la no eliminación de silencios en la extracción de características.
- De las cuatro configuraciones propuestas, la configuración 12+12 es la que ofrece mejores resultados. Alcanza los menores ratios de error en los escenarios 1 y 2 y ofrece las mejores prestaciones en el escenario 3 para los locutores masculinos.

7 Agradecimientos

Este trabajo ha sido posible gracias a la financiación del Ministerio de Educación y Ciencia TEC2006-12365-C02-01.

8 Referencias

1. Porée, F., Mariéthoz, J., Bengio S. et al: The BANCA Database and Experimental Protocol for Speaker Verification. In: 4th Int. Conf. Audio and Video Based Biometric Person Authentication, AVBPA'03. LNCS, vol. 2688. pp. ns-ns. Springer-Verlag. (2003)
2. Reynolds, D.A.: Speaker Identification and Verification using Gaussian Mixture Speaker Models". In: Speech Communication, vol. 17. pp. 91-108. (1995)
3. Dempster, A.P., Laird, N. M. and Rubin, D. B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. In: Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, No. 1. pp. 1-38. (1977)
4. Reynolds, D.A.: An Overview of Automatic Speaker Recognition Technology. In: Proceedings of ICASSP, vol. IV. pp. 4072-4075. (2002)
5. Duda, R.O.: Pattern Classification. John Wiley & Sons, 2nd Ed. (2001)
6. Redner, R.A. and Walker, H.F.: Mixture Densities, Maximum Likelihood and the E-M Algorithm. In: SIAM Review, vol. 26, No. 2. pp. 195-239. (1984)
7. Reynolds, D.A. and Rose, R.: Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. In: IEEE Transactions on Speech and Audio Processing, vol. 3. pp. 72-83. (1994)
8. Tsui, P.: An Expectation Maximization Algorithm for Learning a Multi-Dimensional Gaussian Mixture. In: MATLAB Central. Disponible en: http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=8636&objectType=file
10. Leung, C.C. and Moon, Y.S.: Effect of Window Size and Shift Period in Mel-Warped Cepstral Feature Extraction on GMM-Based Speaker Verification. In: LNCS, vol. 2688. pp. 438-445. Springer. (2003)
11. Song, M. and Rajasekaran, S.: Fast k-Means Algorithms with Constant Approximation. In: ISAAC 2005. LNCS, vol. 3287. pp 1029-1038. Springer-Verlag. (2005)
12. Alsabti, K., Ranka, S. and Singh, V.: An Efficient k-Means Clustering Algorithm. Disponible en: http://www.cise.ufl.edu/~ranka/
13. HTK Speech Recognition Toolkit. Disponible en: http://htk.eng.cam.ac.uk/
14. Han, W,. et al.: An efficient MFCC extraction method in speech recognition. In: IEEE International Symposium on Circuits and Systems, 2006. ISCAS 2006. pp. 145-148 (2006)

Speaker's Gender Detection from Glottal Biometry

Pedro Gómez-Vilda, Roberto Fernández-Baillo, Agustín Álvarez-Marquina, Luis Miguel Mazaira-Fernández, Rafael Martínez-Olalla, Victoria Rodellar-Biarge

Grupo de Informática Aplicada al Tratamiento de Señal e Imagen, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, s/n, 28660 Madrid

e-mail: pedro@pino.datsi.fi.upm.es

Abstract. Through the present work a biometric signature of a speaker's voice is proposed for the detection of the speaker's gender. The estimation method relies on the extraction of the glottal flow derivative from voice after removing the vocal tract transfer function by inverse filtering. This spectral density is related to the vocal fold cover biomechanics, and it is well known that certain speaker's features as gender, age or pathologic condition are present in it. For such a database of 100 pathology-free speakers equally balanced in gender and age is used as an experimental framework to draft the results exposed in the work. As the estimated biometric parameters show a certain degree of cross-correlation Principal Component Analysis (PCA) is used to reduce parameter dimension. The principal components are used in unsupervised *k-means* clustering of speakers (unsupervised gender detection). The outcome grouping shows an almost complete separation of speakers by gender in terms of the most relevant parameters derived from a statistical dispersion study. Possible applications of the study can be found in forensic acoustics as well as in speaker identification and verification tasks.

Keywords: Voice Biometry, Speaker's Identification, Speaker Biometrical Characterization, Forensic Acoustics, Glottal Source

1. Introduction

The present work is oriented to voice characterization to determine a biometric signature of voice based on the parameterization of the glottal biomechanics for voice characterization (gender, age and speaker's pathological voice condition being the primary targets, among others). A comprehensive review of the characterization of voice may be found in [1]. Traditionally the characterization of the speaker has been oriented to gender and age as the main goals. Good studies have been published in this sense during the last two decades [2][3] [4][5][6]. These works show the way to establish a more structured study regarding voice characterization. On one side they point out to the use of time or frequency domain parameters as the basis of the study. On the other side, they deal with Vocal Tract or Glottal Source biometry. In the present approach the Glottal Source has been selected as the object of the research. A generalized signature is proposed on a full description of the Glottal Source spectral envelope, concentrating on the singularities appearing on this pattern (peaks and troughs). This generalization is based on the biomechanical foundations of the Glottal Source spectral

envelope [7], whose singularities may be shown to be strongly conditioned by the biomechanical relations among parameters in well-known *k-mass* models [8]. Principal Component Analysis [9] is proposed to produce more compact data sets which can improve detection and classification results. This first approach to a more general study is oriented to the detection of specific speaker's characteristics as gender on the glottal biometrical signature of voice. Classically most works dealing with the biometry of voice have considered the voice signal as a whole, not establishing a clear separation among the roles played by the different organs implied in voice production (vocal tract vs vocal folds) [10]. After the early work of Brookes and Chan [11] it has been only in the last years when an interest has appeared in studying the characteristics of the Glottal Source separately from the vocal tract for speaker recognition [12][13]. Nevertheless, it seems intuitive that in treating voice biometry following a deconstructive way, important improvements could be obtained. This means that glottal parameters have to be treated separately accordingly to their statistical inter-speaker and intra-speaker characteristic distributions. Considering the classical source-filter voice generation model [14] composed by an excitation (Glottal Source) and a modulating structure (Vocal Tract), it may be expected that the excitation will depend on the biometric low-level characteristics of the speaker (glottal system physiology) being weakly influenced by the message (text), but strongly conditioned by the production process (physiological and emotional conditions, prosody, tonal height, production gesture, pathology, etc.). An analytical description of voice biometry is proposed in Figure 1.

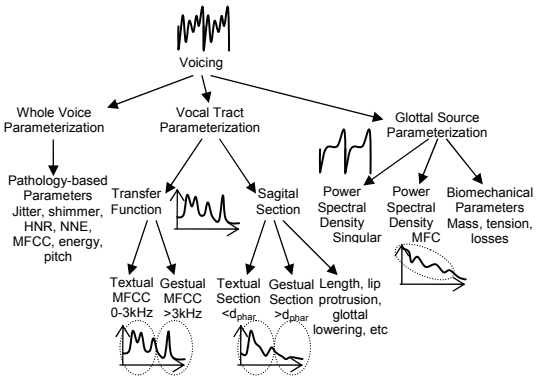


Figure 1. Analytic description of voice biometry in terms of vocal characteristics (mainly message dependent) or glottal characteristics (mainly biometric).

The parameterization of voice may be carried out using estimates of one of the following main categories (from left to right in the picture):

- The Whole Voice Power Spectral Density (WVPSD), estimated by FFT or LPC. The short-time power spectrum is coded as Mel-Frequency Cepstral Coefficients (MFCC) [15].
- The Vocal Tract Transfer Function Modulus (VTTFM). The WVPSD reflects the influence of the Glottal Source spectral envelope as a *1/f* spectral tilt, which distorts the Vocal Tract Transfer Function. A separation between Vocal Tract and Glottal Source could render better results in the decoding of message (Speech Recognition) as well as in the characterization of the source (Speaker Recognition).
- The Glottal Source Power Spectral Density (GSPSD). The Glottal Source can be parameterized in the time or in the frequency domain. Time domain methods are based in the well-known Liljiencrants-Fant model [16]. Frequency domain methods are preferred as they tend to be more robust facing noise or pathology [5].

Within the VTTFM a clear distinction could be made between frequency regions below 3000 Hz, which are more influenced by the message, and above 3000 Hz which are more

influenced by the speaker's gesture and personality. The parameterization of the Vocal Tract can be given as well in terms of its associated area functions (Sagital Section). In this case it is also possible to establish two segments: the oral part and the glottal part. The former is more influenced by articulation, the later is more related to the speaker's characteristics. Concerning the parameterization of the Glottal Source the time domain methods are oriented to the estimation of OC, SC, ClQ, RQ and NAQ (Open, Speed, Closing, Return and Normalized Amplitude Quotients). The frequency domain (GSPSD) is oriented to the estimation of H_1-H_2 which is known to be related to the CQ (Close Quotient), as well as the Maximum Flow Declination Rate (MFDR) and the Spectral Slope. Other methods are based on MFCC or LPCC parameterization of the power spectral density of the glottal signals (glottal flow derivative, glottal source derivative, etc.) similarly to VTTFM. Another line is related with the parameterization of the Glottal Source frequency envelope and the extraction of the biomechanical parameters of a k-mass glottal model by inversion as in [21].

2. Estimation of the glottal source

The methodology proposed in this work is based in a frequency domain parameterization of the glottal source power spectral density, with the following distinctive characteristics:

- It is carried out either on the Glottal Source or on the Mucosal Wave Correlate (MWC), derived from the Glottal Source by removing the Acoustic Average Wave (AAW) [17].
- It estimates the singularities of the Mucosal Wave Correlate Power Spectral Density (MWCPDS) as sets of peaks and notches relative to F_0. Therefore it can be considered as a generalization of the parameters used in [5].

Its biometrical character is granted by its inter- and intra-speaker statistical variability mainly conditioned by the personal characteristics of the speaker (gender, age, tension, glottal gesture, etc). The methodology used for the estimation of the Glottal Source is based on the elimination of the vocal tract by inverse filtering by well-known methods [18], and in the separation of the Glottal Source into the two referred components (AAW and MWC). An example of the glottal signal estimation results from inverse filtering may be seen in Figure 2 from quasi-stationary utterances of the vowel /a/ by typical male and female speakers.

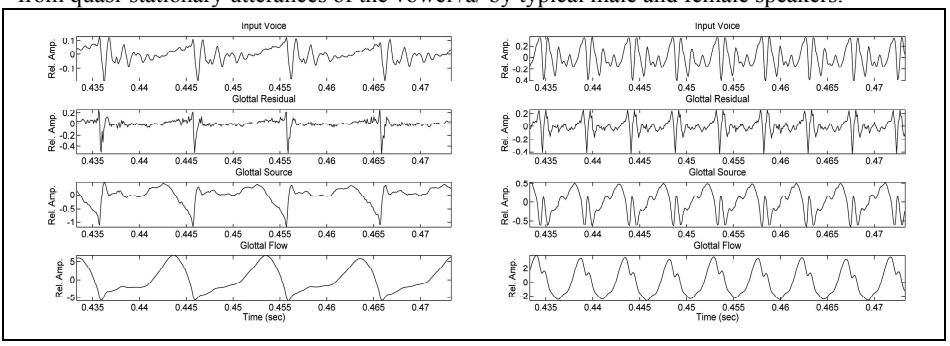


Figure 2. Examples of reconstructed glottal signals from vowel /a/ for prototype male and female speakers (*#185* –left-, and *#158* -right). From top to bottom: input voice, glottal residual, source and flow (four left templates: male prototype; four right templates: female prototype). Horizontal axes are given in sec for a sampling frequency of 11,050 Hz.

The plot in Figure 3 reproduces in detail the time evolution of a cycle of the Glottal Source (full line) where the four phonation phases may be observed separated by vertical dot lines from left to right: return, closure, open and closing phases. Two other variables are plotted as

well: the Average Acoustic Wave (dash-dot) and the Mucosal Wave Correlate (dash). The dash-dot plot corresponds to the ideal Glottal Source if no Vocal and Pharyngeal Tracts were present under non inertial load conditions assuming that each Vocal Fold could be represented by a single body mass (1-mass model). This would be equivalent to two ideal Vocal Folds with a single mass behaviour attached to the walls of the tract by single elastic springs. The vibration would describe perfect semi-sinusoidal arches accordingly to the relation between mass and spring constants. This signal coincides with the Acoustic Average Wave (AAW) and has been evaluated by optimally fitting a semi-sinusoid arch to the Liljencrants-Fant pattern [7].

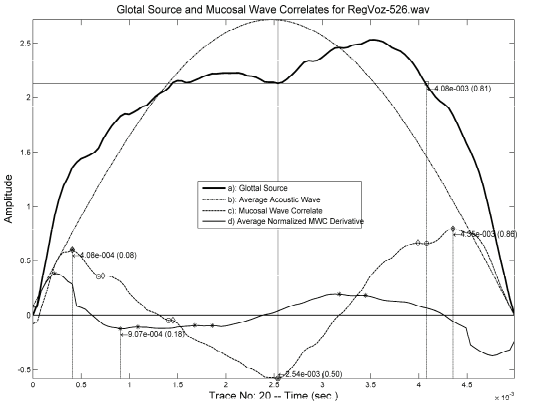


Figure 3. Splitting the Glottal Source (a) into the Average Acoustic Wave (b) and the Mucosal Wave Correlate (c). The minimum in the derivative of the MWC (d) marks the end of the return phase. The vertical middle dot line divides the phonation cycle into the close phase (left) and the open phase (right). The vertical side dot lines mark the return and closing points.

The dash plot corresponds to the difference between the AAW and the L-F Glottal Source plots, and is thus referred as the Mucosal Wave Correlate. This signal shows interesting properties, such as the ability of pointing out the start of the open phase, which takes place at its minimum (middle vertical dot line). This property may be used in detecting the open and close intervals of the phonation cycle.

3. Glottal-Source based Biometric Signature

Through the present approach a methodology to derive biometrical parameters of the Glottal Source in the frequency domain is proposed. The biometrical parameters are estimated on the power spectral density of either the Glottal Source or the Mucosal Wave Correlate. The signature obtained from the Mucosal Wave Correlate is more specifically related to the biomechanics of the vocal fold cover, while that from the Glottal Source includes the biomechanics of both the body and the cover of the vocal fold. The estimates based on this last approach are more suitable for biometric applications, the estimates from the Mucosal Wave Correlate being more suitable for studies in vocal fold pathology. In both cases the parameter estimation methodology to be applied is the same. The power spectral densities shown in Figure 4 correspond to the Glottal Source from prototype male and female voices. A common behaviour may be observed in both cases regarding the envelopes of the power spectral densities: a fast raise from low frequencies to a maximum and a decay towards lower frequencies with a general trend of *12 dB/oct*. In between a series of valleys or local minima may be appreciated surrounded by peaks.

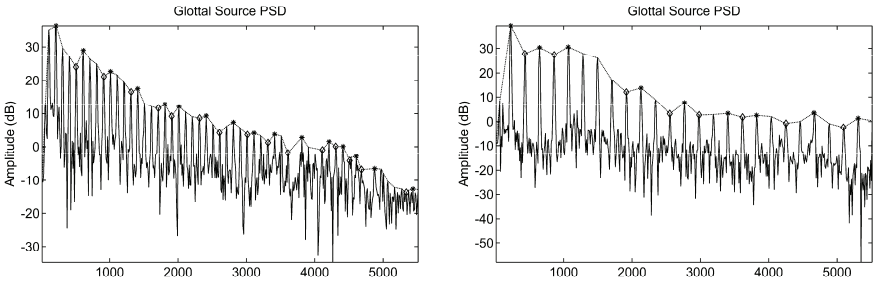


Figure 4. a) Power spectral density of the glottal source from vowel /a/ for prototype male and female speakers (*#185* –left-, and *#158* -right) showing the singularities superimposed: *-maxima; ◊-minima. Relative amplitude is given in dB. Horizontal axes are given in Hz.

In Figure 5 the envelope of the glottal source power spectral density of the male prototype has been extracted showing a first maximum T_{M1} centered at a frequency f_{M1} followed by a descent to a minimum T_{m1} in f_{m1} and to a new maximum T_{M2} at a frequency f_{M2}. This type of notch may appear several more times as the general trend of the power spectral density is decaying. The presence of two maxima enclosing a minimum is explained by the resonances and anti-resonances in the system of masses and springs on the vocal fold body and cover structures [8].

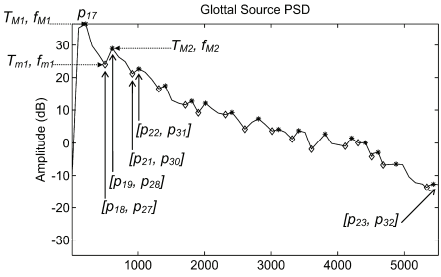


Figure 5. Power spectral density envelope of the glottal source for speaker #185 showing the first notch profile $\{T_{M1}, f_{M1}\}$, $\{T_{m1}, f_{m1}\}$ and $\{T_{M2}, f_{M2}\}$, and the meaning of 10 of the singularity parameters used in the study $\{p_{17}, p_{18}, p_{19}, p_{21}, p_{22}, p_{27}, p_{28}, p_{30}, p_{31}$ and $p_{32}\}$. Relative amplitude is given in dB. Horizontal axis is given in Hz.

Therefore a glottal signature of voice may be established detecting each notch by estimating the amplitude and position of its singularity points and its slenderness factor as described in [7]. In a practical case the biometrical signature is estimated from the singularities of the power spectral density of either the MWC or the Glottal Source as follows

- The Glottal Source is windowed in 512-sample frames and the power spectral density of each window is estimated by FFT in dB as in Figure 4.
- The envelopes of the power spectral densities are estimated for each frame.
- The maxima (*) and minima (◊) on the respective envelopes are detected and their amplitudes and frequencies collected as two lists of ordered pairs: $\{T_{Mk}, f_{Mk}\}$ and $\{T_{mk}, f_{mk}\}$, with k the ordering index.
- The first (and usually the largest of all maxima) (T_{M1}, f_{M1}) is used as a normalization reference both in amplitude and in frequency.
- The normalized singularity points and the approximate envelope of the power spectral densities for the MWC are assigned to the parameters for the study accordingly to Table 1.

Table 1. MWC singularity parameters used in the study.

Parameter No. and Description	Parameter No. and Description
p_{17} - Amplitude of the first maximum in dB T_{M1}	p_{26} - Absolute pos. of first maximum f_{M1}
p_{18} - Normalized ampl. of first minimum in dB τ_{m1}	p_{27} - Norm. pos. of first minimum φ_{m1}
p_{19} - Norm. ampl. of second maximum in dB τ_{M2}	p_{28} - Norm. pos. of second maximum φ_{M2}
p_{21} - Norm. ampl. of second minimum in dB τ_{m2}	p_{30} - Norm. pos. of second minimum φ_{m2}
p_{22} - Norm. ampl. of third maximum in dB τ_{M3}	p_{31} - Norm. position of third maximum φ_{M3}
p_{23} - Norm. ampl. of spec. prof. at max. freq. in dB τ_{fm}	p_{32} - Norm. position of end value φ_{fm}
p_{24} - Norm. position of initial value in freq. φ_i	p_{33} - Slenderness of first notch σ_{m1}
p_{25} - Norm. pos. of first min. before the first max. φ_{m0}	p_{34} - Slenderness of second notch σ_{m2}

Another possible parameterization strategy would be based on clipping voicing frames in segments aligned with the pitch cycle. In this way a different estimation would be produced for each cycle-like segment. In the present study voice frame durations of *0.2* sec. long are used producing different numbers of pitch cycles for male and female voice (typically ranging from 20-40). The number of pitch cycles used is designated as *M*. Assuming reasonable stationary conditions along the frame duration (considering that a stable vowel is produced) estimations of the parameter means and standard deviations could be used in classification as

$$x_{ij} = \frac{1}{M}\sum_{m=1}^{M} p_{ijm} \tag{1}$$

$$\sigma_{ij} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(p_{ijm} - x_{ij}\right)^2} \tag{2}$$

where *i* is the parameter index and *m* is the cycle index. In this way the estimations are more robust to intra-speaker variability as will be shown in the sequel.

4. Materials and methods

A corpus of 100 normal speakers equally distributed by gender was randomly recruited from a wider database recorded during the life of project MAPACI [19]. Speaker ages ranged from *19* to *39*, with an average of *26.77* years and a standard deviation of *5.75* years. The normal phonation condition of speakers was determined by electroglottography, video-endoscopiy and GRBAS evaluation [20]. The recordings consisted in utterances of the vowel */a/* of about *3 sec* per record. A *0.2 sec* frame from the record centre was used in the estimations. The spectral profile parameters $\{p_{17\text{-}34}\}$ As each parameter was estimated on a phonation cycle basis, for a prototype male voice (with pitch around *100 Hz*) an average of *M=20* values was obtained, which for female voice (with a typical pitch of *200 Hz*) should be around *M=40*. In this way *J=46* observation parameters x_{ij} were obtained as the average of each observation parameter p_{im} over $1 \le m \le M$ phonation cycles following (1) with $1 \le j \le J$ for each speaker $1 \le i \le I$ in the set of *I=100* speakers. The estimations of observation parameter *j* for all the speakers $1 \le i \le I$ in the set are stacked as a column vector

$$\boldsymbol{x}_j = \left[x_{1j}, x_{2j} \ldots x_{ij}, \ldots x_{Ij}\right]^T \tag{3}$$

Similarly the estimations for the whole set of parameters are piled as a matrix of observations

$$\boldsymbol{X} = \left[\boldsymbol{x}_1, \ldots \boldsymbol{x}_j, \ldots \boldsymbol{x}_J\right] \tag{4}$$

Principal Component Analysis was applied to this dataset as described in [9] and [22] to re-evaluate the set of observation parameters as

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{e}_j; \quad 1 \le j \le J \tag{5}$$

where the vectors $\boldsymbol{y}_j$ contain the new parameters (principal components) for each speaker in the list $1 \le i \le I$ their variance diminishing with component order. PCA was applied as follows:

a) Pre-selection of a database $\boldsymbol{X}_{17\text{-}34}$ from the original parameter set $\boldsymbol{S}_0 = \{x_{1\text{-}46}\}$ for the whole set of speakers. The resulting subset of parameters $\boldsymbol{S}_1 = \{x_{17\text{-}19}, x_{21\text{-}28}, x_{30\text{-}34}\}$ included the normalized estimates of the power spectral density singularities, as given in Table 1.
b) Z-score the database $\boldsymbol{X}_{17\text{-}34}$ by subtracting means and normalizing to standard deviations.
c) Split the database $\boldsymbol{X}(\boldsymbol{S}_1)$ in two clusters by *k-means* blindly (unsupervisedly).
d) Apply PCA on $\boldsymbol{X}(\boldsymbol{S}_1)$ to transform it to a new manifold for *16* principal components producing a matrix $\boldsymbol{Y}_{1\text{-}16}$ ordered by component relevance.
e) Select the three first components for 3-D presentation purposes.

5. Results and discussion

The results in $\boldsymbol{Y}_{1\text{-}16}$ have been plotted in terms of the three first principal components, as described in a)-e) as given in Figure 6.

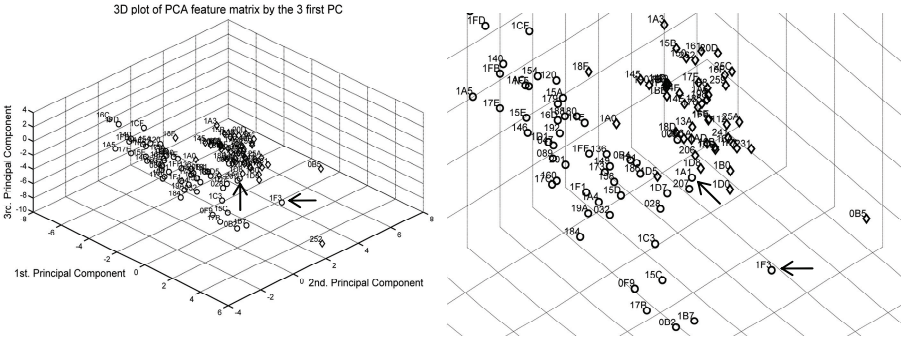


Figure 6. Classification results in the PCA manifold in terms of the first 3 principal components. Left: The set of samples is clustered into two main groups. Samples labelled (◊) are from male subjects, whereas those labelled (o) are from female subjects with two exceptions for #1A1 and #1F3, pinpointed by arrows. Right: Close-up view of the same plot.

These results show that the unsupervised clustering succeeded in accurately separating speakers by gender with the exception of the two male subjects grouped within the female cluster (*#1A1* and *#1F3*). The female cluster shows a broader branch-like inter-speaker variability than the male cluster, which is less spread-out. This may imply that different branches may be found within the mainly-female cluster and would deserve a further investigation. As a consequence it may be said that it will be easier to establish classifications within female than in male groups.

The main question to be answered at this point is which parameters will be more sensitive to gender, as there is a clear dependence of sample distributions on gender. To gain a better view on intra- and inter-speaker variability the following steps were covered:

- The average values and standard deviations of each speaker were evaluated for each parameter in their respective templates in terms of *M* accordingly with (1) and (2), thus serving as estimates of intra-speaker variability.
- The statistical dispersion of the parameter templates for the set of male and female subjects was presented as box plots, thus serving as estimates of inter-speaker variability.

Results are presented in Figure 7 (contrasted to the ones from the male and female prototypes).

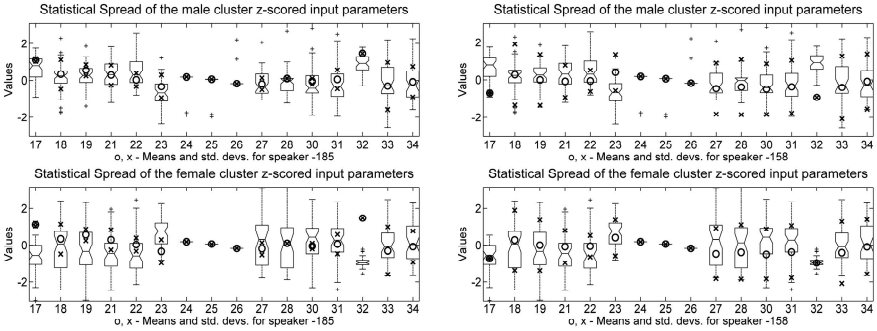


Figure 7. Statistical dispersion of the profile parameters used in the study from the male (left) and female (right) groups compared against their respective prototype male (#185) and female (#158) templates. The intra-speaker variability is expressed by the marks (o: mean) and (×: standard deviation). The inter-speaker variability is represented as notch-box plots.

First of all it must be mentioned that certain parameters (as x_{24}, x_{25} and x_{26}) show almost no variability, therefore they are not of specific interest for our study. The impression derived from Figure 6 about the wider spread presented by female distributions when comparing male and female clusters is clearly confirmed. This dispersion is especially relevant regarding parameters x_{18}, x_{19}, x_{27} and x_{28}, which are related to estimates of the first minimum and second maximum. Besides, certain parameter distributions from male and female groups do not overlap, or do so slightly, as is the case of x_{17}, x_{23} and x_{32}. This means that if used in differential clustering experiments they will render the best results, as will be shown in the sequel. Another interesting result is derived from the comparison between the prototype intra-speaker male template (*#185*) against the spread of the male and female groups. It may be seen that the prototype male template fits within the male parameter spread (showing slight deviations for x_{19}, x_{23}, x_{28} and x_{32}). The same template when compared against the female distribution is in clear disagreement with respect to parameters x_{17}, x_{23} and x_{32}. A similar comparison may be carried out on the female prototype (*#158*) against male inter-speaker variability, showing strong disagreements again with respect to parameters x_{17}, x_{23} and x_{32}. On the contrary only parameters x_{30} and x_{31} show slight deviations between the prototype female and the female inter-speaker dispersion. A further confirmation of these observations is obtained using Fisher's Discriminant Ratio

$$fdr_j = \frac{\left(\mu_{mj} - \mu_{fj}\right)^2}{\sigma_{mj}^2 + \sigma_{fj}^2}; \quad 1 \le j \le J \qquad (6)$$

where (μ_{mj}, σ_{mj}) and (μ_{fj}, σ_{fj}) are the means and standard deviations of the male and female distributions for parameter j. The results in Table 2 confirm the observations in the sense that x_{17}, x_{23} and x_{32} are the most relevant parameters in gender detection.

Table 2. Relevance of singularity parameters from FDR

Parameter index and name	Relevance	Parameter index and name	Relevance
32. MW PSD End Val. Pos. rel.	0.2083	23. MW PSD End Val. rel.	0.1430
17. MW PSD 1st Max. ABS.	0.1365	33. MW PSD 1st Min NSF	0.0213
18. MW PSD 1st Min. rel.	0.0146	31. MW PSD 4th Max. Pos. rel.	0.0092
19. MW PSD 2nd Max. rel.	0.0078	30. MW PSD 2nd Min. Pos. rel.	0.0040
34. MW PSD 2nd Min NSF	0.0029	28. MW PSD 2nd Max. Pos. rel.	0.0008
27. MW PSD 1st Min. Pos. rel.	0.0006	21. MW PSD 2nd Min. rel.	0.0005
22. MW PSD 4th Max. rel.	0.0003	24. MW PSD Origin Pos. rel.	0.0000
25. MW PSD In. Min. Pos. rel.	0.0000	26. MW PSD 1st Max. Pos. ABS.	0.0000

These results may be used in improving clustering strategies as shown in Figure 8, where clear differential groupings may be obtained when a non-overlapping parameter as x_{32} is used in a typical differentiating experiment.

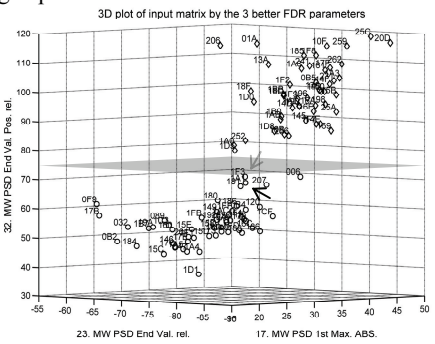


Figure 8. Differential clustering in terms of the most relevant parameters according to FDR: x_{17}, x_{23} and x_{32} where clear gender separation is produced by the plane at x_{32}=*75*.

The fact that gender splitting is feasible by hyperplane separation (x_{32}=75) suggests that more sophisticated techniques as Support Vector Machines may be used in complex multiple-feature separation (as gender, age or phonation modality).

6. Conclusions

From what has been shown the following important conclusions may be derived:

- The general decay trend of the glottal signal is coded in parameters p_{17}, p_{23} and p_{32}.
- Parameters p_{17}, p_{23}, p_{32} and p_{33} are the most sensitive ones to gender.
- Genders show different parameter dispersions, being broader in female than in male voice.
- In general intra-speaker parameter dispersion is lower than inter-speaker dispersion.
- PCA helps interpreting results by dimensionality reduction.

As a general conclusion it may be said that a structured classification of the biometry of voice is a real need, as specific and clearly differentiated biometric information is present in the glottal components of voice, independently from features observed in vocal tract features. Therefore splitting voice into vocal and glottal components is a reasonable technique when articulation and biometry are two different objectives, as for example, in forensic applications of voice. As speaker identification and characterization algorithms strongly rely on joint probability densities of the parameters used in the experiments the production of glottal and vocal parameter descriptions statistically independent may be the clue to more accurate speaker recognition methods. This is especially important as far as the False Acceptance rates in security applications are critical to determine the suitability of these techniques in a given scenario. In this respect a combination of vocal and glottal feature descriptors independently and in fusion experiments may help in establishing efficient strategies for the improvement of detection rates. The implementation of this methodology may rely in pitch-synchronous or pitch-independent strategies, both having been tested with similar results. This makes it suitable for its application in real scenarios in forensic and security frameworks. The methodology presented may be also generalized to the study of speaker features as age, voice profile, emotional features and others alike.

Acknowledgments

This work is being funded by grants TIC2003-08756 and TEC2006-12887-C02-01/02 from Plan Nacional de I+D+i, Ministry of Education and Science, by grant CCG06-UPM/TIC-0028 from CAM/UPM, and by project HESPERIA (http.//www.proyecto-hesperia.org) from the Programme CENIT, Centro para el Desarrollo Tecnológico Industrial, Ministry of Industry, Spain.

References

[1] R. M. Nickel, Automatic Speech Character Identification, IEEE Circuits and Systems Magazine 6 (4) (2006) 8-29.
[2] P. J. Price, Male and female voice source characteristics: Inverse filtering results, Speech Comm. 8 (1989) 261-277.
[3] D. G. Childers et al., Automatic recognition of gender by voice, Proc. of the ICASSP 1 (1988) 603-606.
[4] A. R. Sulter and H. P. Wit, Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age, J. Acoust. Soc. Am. 100 (1996) 3360-3373.
[5] P. Alku, Parameterisation Methods of the Glottal Flow Estimated by Inverse Filtering, Proc. of VOQUAL'03, Geneva, August 27-29 (2003) 81-87.
[6] S. P. Whiteside, Sex-specific fundamental and formant frequency patterns in a cross-sectional study, J. Acoust. Soc. Am. 110 (1) (2001) 464–478.
[7] P. Gómez et al., Evaluation of voice pathology based on the estimation of vocal fold biomechanical parameters, J. Voice 21 (4) (2007) 450-476.
[8] D. A. Berry, Mechanisms of modal and non-modal phonation, J. Phonetics 29 (2001) 431-450.
[9] Johnson, R. A., Wichern, D. W., Applied Multivariate Statistical Analysis, Prentice-Hall, Upper Saddle River, NJ, 2002.
[10] F. Bimbot et al., A tutorial on text-independent speaker verification, Eurasip Journal on Applied Signal Processing (4) (2004) 430–451.
[11] D. M. Brookes and D. S. F. Chan, Speaker characteristics from a glottal airflow model using robust inverse filtering, Proc. Inst. of Acoustics 16 (5) (1994) 501–508.
[12] M. D. Plumpe et al., Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification, IEEE Trans. on Speech and Audio Proc. 7 (5) (1999) 569-586.
[13] N. Zheng et al., Integration of Complementary Acoustic Features for Speaker Recognition, IEEE Signal Proc. Letters 14 (3) (2007) 181-184.
[14] G. Fant, Theory of Speech Production, Mouton, The Hague, Netherlands (1960).
[15] J. R. Deller, J. G. Proakis J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan, NY (1993).
[16] G. Fant et al., A four-parameter model of glottal flow, STL-QSPR 4 (1985) 1-13. Reprinted in: Speech Acoustics and Phonetics: Selected Writings, G. Fant, Kluwer Academic Publishers, Dordrecht (2004) 95-108.
[17] I. R. Titze, Summary Statement, Workshop on Acoustic Voice Analysis, National Center for Voice and Speech (1994).
[18] P. Alku, An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulseform, Proc. of the ICASSP'92 (1992) II/29-32.
[19] Project MAPACI: http://www.mapaci.com.
[20] M. Hirano et al., Acoustic analysis of pathological voice. Some results of clinical application, Acta Otolaryngologica 105 (5-6) (1988) 432-438.
[21] P. Gómez et al., Biometrical Speaker Description from Vocal Cord Parameterization, Proc. of ICASSP'06, Toulouse, France, (2006) 1036-1039.
[22] P. Gómez et al., PCA of Perturbation Parameters in Voice Pathology Detection, Proc. of INTERSPEECH'05 (2005) 645-648.

Experiencia del I3A en la Evaluación de Reconocimiento de Locutor NIST 2008

Jesús A. Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel, José E. García, Luís Buera, Óscar Saz

Grupo de Tecnologías de las Comunicaciones (GTC), Instituto de Investigación en Ingeniería de Aragón (I3A), Zaragoza, España
{villalba,cvaquero,lleida,ortega,amiguel,jegarlai,lbuera,oskarsaz}@unizar.es

Abstract. En este artículo se describe el sistema de reconocimiento de locutor implementado por el I3A para la evaluación del NIST 2008. Se dispone de dos sistemas básicos: GMM-UBM likelihood ratio y GMM-SVM. Las señales proporcionadas por el NIST para la evaluación han sido adquiridas a través de diferentes micrófonos y canales de comunicación, se discutirá como afectan las diferentes técnicas de compensación de canal al funcionamiento del sistema. Se presentan los resultados obtenidos durante el desarrollo del sistema sobre la base de datos de 2006 y los obtenidos en 2008.

Palabras Clave: Speaker Recognition, Gaussian Mixture Model (GMM), Support Vector Machine (SVM), Feature Warping, Nuisance Attribute Projection (NAP), NIST.

1 Introducción

Periódicamente el Instituto Nacional de Estándares y Tecnología Americano (NIST) lleva a cabo una evaluación de sistemas de reconocimiento de locutor con el fin de comparar las prestaciones de diferentes técnicas sobre un corpus de test común. Como en años anteriores, la tarea en la que se basa la evaluación 2008 [1] es detección de locutor, es decir, determinar si un determinado locutor está presente en un segmento de voz dado.

Mientras que en años anteriores la evaluación estaba basada básicamente en señales de canal telefónico, en 2008 se incluyen también dentro de la condición principal conversaciones telefónicas y entrevistas grabadas usando diferentes tipos de micrófonos. Para cada uno de los intentos se conoce, para el segmento de entrenamiento y test, si es canal telefónico o micrófono, si es conversación telefónica o entrevista, el idioma y el sexo de locutor.

En la preparación de esta evaluación se han implementado dos sistemas básicos basados en el modelado de las características cepstrales (MFCC) de la señal de voz mediante modelos de mezclas de gaussianas. El primero de ellos es el clásico GMM-UBM [2] consistente en evaluar el log-ratio de verosimilitud entre el modelo del locutor de test y el modelo universal (UBM) que representa al locutor medio. El segundo es el GMM-SVM [5], que utiliza la capacidad discriminativa de las support vector

machines para comparar los supervectores obtenidos concatenando las medias de los GMM de los segmentos de entrenamiento y test, consiguiendo resultados que mejoran a los del GMM-UBM. La presencia en la evaluación de diferentes canales telefónicos y micrófonos degrada considerablemente las prestaciones de estos sistemas, por ello se hace necesaria la aplicación de técnicas de compensación de canal y normalización de scores entre las que se encuentran: sustracción de la media del cepstrum (CMS), Feature Warping [3], Nuisance Attribute Projection (NAP) [6], T-Norm [4] y Z-Norm.

Este artículo se divide en las siguientes secciones: en la sección 2 se describen los sistemas implementados incluyendo la extracción de características, los tipos de clasificadores y las diferentes técnicas de compensación de canal; en la sección 3 se describen los experimentos realizados y bases de datos utilizadas para desarrollar el sistema y los resultados conseguidos con la base de datos NIST 2006 y en la presente evaluación NIST 2008; finalmente en la sección 4 se exponen la conclusiones y los pasos a dar en el futuro.

2 Descripción del Sistema

2.1 Extracción de Características

El Front-End extrae los 16 primeros Mel Frequency Cepstral Coefficients (MFCC) incluyendo el C0 a los que se añaden sus primeras y segundas derivadas tomando tramas de 25 msg. con un desplazamiento de 10 msg. El C0 se elimina manteniendo sólo sus derivadas quedando finalmente un vector de dimensión 47. El banco de filtros triangulares se modifica para encajar con el ancho de banda de canal telefónico 0.3-3.4 kHz y limitar la influencia del ruido de baja frecuencia presente en la mayoría de señales. Para las condiciones que incluyen señal de micrófono, mucho más ruidosa que la telefónica, ya sea en entrenamiento o en test se utiliza el Advanced Front-End del ETSI (AFE) [9], que a las características anteriores añade un filtrado de Wiener.

La selección de tramas de voz se realiza mediante un umbral sobre la log-energía. Para estimar el umbral se modela la distribución de la log-energía de la señal mediante un modelo bigaussiano, La gaussiana de mayor energía se supone asociada al habla del locutor y la de menor energía asociada al ruido. Buscando el valle entre las mismas se puede determinar el umbral de energía que decidirá si una trama es de voz o es ruidosa. Además se selecciona otro umbral 30dB por debajo de la energía de pico del segmento, por si el nivel de ruido es demasiado bajo para estimar correctamente la gaussiana de menor energía. El umbral seleccionado es el mayor de ambos. Este algoritmo funciona correctamente con SNR aceptables pero puede fallar en el caso de grabaciones con micrófonos de campo lejano mucho más ruidosas, en el caso de que el tanto por ciento de tramas escogidas no sea suficiente para modelar a un locutor se selecciona el umbral que deja pasar el 30 % de las tramas de más energía. A esta selección de tramas se aplica un filtro de mediana, para eliminar tramas de voz aisladas, que puedan deberse a ruidos impulsivos. Dicho filtro es asimétrico: Descarta tramas de voz rodeadas de silencio, pero nunca recupera una trama de silencio convirtiéndola en voz.

2.2 GMM-UBM

Como se ha dicho anteriormente, cada locutor se modela mediante una mezcla de gaussianas. Dicho modelo se obtiene mediante adaptación MAP de las medias de un UBM [2].

$$\mu_k = \alpha_k E_k[x] + (1-\alpha_k)\mu_k^{UBM} \quad \textbf{(1)}$$

$$\alpha_k = \frac{c_k}{c_k + \tau} \quad c_k = \sum_t c_{kt} \quad E_k[x] = \frac{\sum_t c_{kt} x_t}{c_k} \quad \textbf{(2)}$$

siendo c_{kt} la probabilidad a posteriori de que la muestra x_t pertenezca a la gaussiana k y α_k el coeficiente de adaptación MAP.

El UBM se estima previamente mediante el algoritmo EM utilizando gran cantidad de señal de diferentes locutores, y representa el modelo del locutor medio. Como en la evaluación del NIST no hay test cruzados hombre-mujer se obtienen modelos universales de hombre y mujer por separado.

La evaluación de cada intento se realiza calculando el valor del ratio de verosimilitud entre el modelo Target y el UBM.

$$LLR = \log\left[p(O \mid TARGET)\right] - \log\left[p(O \mid UBM)\right] \quad \textbf{(3)}$$

Para evitar tener que evaluar todas las gaussianas de UBM y Target se aprovecha que las gaussianas de ambos modelos son correspondientes para obtener del UBM las N gaussianas de mayor probabilidad y solo evaluar esas en el Target. Esta técnica se conoce con el nombre de Fast-Scoring. Las pruebas indican que es necesario evaluar al menos 10 gaussianas para no perder prestaciones.

2.3 GMM-SVM

Un SVM es un clasificador binario formado por sumas de una función kernel:

$$f(x) = \sum_{i=1}^{L} \alpha_i y_i K(x_i, x) + b \quad \textbf{(4)}$$

El proceso de optimización coloca un hiperplano capaz de separar ambas clases en el espacio de alta dimensionalidad definido por el kernel. Los supervectores de entrenamiento que se encuentran en la frontera de separación constituyen los vectores soporte. El proceso de entrenamiento consiste en la obtención de estos vectores soporte que modelan la frontera de separación.

Como aparece en [5] a partir de la aproximación de la distancia KL se puede obtener el siguiente kernel que es un producto escalar de dos supervectores:

$$K(i,j) = \sum_k w_k (\mu_k^i)^T \Sigma_k^{-1} \mu_k^j = \sum_k (\sqrt{w_k}\Sigma_k^{-1/2}\mu_k^i)^T (\sqrt{w_k}\Sigma_k^{-1/2}\mu_k^j) = \phi(i)\phi(j) \quad \textbf{(5)}$$

De este modo para cada modelo construimos un supervector concatenando sus medias normalizadas por su desviación típica, y ponderadas por la raíz cuadrada del peso de su gaussiana. El kernel es el producto escalar de los supervectores que queremos comparar.

Se entrena un SVM para cada modelo target utilizando la librería SVMTorch [8]. Se pasa a la librería el modelo del locutor como único ejemplo positivo y varios modelos de locutores de background como ejemplos negativos. Para evaluar se utiliza la siguiente función:

$$f(x) = \sum_{i=1}^{L} \alpha_i y_i K(x_i, x) + b = \left(\sum_{i=1}^{L} \alpha_i y_i \phi(x_i) \right)^T \phi(x) + b = w^T \phi(x) + b \quad \textbf{(6)}$$

Donde $\phi(x)$ es el supervector obtenido a partir del modelo que se estima mediante adaptación MAP al fichero de test, $\phi(x_i)$ son los vectores soporte que da como resultado el algoritmo de optimización cuadrática implementado por SVMTorch y $\alpha_i y_i$ los pesos de los mismos. Al ser el kernel un producto escalar no es necesario almacenar todos los vectores soporte como ocurre con otros kernels sino que podemos limitarnos a calcular el vector del hiperplano que separa ambas clases y la evaluación consiste simplemente en el producto escalar del vector del plano por el supervector de test.

2.4 Normalización de Parámetros

Sustracción de la Media del Cepstrum (CMS).
Una gran ventaja de los MFCC es que constituyen una transformación homomórfica, de forma que, en teoría, las convoluciones y efectos de filtrado en el dominio temporal se convierten en sumas en el dominio cepstral. Suponiendo que el canal no varía, la contribución del mismo al valor de los MFCC se convierte, principalmente, en una constante aditiva. Dicha constante se puede eliminar fácilmente restando directamente la media temporal del valor de los MFCC a cada vector de los mismos [10].

Feature Warping.
En el caso de que además de ruido convolucional, haya presente distorsión no lineal y ruido en el canal, el CMS no es capaz de compensar estos efectos. Una técnica que se probado eficaz en estos casos consiste en aplicar una ecualización de histograma a cada componente de tal manera que tenga una distribución fija, generalmente una gaussiana de media cero y varianza unidad. Se ha demostrado experimentalmente [3] que para la tarea de verificación de locutor, la ventana de análisis óptima es de 3 segundos, de forma que dicha ventana se desplaza trama a trama y únicamente se aplica la transformación a la trama situada en el centro de la ventana, considerando únicamente las tramas de voz.

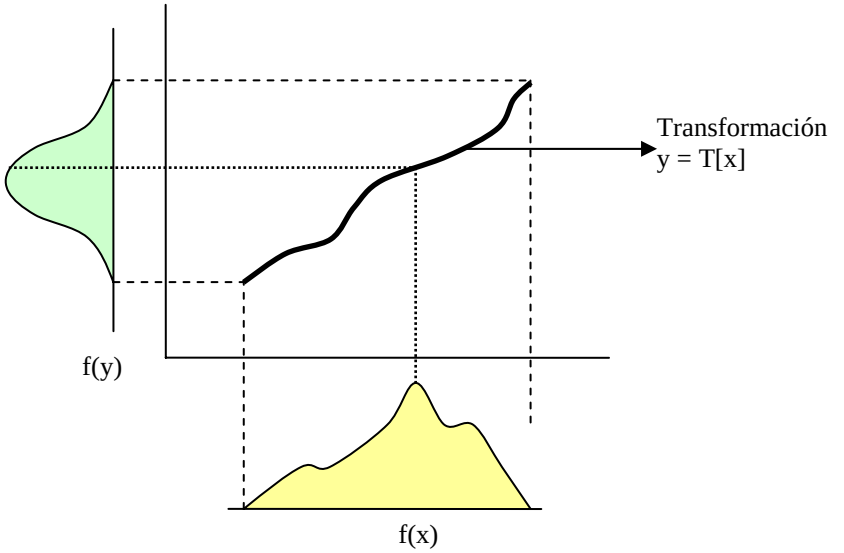


Fig. 1. Feature Warping

2.5 Nuisance Attribute Projection (NAP)

Como técnica de compensación de canal para el sistema GMM-SVM se ha implementado Nuisance Attribute Projection [5][6]. Esta técnica consiste en encontrar una matriz de transformación del tipo $P = I - vv^T$ que proyecte los supervectores en un subespacio vectorial más resistente a las variaciones intra-locutor y de canal. Las columnas de **v** son los k vectores que representan las direcciones de variación del supervector con los cambios de canal o de otros factores. El criterio para obtener P consiste en:

$$v^* = \arg\min_{v,\|v\|_2=1} \sum_{i,j} W_{ij} \left\| P\phi(x_i) - P\phi(x_j) \right\|_2^2 \quad (7)$$

donde W_{ij} debe escogerse de diferente manera en función de las direcciones de variación que se quieran eliminar, W_{ij}=1 si se quiere $\phi(x_i)$ se parezca a $\phi(x_j)$, W_{ij}=-1 si se quiere que sean distintos y W_{ij}=0 si no importa. Utilizando que P^2=P se puede demostrar que los vectores de v se obtienen resolviendo el siguiente problema de autovalores:

$$A(diag(W1) - W)A^T v = \Lambda v \quad (8)$$

donde ***A*** es la matriz cuyas columnas son los supervectores $\phi(x_i)$, **1** es el vector de todo unos y $\boldsymbol{W}=(W_{ij})$.

Para eliminar las direcciones de variación debidas a cambios en el locutor o en el canal entre sesiones se fija W_{ij}=1 si $\phi(x_i)$ pertenece al mismo locutor que $\phi(x_j)$ y W_{ij}=0

en caso contrario. Para este caso concreto podemos desarrollar la formula anterior de tal manera que la matriz de la que hay que hallar los autovalores es:

$$S = \sum_{i=1}^{L} n_i \sum_{j=1}^{n_j} (\phi_i^j - \overline{\phi_i})(\phi_i^j - \overline{\phi_i})^T \quad \textbf{(9)}$$

donde n_i es el número de supervectores del locutor i, ϕ_i^j es el supervector del locutor i en la sesión j $\overline{\phi_i}$ es el supervector medio del locutor i y L es el numero de locutores. Para obtener los autovectores de estas matrices de una forma eficiente se ha seguido el siguiente procedimiento. Se crea una matriz ***B*** cuyas filas son los supervectores de cada locutor menos su media multiplicada por la raiz cuadrada del número de vectores del locutor de tal forma que:

$$B = \left(\sqrt{n_i}(\phi_i^j - \overline{\phi_i})\right)^T; \quad S = B^T B \quad \textbf{(10)}$$

Es inviable calcular la matriz S por su tamaño (varios GB) pero se pueden obtener los autovectores a partir de la descomposición en valores singulares de ***B*** sin necesidad de hacer el producto.

2.6 Normalización de Scores

Los scores que se obtienen de la evaluación del modelo pueden presentar gran variabilidad, debida fundamentalmente al desajuste existente entre la fase de entrenamiento y funcionamiento del sistema, pero también debida a otros factores difíciles de eliminar como el locutor en sí.

T-Norm
Se normaliza el score con la media y varianza de scores que obtiene la señal de test impostando varios modelos de background. De esta forma se acota el rango dinámico de scores que produce una señal de test. Estos modelos se escogen de manera que sean similares al del locutor objetivo, para seleccionarlos se ha utilizado una aproximación de la distancia Kullback Leibler [5]:

$$KL(i, j) \leq \sum_k w_k (\mu_k^i - \mu_k^j)^T \Sigma_k^{-1} (\mu_k^i - \mu_k^j) \quad \textbf{(11)}$$

Z-Norm
Se normaliza el score con la media y varianza de scores que obtiene el modelo del locutor objetivo al ser impostado por varios modelos de background. En este caso lo que se acota es el rango dinámico de scores que produce el modelo objetivo.

3 Experimentos y Resultados

3.1 Efecto de las Técnicas de Compensación de Canal y Normalización de Scores.

Se han realizado experimentos sobre la condición principal NIST 2006 (tlfn-tlfn) orientados a comprobar la mejora que aportan cada una de las técnicas de compensación de canal expuestas en la sección 2 por separado. Para ello se ha partido de un baseline consistente en el sistema GMM-UBM básico con 256 gaussianas al que se han ido añadiendo cada una de estas técnicas. Los datos de desarrollo utilizados vienen descritos en la tabla 4. Los resultados obtenidos se resumen en la tabla 1 y la figura 2:

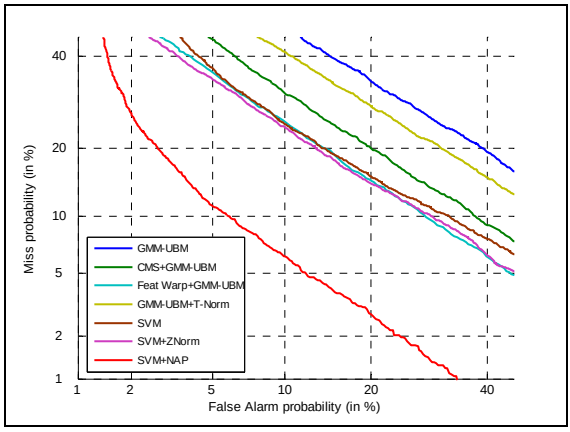


Fig. 2. Curvas DET resumen de técnicas de compensación de canal y normalización de scores.

Tabla 1. Resultados de las Tecnicas de Compensación de Canal.

	ERR(%)	Mejora GMM-UBM (%)	Mejora SVM (%)
Baseline (GMM-UBM)	27.5	0	0
CMS+GMM-UBM	20	27	0
Feat Warp+GMM-UBM	16.8	39	0
GMM-UBM+T-Norm	24.5	11	0
SVM	17.2	37	0
SVM+ZNorm	16.5	40	4
SVM+NAP	7.9	71	54

3.2 Resultados del Sistema Completo en NIST 2006

Una vez comprobado el aporte de cada una de las técnicas por separado se van a proceder a superponer para obtener el sistema completo. Se muestran resultados para la condición principal (tlfno-tlfno) y para la condición cross-channel (tlfno-mic) utilizando modelos de 512 gaussianas.

Tabla 2. Resultados superponiendo diferentes técnicas de compensación y normalización.

Condición	Tlfno-Tlfno		Tlfno-Mic	
	EER (%)	Mejora (%)	EER (%)	Mejora (%)
Warp+GMM-UBM	8.9	0	11.5	0
+TNorm	7.9	+11.2	9.8	+10.6
Warp+SVM	6.9	+12.6	10.7	-9,1
+NAP	5.3	+23.1	5.9	+44.8
+ZNorm	5	+5.6	4.4	+25.4

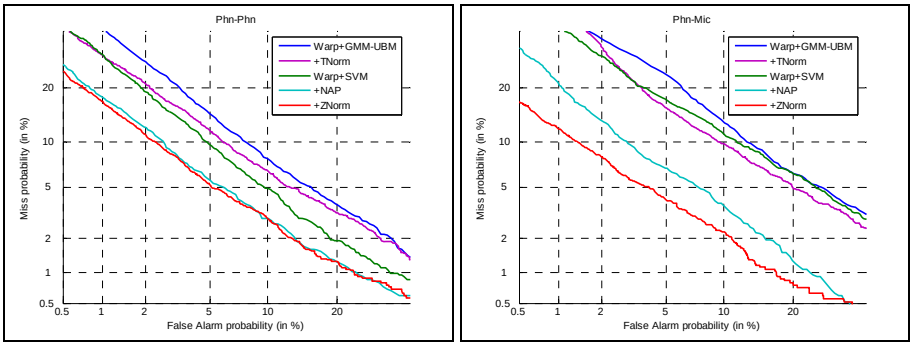


Fig. 3. Curvas DET superponiendo diferentes técnicas de compensación y normalización.

3.3 Resultados del Sistema Completo en NIST 2008

El sistema primario para evaluación 2008 incluye Feature Warping, SVM-NAP y Z-Norm 512 gaussianas. Se ha observado en el caso de que los segmentos tengan pocas tramas, estas no son suficientes para extraer un buen modelo de 512 gaussianas, y se obtienen mejores resultados utilizando modelos de mejor orden. Por ello, uno de los sistemas secundarios enviados consiste en la fusión del sistema primario con 256 y 512 gaussianas utilizando regresión logística lineal [11]. A continuación se presentan los resultados obtenidos en NIST 2008 para las diferentes condiciones existentes.

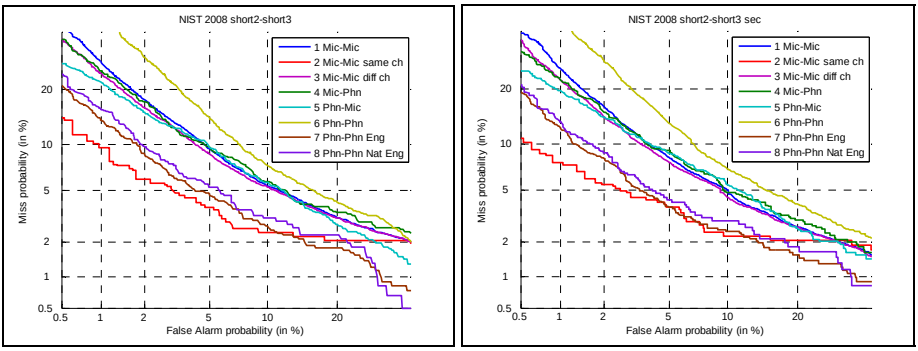


Fig. 4. Curvas DET del sistema en la evaluación NIST 2008 para todas las condiciones.

Tabla 3. Resultados del sistema en la evaluación NIST 2008 para todas las condiciones.

Condición	1	2	3	4	5	6	7	8
EER(%) Primario	7.1	4.1	6.8	7.6	7.2	8.5	4.8	5.2
EER(%) Secundario	6.6	4.2	6.4	6.9	6.9	8.2	4.3	4.6

3.4 Datos de Desarrollo

En función del tipo de condición de entrenamiento y test de cada intento se han utilizado diferentes datos de desarrollo para entrenar modelos de background, proyecciones NAP, etc. Todos los datos empleados pertenecen a evaluaciones previas del NIST. En la tabla siguiente se resumen los datos utilizados en cada caso.

Tabla 4. Datos de desarrollo para las diferentes condiciones.

<table>
<tr><th></th><th>Tlfno-Tlfno</th><th>Tlfno-Mic</th><th>Mic-Tlfno</th><th>Mic-Mic</th></tr>
<tr><td>UBM</td><td>Entrenamiento 2004</td><td colspan="2">Entrenamiento 2004 + xchannel 2006 y 2005</td><td>xchannel 2006 y 2005</td></tr>
<tr><td>NAP</td><td>Locutores 2004 con más de una sesión (1500 sesiones)</td><td colspan="2">Tlfno+xchannel 2005 (3000 sesiones)</td><td>xchannel 2006 y 2005 (2500 sesiones)</td></tr>
<tr><td>T-Norm/SVM Background</td><td colspan="2">120 Loc. entrenamiento 2004</td><td colspan="2">50 loc x 8 canales xchannel 2005</td></tr>
<tr><td>Z-Norm</td><td>120 Loc. Test 2004</td><td>50 loc x 8 canales xchannel 2005</td><td>120 Loc. Test 2004</td><td>50 loc x 8 canales xchannel 2005</td></tr>
</table>

4 Conclusiones

Se ha presentado la evolución del sistema de reconocimiento de locutor del I3A para la evaluación NIST 2008 junto con las prestaciones de diferentes métodos de compensación de canal y normalización de scores. Se ha comprobado que el sistema es robusto al cambio de bases de datos dando tasas de error comparables entre 2006 y 2008. También lo es entre las distintas condiciones de entrenamiento y test de 2008, a pesar de que las señales de micrófono son bastante más ruidosas que las telefónicas, el Feature Warping y el NAP han sido capaces de compensar una gran parte de la variación inter-canal.

En los resultados de 2008 se aprecia una degradación considerable de las prestaciones entre la condición tlfno-tlfno genérica y las que solo incluyen voz en ingles. Esto se debe a que en esta evaluación se han introducido multitud de idiomas que no estaban presentes en evaluaciones anteriores y que por tanto no tienen modeladas sus características en el UBM, NAP, etc. La problemática del idioma tendrá que ser muy tenida en cuenta de cara a próximas evaluaciones a la hora de diseñar el sistema y de escoger los datos de desarrollo, ya que, como muestran los resultados, puede degradar el funcionamiento del sistema tanto o más que las diferencias de canal. También se ha

visto que apenas hay diferencia entre los resultados de la condición ingles genérica y la de ingles sólo hablado por nativos.

Una de las dificultades que se encuentran al construir un sistema de estas características es contar con datos de desarrollo suficientes para entrenar todos modelos de background, NAP y normalizaciones necesarios. Es necesario dedicar esfuerzos a la optimización en la selección de estos datos.

5 Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Educación y Ciencia español a través del Proyecto Nacional TIN 2005-08660-C04-01.

Referencias

1. http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf
2. D. Reynolds, T. Quatieri, R. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models” Digital Signal Processing 10, pp 19-41 (2000).
3. Jason Pelecanos, Sridha Sridharan, “Feature Warping for Robust Speaker Verification”, Odissey 2001.
4. D. Ramos, D. Garcia, I. Moreno, J. Gonzalez, “Speaker Verification Using Fast Adaptive TNorm Based On Kullback Leibler Divergence” Third Cost 275 Worshop, Biometrics On The Internet (2005).
5. W. M. Campbell, D. Sturim, and D. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” *IEEE Signal Processing Letters*, vol. 13, no. 5, May 2006.
6. W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, “SVM based speaker verification using a GMM supervector kernel and NAP variability compensation,” in *IEEE-ICASSP*, Toulouse, France, 2006.
7. Alex Solomonoff, W. M. Campbell, and I. Boardman, “Advances in channel compensation for SVM speaker recognition,” in *Proceedings of ICASSP*, 2005.
8. R. Collobert and S. Bengio, “SVMTorch: Support vector machines for large-scale regression problems,” *J. Mach. Learn. Res.*, vol. 1, pp.143–160, 2001.
9. ETSI ES 202 050 recommendation, 2002. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms.
10. Furui Sadaoki. ”Cepstral analysis technique for automatic speaker verification”. IEEE Transactions on speech and audio processing, Vol. ASSP-29, No.2. April 1981
11. http://www.dsp.sun.ac.za/~nbrummer/focal

Identificación de locutores en entornos multilingües

Iker Luengo, Eva Navas, Iñaki Sainz, Ibon Saratxaga, Jon Sanchez, Igor Odriozola, Inmaculada Hernaez

Universidad del País Vasco, Alda. Urquijo s/n,
48013 Bilbao, Spain
{ikerl, eva, inaki, ibon, ion, igor, inma}@aholab.ehu.es

Abstract: Los sistemas de identificación y verificación de locutor tienen resultados pobres cuando el modelo se entrena en un idioma mientras que las pruebas se realizan en otro. Esta situación es bastante común en entornos multilingües, en donde los usuarios deberían poder utilizar el sistema en el idioma que prefieran en cada momento, sin notar una reducción en la fiabilidad del mismo. En este trabajo se estudia la posibilidad de utilizar parámetros derivados de características prosódicas con el objetivo de reforzar la independencia del idioma de estos sistemas. Un análisis previo de las características de los parámetros en términos de variabilidad frente al idioma y la sesión predice un incremento en la robustez frente al idioma cuando parámetros MFCC tradicionales se combinan con valores de energía y entonación extraídos para cada trama. Los resultados experimentales confirman que estos parámetros proporcionan una mejor tasa de reconocimiento de locutor cuando entrenamiento y prueba se realizan con idiomas diferentes.

1 Introducción

En los últimos años varios grupos de investigación han centrado su atención en los sistemas de reconocimiento de locutor en entornos multilingües, donde puede ocurrir que los modelos de locutor se entrenen en un idioma pero que se usen en otro. Trabajos como los realizados por Faúndez y Satué-Villar [1] y Durou [2] demuestran que hay una reducción en la precisión del sistema en estas condiciones, pero no aportan alternativas para aliviar el problema. Otros trabajos como los de Akbacak y Hansen [3] y Ma y Meng [4] proponen algún tipo de solución, pero sus propuestas siempre implican conocer de antemano los posibles idiomas que van a ser utilizados, lo cual no siempre es posible.

En este trabajo se trata de encontrar una solución a nivel de parámetros, es decir, buscar una parametrización que ayude a mantener la tasa de acierto bajo condiciones de desadaptación de idiomas. Al ser una solución a nivel de parámetros, debería ser completamente generalizable a cualquier idioma no visto durante el entrenamiento. Para ello, el presente estudio se centra en el País Vasco, en donde coexisten dos idiomas oficiales: castellano y euskera. Ambos idiomas tienen muy poco en común, ya que el euskera no es un idioma indo-europeo como el castellano. De hecho, el euskera está considerado un idioma aislado, que no tiene ninguna relación con ninguna otra

lengua viva o muerta. Esto proporciona una situación real que puede considerarse como el peor caso para un sistema de reconocimiento de locutor, ya que las diferencias existentes son mucho mayores que las puramente dialectales.

2 Definición del Problema

2.1 Identificación de Locutores en Desadaptación de Idiomas

El método más habitual de diseñar sistemas de reconocimiento de locutor es utilizar modelos de mezclas gaussianas (GMM) [5] para modelar la distribución de parámetros espectrales a corto plazo, tales como MFCC y LPCC [6][7]. Estos parámetros espectrales caracterizan el filtro que modela el tracto vocal de cada locutor en el momento de articulación, capturando por tanto no sólo las características del tracto vocal (permitiendo por tanto la identificación del locutor), sino también las características del tracto vocal para cada fonema. Esto supone que este tipo de parametrizaciones contienen también información acerca del contenido fonético de la locución.

En un sistema de reconocimiento de locutores independiente del texto los problemas surgen cuando, en un entorno multilingüe, el modelo se entrena en un idioma pero las pruebas se realizan en otro. Normalmente el contenido fonético de ambos idiomas no coincide, por lo que los scores de las locuciones de prueba no serán fiables, incrementando la tasa de error del sistema.

2.2 Solución Propuesta

Un método inmediato para reducir la discrepancia entre las locuciones de prueba y el modelo es realizar el entrenamiento con grabaciones en ambos idiomas. De esta forma, es probable que el modelo consiga aprender las características de todos los fonemas. Esta solución es la adoptada por Ma y Meng en su trabajo [4]. Otra posible vía es entrenar un modelo diferente para cada locutor e idioma, y usar un detector de idioma para decidir qué modelo utilizar durante la prueba, tal y como proponen Akbacak y Hansen [3]. Pero este tipo de soluciones requiere conocer de antemano los idiomas que van a ser utilizados, ya que no son generalizables a idiomas no vistos durante la fase de entrenamiento. Por tanto, sería conveniente disponer de una solución más independiente de idioma.

En los últimos años se ha propuesto e implementado con éxito el uso de parámetros de alto nivel para problemas de reconocimiento de locutor en entornos monolingües [8]. Entre estos parámetros, las características prosódicas, que están relacionadas con la entonación, la energía y la velocidad del habla, parecen una buena alternativa [9], ya que pueden ser estimados fácilmente mediante algoritmos automáticos de procesado de señal y pueden ser calculados incluso para señales muy cortas. Al igual que las parametrizaciones espectrales, estas características prosódicas contienen información tanto del locutor como del idioma utilizado. En el caso de sistemas multilingües, utilizar estos parámetros prosódicos será conveniente si su variabilidad entre locutores es

mayor que su variabilidad entre idiomas. En este caso es razonable decir que la prosodia es menos dependiente de idioma que dependiente de locutor. Con el objetivo de ver si las características prosódicas pueden ser aplicadas con éxito para reducir la tasa de error en el caso bilingüe castellano-euskera, se han realizado medidas de separabilidad de locutor e idioma tanto para parámetros MFCC como para estas características prosódicas. Estas medidas se detallan en la sección 4.

Por tanto, la solución propuesta utiliza dos tipos de parámetros: Espectrales y prosódicos. Como representativa de la información espectral se ha seleccionado la parametrización MFCC de 18 componentes, junto con sus primeras y segundas diferencias. Se ha calculado un vector MFCC cada 10 ms y se ha aplicado normalización de media y varianza (MVN) sobre cada grabación con el objetivo de reducir efectos de canal.

Las características prosodicas utilizadas han sido los valores de entonación y energía absoluta extraídos cada 10 ms, junto con sus primeras y segundas derivadas. También se ha aplicado MVN con el objetivo de reducir la gran variabilidad entre sesiones que presentan este tipo de parámetros. Esta aproximación hace posible concatenar los vectores MFCC anteriormente calculados y los valores prosódicos, combinando fácilmente ambos tipos de parámetros. Así pues, la parametrización propuesta consiste en vectores MFCC con valores de entonación y energía añadidos. Puesto que no existe información de entonación en las tramas sordas, esta parametrización se realiza utilizando sólo las tramas sonoras.

3 Descripción de la Base de Datos

Para los experimentos se utilizó una nueva base de datos bilingüe castellano-euskera [10]. Esta base de datos contiene grabaciones de 22 locutores bilingües (11 hombres y 11 mujeres) en un entorno semi-silencioso. Las grabaciones se realizaron con un micrófono Plantronics DSP-400, utilizando una frecuencia de muestreo de 44,1 kHz y 16 bits por muestra. Cada locutor realizó cuatro sesiones de grabación espaciadas en el tiempo, con el objetivo de capturar la variación de la voz a lo largo del tiempo. Esta base de datos bilingüe fue adquirida junto con una base de datos biométrica multimodal [11] y el calendario de captura diseñado para esta base de datos biométrica fue utilizado para la nueva base de datos bilingüe. Hay una diferencia de dos semanas entre la primera y segunda sesión, cuatro entre la segunda y la tercera y seis semanas entre la tercera y la cuarta.

Cada locutor grabó 7 secuencias numéricas formadas por 8 dígitos, pudiendo leerlas según su preferencia. Todas las secuencias numéricas son comunes para el castellano y el euskera.

4 Estudio de la Dependencia de los Parámetros con el Idioma

Una parametrización adecuada para el reconocimiento de locutores debe tener una gran variabilidad entre locutores (para permitir discriminarlos) y una reducida variabilidad intra locutor (de forma que la distribución de los parámetros no cambie mucho

entre las condiciones de entrenamiento y prueba). Para verificar que la parametrización propuesta es adecuada, se ha estimado su variabilidad utilizando la divergencia de Kullback-Leibler [12] como medida de distancia entre distribuciones.

Para la variabilidad inter locutor, se ha calculado la divergencia K-L entre todas las posibles parejas de locutores. El valor medio de todas estas medidas es representativo de la divergencia media entre dos locutores cualquiera, y por ello se ha utilizado como estimación de la variabilidad global inter locutor. Este cálculo se ha llevado a cabo de forma separada para el castellano y el euskera.

De forma similar, se ha estimado la variabilidad entre sesiones para cada locutor como la divergencia K-L media entre todas las posibles parejas de sesiones disponibles para ese locutor. La variabilidad inter sesión global se ha estimado como la variabilidad media para todos los locutores. Este cálculo también se ha realizado de forma separada para el castellano y el euskera.

Por último para cada locutor se ha calculado la divergencia K-L entre las parametrizaciones de las grabaciones en castellano y euskera. El valor medio entre todos los locutores se ha utilizado otra vez como una medida global de la variabilidad inter idioma.

La relación entre la variabilidad inter locutor e inter idioma puede usarse como medida de la robustez de una parametrización frente al idioma. Similarmente, la relación entre la variabilidad inter locutor e inter sesión puede usarse como medida de la robustez de la parametrización frente a la sesión. Lo ideal es que estas dos medidas sean tan grandes como sea posible.

Los resultados de estas medidas se resumen en la Tabla 1 para parámetros MFCC tradicionales y la parametrización propuesta. Tal y como se esperaba, al añadir los nuevos parámetros prosódicos a los vectores MFCC se incrementan todas las variabilidades, puesto que incrementar el número de dimensiones de los vectores sólo puede aumentar la distancia entre dos distribuciones. Pero mientras que la variabilidad inter locutor se incrementa en un 30%, la variabilidad inter idioma sólo se incrementa un 12%. Tal y como se refleja en la relación entre la variabilidad inter locutor frente a la variabilidad inter idioma, esto supone un incremento de la robustez frente al idioma de un 15%.

Tabla 1: Divergencia K-L frente al locutor, sesión e idioma para MFCC tradicional y con parámetros prosódicos (MFCC+P), para las grabaciones en castellano (C) y euskera (E).

		MFCC	***MFCC+P***	***Ganancia***
locutor	**C**	6.34	8.25	30%
	E	6.82	8.77	29%
sesión	**C**	3.62	4.81	33%
	E	3.52	4.64	32%
idioma	**-**	4.09	4.61	12%
locutor/idioma	**C**	1.55	1.79	15%
	E	1.67	1.90	14%
locutor/sesión	**C**	1.75	1.72	-2%
	E	1.94	1.89	-3%

La otra cara de la moneda es que las medidas de entonación y energía tienen una gran variabilidad inter sesión, por lo que al menos parte de la ganancia obtenida en robustez frente al idioma se perderá debido a la sensibilidad frente a la sesión. La relación entre la variabilidad inter locutor frente a variabilidad inter sesión se reduce alrededor de un 2-3%. Esto significa que cuando las pruebas se realicen en el mismo idioma que el entrenamiento (es decir, no hay variabilidad inter idioma), es previsible que los resultados sean un poco peores con la parametrización propuesta que utilizando únicamente parámetros MFCC tradicionales.

5 Condición de los Experimentos

Se han utilizado modelos GMM entrenados mediante el algoritmo EM [13], tanto para la parametrización MFCC tradicional como para los vectores MFCC con características prosódicas a corto plazo añadidas. Los parámetros MFCC tradicionales se han entrenado usando tanto tramas sordas como sonoras, mientras que los modelos de MFCC con características prosódicas utilizan sólo las tramas sonoras. También se han evaluado modelos MFCC con sólo tramas sonoras para facilitar la comparación.

Las grabaciones se han diezmado a 8kHz. Se ha utilizado un detector de actividad vocal (VAD) basado en la desviación espectral a largo plazo [14] con el objeto de eliminar las regiones de silencio de las grabaciones antes de aplicar la parametrización. Las cuatro sesiones disponibles en la base de datos se han utilizado en los experimentos en un esquema leave-one-out. El modelo de cada locutor se ha entrenado utilizando dos sesiones completas (aproximadamente 45 segundos de voz), mientras que una tercera sesión se ha utilizado para pruebas de desarrollo, con el objetivo de estimar los meta-parámetros del modelo (en este caso, el número de componentes gaussianas). La cuarta sesión se ha reservado para las pruebas finales. Este procedimiento se ha repetido cuatro veces, cambiando en cada caso la función de cada sesión. Por último, se ha calculado la precisión global del sistema como la precisión media de todas las iteraciones. En los casos en los que se ha realizado un entrenamiento bilingüe (utilizando ambos idiomas en el entrenamiento), se ha tomado una sesión de entrenamiento en castellano y la otra en euskera, de forma que se siguen utilizando dos sesiones de entrenamiento. Esto permite una comparación directa entre los sistemas, ya que todos ellos han sido entrenados utilizando aproximadamente la misma cantidad de voz.

6 Resultados Experimentales

Como referencia la Tabla 2 muestra las tasas de acierto de los modelos GMM con parametrización MFCC tradicional y utilizando el mismo idioma tanto para entrenamiento como para prueba (C=castellano, E=euskera). La precisión del sistema disminuye para un número de componentes gaussianas mayor que 64, a causa del sobreentrenamiento de los modelos debido a la reducida cantidad de material de entrenamiento disponible. La Tabla 3 muestra las tasas de acierto de los modelos de 64 componentes cuando el entrenamiento y las pruebas se realizan con idiomas diferen-

tes. Como puede apreciarse, bajo estas condiciones la precisión se reduce significativamente. También se aprecia que si se realiza un entrenamiento bilingüe, los resultados vuelven a estar cerca de los obtenidos en el caso de un único idioma.

Tabla 2: Tasa de identificación correcta para diferente número de componentes gaussianas utilizando parámetros espectrales con entrenamiento y pruebas realizadas con el mismo idioma

# mix	C-train; C-test	E-train, E-test
2	81.12	79.25
4	89.29	87.93
8	94.05	92.35
16	96.60	95.24
32	97.62	95.41
64	98.34	97.29

Tabla 3: Tasas de identificación correcta para los modelos de 64 componentes gaussianas con entrenamiento y prueba en diferente idioma. CE significa entrenamiento bilingüe.

C-E	E-C	CE-C	CE-E
63.55	67.34	96.77	95.58

Sin embargo, esta solución de entrenamiento bilingüe no es generalizable a idiomas no vistos durante el entrenamiento, y sería preferible utilizar una parametrización más robusta frente al cambio de idioma. La Tabla 4 muestra los resultados obtenidos con la parametrización propuesta, usando sólo tramas sonoras. Con el objetivo de facilitar la comparación, también se muestran los resultados de un sistema MFCC tradicional usando sólo las tramas sonoras. Si se comparan los resultados de la Tabla 2 y la Tabla 3 para modelos de 64 componentes con los valores de la Tabla 4 para parámetros MFCC tradicionales, puede comprobarse que el hecho de descartar las tramas sordas tiene poca influencia en los resultados finales para esta parametrización.

Cuando se añaden los parámetros prosódicos a corto plazo, la precisión del sistema en condiciones de idioma único se reduce ligeramente, tal y como predicen las medidas de variabilidad de la sección 4. Sin embargo las tasas de acierto aumentan significativamente en el caso de usar un idioma diferente para el entrenamiento y las pruebas, debido a que la mejora en la robustez frente al idioma es mayor que la pérdida de robustez frente a la sesión.

Tabla 4: Tasa de identificación correcta para modelos de 64 componentes usando sólo tramas sonoras, para parametrización MFCC y MFCC con parámetros prosódicos a corto plazo. También se detalla el incremento de precisión obtenido al añadir los parámetros prosódicos.

	C-C	E-E	C-E	E-C	CE-C	CE-E
MFCC	97.6	96.8	62.6	67.0	96.6	95.6
MFCC+P	97.1	96.3	71.0	73.0	96.1	94.4
Ganancia (%)	-0.5	-0.5	13.4	8.9	-0.5	-1.3

Cuando se realiza un entrenamiento bilingüe de los modelos la precisión también se reduce un poco en el caso de los parámetros prosódicos añadidos. Estos modelos ya han adquirido una robustez frente al idioma gracias a este entrenamiento bilingüe. Sin embargo, esta robustez sólo es válida para los dos idiomas considerados en el entrenamiento (castellano y euskera), y la precisión del sistema volvería a caer si se utilizara un tercer idioma para las pruebas.

7 Conclusiones

En este trabajo se han estudiado las ventajas de añadir información de energía y entonación a corto plazo a parámetros MFCC para obtener una parametrización más robusta frente al idioma en sistemas de reconocimiento de locutores. En una primera etapa se han estimado las variabilidades frente a locutor, sesión e idioma de estos parámetros. Estas medidas han permitido prever una mejora en la precisión del reconocimiento cuando la prueba se realiza un idioma no visto durante el entrenamiento. Los resultados experimentales confirman esta predicción, mostrando una mejora significativa de la tasa de acierto bajo condiciones de desadaptación de idiomas.

Estos resultados experimentales también muestran una pequeña pérdida de precisión cuando el entrenamiento y las pruebas se realizan utilizando un único idioma, debido a la gran variabilidad inter sesión de los parámetros prosódicos. En cualquier caso, esta pérdida puede ser perfectamente asumible cuando el sistema es utilizado en un entorno multilingüe y no puede realizarse un entrenamiento con varios idiomas, o cuando no es posible conocer de antemano el idioma que el locutor va a utilizar al usar el sistema.

Aunque las características prosódicas a corto plazo mejoran la robustez frente al idioma de los sistemas de reconocimiento de locutor, los resultados todavía están lejos de ser totalmente independientes del idioma. Se necesitan nuevos parámetros o nuevas técnicas de normalización de idioma para poder construir un sistema que mantenga una precisión similar independientemente de los idiomas de entrenamiento y prueba.

Agradecimientos: Este trabajo ha sido financiado parcialmente por el Gobierno Vasco bajo la subvención IE06-185 (proyecto ANHITZ, http://www.anhitz.com) y por la Universidad del País Vasco y EJIE S.A. bajo la subvención EJIE07/02 (proyecto MULTILOK).

Referencias

1. Faundez, M., Satue-Villar, A.: Speaker recognition experiments on a bilingual database. In: IV Jornadas en Tecnologias del Habla (4JTH), pp. 261--264 (2006).
2. Durou, D.: Multilingual text-independent speaker identification. In: Multi-lingual Interoperability in Speech Technology (MIST), pp. 115--118 (1999).

3. Akbacak, M., Hansen, J. H. L.: Language normalization for bilingual speaker recognition systems. In: International Conference on Acoustics, Speech, and Signal Processing (ICCASP), pp. 257--260 (2007).
4. Ma, B., Meng, H.: English-Chinese bilingual text-independent speaker verification. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'04), pp. 293--296 (2004).
5. Paalanen, P., Kamarainen, J. K., Ilonen, J., Kälviäinen, H.: Feature representation and discrimination based on Gaussian mixture model probability densities – Practices and algorithms. Pattern Recognition 39, 1346--1358 (2006).
6. Young,S.: Large vocabulary speech recognition: A review. In: IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 3--28 (1995).
7. Reynolds, D. A., Rose, R. C.: Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models. IEEE transactions on Speech and Audio Processing 3, 72--83 (1995).
8. Reynolds, D. A., Campbell, J. P., Dunn, R. B., Gleason, T., Jones, D., Quatieri, T. F., Carl, Q., Sturim, D., Torres-Carrasquillo, P.: Beyond Cepstra: Exploiting High-Level Information in Speaker Recognition. In: Workshop on Multimodal User Authentication, pp. 223--229 (2003).
9. Dehak, N., Dumouchel, P., Kenny, P.: Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification. IEEE Transactions On Audio, Speech, And Language Processing 15, 2095--2103 (2007).
10. Luengo, I., Navas, E., Sainz, I., Saratxaga, I., Sanchez, J., Odriozola, I., Igarza J.J., Hernaez, I.: Building a Basque/Spanish bilingual database for speaker verification. In Workshop Collaboration: interoperability between people in the creation of language resources for less-resourced languages, pp. 23--26, (2008).
11. Galbally, J., Fierrez, J., Ortega-Garcia, J., Freire, M. R., Alonso-Fernandez, F., Siguenza, J.A., Garrido-Salas, J., Anguiano-Rey, E., Gonzalez-de-Rivera, G., Ribalda, R., Faundez-Zanuy, M., Ortega, J.A., Cardeñoso-Payo, V., Viloria, A., Vivaracho, C. E., Moro, Q. I., Igarza, J.J. Sanchez, J., Hernaez I., Orrite-Uruñuela, C.: BiosecurID: a Multimodal Biometric Database. In: MADRINET Workshop, pp. 68--76 (2007).
12. Kullback, S., Leibler, R. A.: On information and sufficiency. Annal of Mathematical Statistics 22, 79--86 (1951).
13. Duda, R. O., Hart, P. E., Stork, D. G.: Pattern Classification. Wiley, John and Sons (2001).
14. Ramirez, J., Segura, J. C., Benitez, C., de la Torre, A., Rubio, A.: Efficient Voice Activity Detection Algorithms Using Long Term Speech Information. Speech Communication 42, 271--287 (2004).

Un Estudio sobre la Identificación de Personas basada en su Movimiento al Caminar (*Gait*)

Ángel Sánchez, Juan José Pantrigo, Alberto Rubio and Jesús Virseda

Departamento de Ciencias de la Computación
Universidad Rey Juan Carlos, C/Tulipán, s/n,
28933 Móstoles, Madrid, Spain
{angel.sanchez, juanjose.pantrigo, a.rubio, j.virseda}@urjc.es

Abstract. Este trabajo presenta un prototipo de sistema basado en el conjunto de la postura y el movimiento al caminar (en inglés, *gait*) con el objetivo de reconocer personas. Esta modalidad biométrica presenta las ventajas de ser poco invasiva y de resultar relativamente fácil capturar las secuencias de datos para la experimentación. En nuestro caso particular, se han obtenido resultados de reconocimiento muy satisfactorios usando secuencias de vídeo muy cortas (en promedio, 51 fotogramas por secuencia), con muy pocos individuos para la experimentación (sólo 6 personas) y también usando pocas características discriminantes (en total, 5).

1 Introducción

El reconocimiento de personas basado en su postura y forma de caminar (*gait*) es un método de identificación de individuos estudiada desde el siglo XIX por la Medicina y la Biomecánica [1]. Desde un punto de vista más general, el patrón de movimiento al caminar de una persona puede indicar algún tipo de patología. Además, este patrón podría ser analizado para el diseño de tipos de calzados y pavimentos deportivos. El análisis del movimiento humano (y del *gait*) desde la Visión Artificial es un área de investigación mucho más reciente. El concepto de biometría basada en *gait* aparece hacia 1994 [2]. Desde entonces esta modalidad conductual de reconocimiento biométrico recibido también una mayor atención desde la Visión Artificial. En 2006 aparece un libro de Nixon, Tan y Chellappa [1] que incluye los principales técnicas, sistemas, bases de datos y trabajos realizados hasta esa fecha dentro de la identificación humana basada en el *gait*. Un trabajo destacado es el de Sarkar y otros [3] que trata de medir la evolución y caracterizar las propiedades del *gait* usando un conjunto de 12 experimentos y una gran base de datos (este abordaje ha sido denominado por los autores de trabajo como *HumanID Gait Challenge Problem*). Otro trabajo a destacar es la propuesta recopilatoria de Boyd y Little [4] que describe los factores que afectan al reconocimiento basado en *gait*, los métodos de evaluación usados y también compara diferentes sistemas de reconocimiento basados en este tipo de biometría. Una propuesta reciente en esta área es la realizada por Boulgouris y

Chi [5], basada en el ajuste (por separado) entre las componentes de la siluetas corporales a reconocer. En dicho trabajo se identifica también la contribución de cada una de las componentes corporales consideradas (p. ej. cabeza, torso, brazos, piernas, etc) en el rendimiento del sistema de reconocimiento propuesto. Entre las bases de datos de secuencia de *gait* presentadas en la literatura, deben destacarse las siguientes por su tamaño y variabilidad: *CASIA Gait Database* [6], *Southampton Human ID at a Distance Gait Database* [1] y *UMD Surveillance Data* [7].

Nuestro trabajo presenta un estudio preliminar y un prototipo de sistema de reconocimiento biométrico basado en la postura y la forma de caminar. En este caso particular, y al tratarse de un primer abordaje al problema, se ha creado una base de datos propia que contiene exclusivamente secuencias de vídeo grabadas en interior y de corta longitud, con sólo 6 secuencias para la experimentación (una secuencia de vídeo por individuo) y pocas características discriminantes (en total, 5). Con estos datos se ha conseguido identificar satisfactoriamente al 100% de los individuos analizados.

2 Descripción de la solución propuesta

La Figura 1 describe gráficamente las etapas seguidas para resolver el problema de reconocimiento planteado. En las secciones sucesivas se explican con detalle cada una de las etapas consideradas.

3 Preproceso de los datos

La etapa de preproceso puede descomponerse en la secuencia de pasos intermedios que se describen a continuación.

- **Captura y procesado inicial del vídeo**
  Para este trabajo se han usado secuencias de vídeo propias, por ello es importante una adecuada captura de los datos, ya que el resto del proceso dependerá de la fiabilidad y facilidades que nos proporcione este paso. Se ha trabajado con secuencias de vídeo cortas (en promedio de 51 fotogramas por captura de individuo con, al menos, un ciclo completo de paso) y las grabaciones se han realizado en un estudio fotográfico mediante un esquema de iluminación basado en contraluz, con un fondo blanco y un foco dirigido hacia el individuo caminando, con la idea de obtener su silueta en negro y el fondo totalmente blanco. El procedimiento de captura favorece a la segmentación por umbralización para separar en cada fotograma el sujeto del fondo. Cada secuencia de vídeo capturada se tiene en formato AVI y se ha extraído la componente azul del vídeo, utilizándose como imagen en escala de grises, ya que esta componente del espacio de color es la que mejor representa el contraste de luces. Un ejemplo de imagen en escala de grises obtenida por el procedimiento explicado aparece en la Figura 2.a.

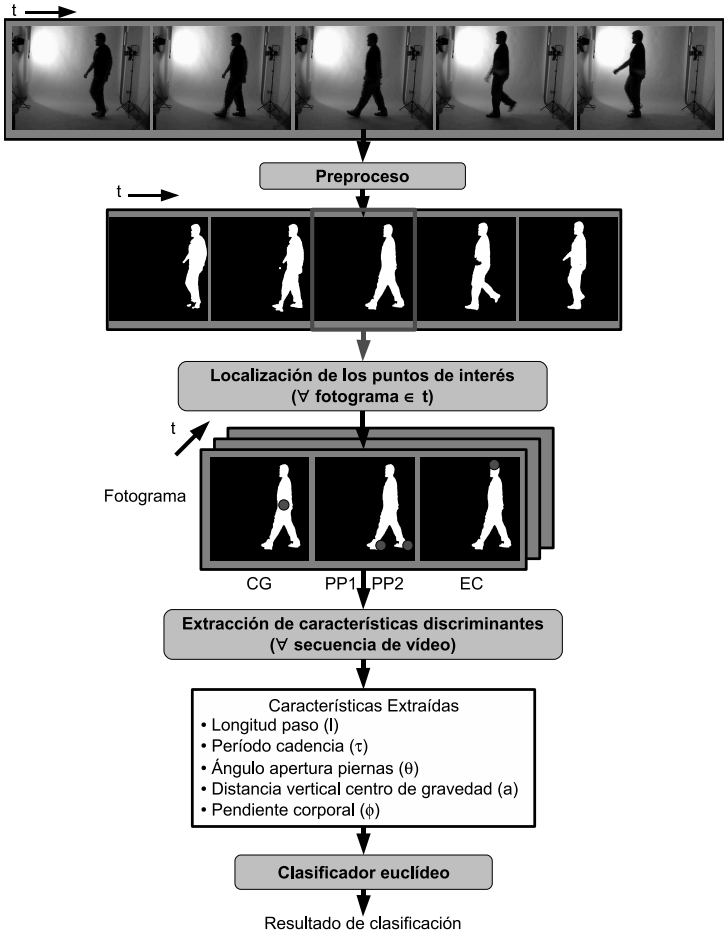


Fig. 1. Arquitectura del sistema propuesto, basado en *gait*.

– **Umbralización adaptativa de los fotogramas**
A pesar de la calidad de las tomas realizadas, y debido a la autorregulación por parte del diafragma de la cámara, la intensidad lumínica en la secuencia del vídeo no es uniforme para toda la escena, por lo que no podemos establecer un umbral fijo de binarización desde el comienzo. Por este motivo, y gracias al contraste y equilibrio que hay entre las zonas claras y obscuras en cada fotograma, se puede utilizar la media del valor de todos los píxeles para fijar un buen umbral para la binarización de los fotogramas. Esta idea, sin embargo, no ha producido resultados correctos ya que, en general, la proporción de zonas obscuras (que originalmente representan individuos y que se quieren separar del fondo) es sólo de un tercio. Por lo tanto, multiplicando el valor medio de los niveles de intensidad en la imagen por un tercio,

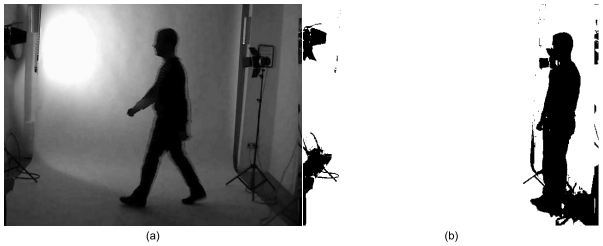


Fig. 2. (a) Imagen en escala de grises obtenida de la componente azul del fotograma y (b) resultado de la binarización adaptativa para un fotograma ejemplo.

obtenemos el correcto valor del umbral. El resultado de esta umbralización para un fotograma ejemplo aparece en la Figura 2.b.

- **Sustracción de fondo aplicada a cada fotograma**
  Para la perfecta localización del sujeto, es necesaria la eliminación del fondo del fotograma, ya que las irregularidades que introduce sobre el sujeto pueden complicar la obtención posterior de puntos de interés, añadiendo ruido e imperfecciones. La substracción de fondo se realiza para cada fotograma del vídeo con el fondo de la escena (es decir, usando una misma imagen sin el sujeto caminando).
- **Eliminación de ruido**
  En esta etapa se trata de "limpiar" la silueta del sujeto, que contiene una serie de líneas horizontales resultantes del entrelazado introducido durante la conversión del formato propio de la cámara al formato AVI. Además, aparece un ruido en el fondo correspondiente a unos puntos blancos donde estaban situados los focos. Este ruido se debe a una inevitable vibración de la cámara a causa del propio paso del sujeto y a variaciones de intensidad durante la secuencia de vídeo. Para eliminar este ruido, se ha aplicado una sencilla operación de apertura morfológica. El resultado de la substracción de fondo junto con la eliminación de ruido se muestran en las Figuras 3.a y 3.b, respectivamente.
- **Recorte y selección de fotogramas**
  Para conseguir secuencias donde aparezcan en todos los fotogramas la silueta completa del individuo caminando, se eliminan manualmente algunos fotogramas al principio y al final de cada secuencia grabada. Estas nuevas secuencias contienen, al menos, dos o tres pasos (por lo que se dispone, al menos, de un ciclo completo de paso por vídeo). Un fotograma correctamente recortado y completo dentro del ciclo de un paso aparece en la Figura 4.a.

4 Obtención de puntos de interés

La etapa de detección de puntos de interés permitirá posteriormente calcular las características discriminantes consideradas. Los puntos de interés necesarios y su proceso de detección se describen a continuación.

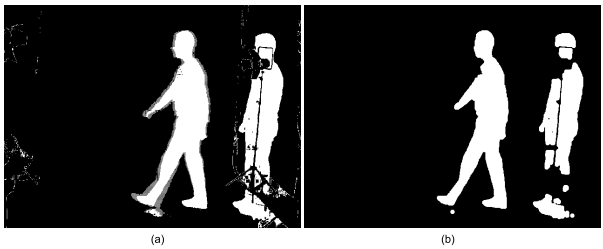


Fig. 3. (a) Resultado de la substracción de fondo y (b) de la eliminación de ruido para un fotograma ejemplo

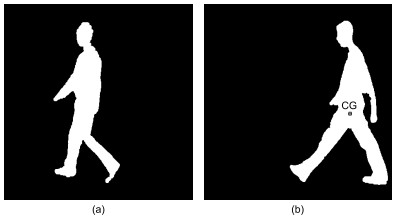


Fig. 4. (a) Silueta recortada para un fotograma ejemplo y (b) localización del centro de gravedad, marcado en el centro de la cadera.

- **Centro de gravedad (punto CG)**
  Este punto permite localizar al sujeto dentro de cada fotograma y nos indica dónde está el centro de la cadera del mismo en cada instante. Su cómputo se realiza a partir de la media de las abscisas y de las ordenadas, respectivamente, de todos los puntos de la silueta del sujeto (píxeles blancos). El resultado de la localización de este punto para un fotograma ejemplo aparece en la Figura 4.b.
- **Plantas de los pies (puntos PP1 y PP2)**
  Para la detección de las posiciones aproximadas de cada pie, se obtiene la imagen especular con respecto al eje de abscisas del fotograma considerado. A continuación, se trazan, por cada una de las columnas de la imagen, líneas descendentes en vertical hasta tocar algún píxel de la silueta invertida del sujeto, obteniéndose un tipo de histograma como el de la Figura 5.a para el ejemplo considerado. De esta manera, un pie se encuentra como el primer máximo relativo que supere un cierto umbral experimentalmente calculado, buscando desde la parte izquierda del histograma. Análogamente, el otro pie se localiza de forma similar pero comenzando ahora la búsqueda desde la parte derecha del histograma.
  El problema que presenta esta solución es, básicamente, que ante la presencia de falsos máximos relativos producidos por las manos en algunas ocasiones (véase la Figura 5.b), existe una probabilidad alta de error en la localización

de las plantas de los pies. Sin embargo, se puede refinar la búsqueda, teniendo en cuenta el valor y la posición relativa de los máximos locales para "filtrar" las posiciones correctas de dichos máximos y resolver situaciones como la de la Figura 5.b. En el caso de que los pies estén juntos, el método determinará que las posiciones de ambos pies coinciden el mismo punto, es decir en el único máximo que habrá en toda la imagen. Finalmente, se deshace la simetría especular para obtener la posición correcta de cada uno de los puntos ($PP1$ y $PP2$) en cada fotograma.

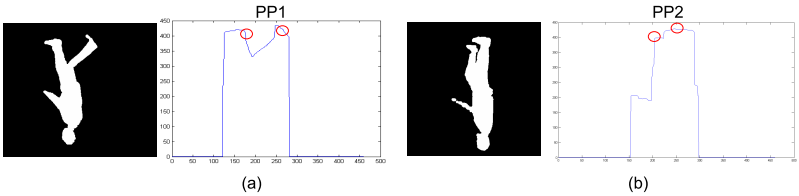


Fig. 5. (a) Contorno invertido del sujeto y búsqueda de los pies y (b) análisis para evitar falsos máximos.

– **Extremo superior de la cabeza (punto EC)**
Se aplica un proceso similar al realizado para determinar la posición de las plantas de los pies. En este caso sólo se busca un único punto y, por una mayor sencillez, no se calcula la imagen especular de la silueta del individuo. El máximo absoluto que se obtiene sobre el histograma calculado determina el punto extremo superior de la cabeza (EC). Un ejemplo de la localización de este punto aparece en la Figura 6.

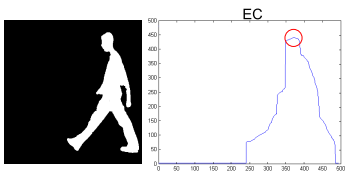


Fig. 6. Localización del punto EC en la cabeza.

5 Obtención de características discriminantes

A partir de la localización automática de los cuatro puntos (CG, $PP1$, $PP2$, EC) extraídos de la silueta del individuo caminando, se procede a calcular un conjunto

de características discriminantes. Para dichas características se elegirá su valor máximo en cada secuencia de vídeo. Todas las características discriminantes consideradas se han calculado para resultar invariantes a escala y a traslación.

- **Longitud máxima (o envergadura) del paso**
  Representa la distancia máxima, en el eje de abscisas, de separación de los pies para una secuencia de vídeo que contiene, al menos, un ciclo de paso. Con el fin de obtener invarianza a escalados, esta característica, expresada en número de píxeles, se normaliza dividiendo por la altura de la silueta del individuo en el primer fotograma. El resultado será un valor en el intervalo [0..1]. La Figura 7 ilustra esta característica con un par de ejemplos.

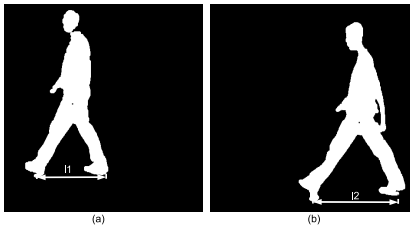


Fig. 7. Envergadura del paso. Dos ejemplos.

- **Periodo de cadencia del paso**
  Si se analiza en una secuencia de vídeo, la distancia existente entre ambos pies en un semi-ciclo del paso (fotogramas que transcurren entre el apoyo consecutivo de cada uno de los pies en el suelo), los valores obtenidos describen un movimiento ondulatorio. Si se representa gráficamente esta distancia de separación de los pies (en píxeles) para un individuo concreto, entre fotogramas consecutivos, se obtiene una curva como aparece en la Figura 8.a. Al existir en esta gráfica pequeñas oscilaciones (causadas por la imperfección en la localización de los puntos), se ha suavizado la curva realizando la media entre el valor de cada punto de la gráfica y el siguiente, obteniéndose como resultado una nueva curva como la de la Figura 8.b. En ella, el valor τ se calcula como el máximo entre dos mínimos locales de la gráfica y define el tiempo (o el número de fotogramas) de la cadencia en un paso del individuo.
- **Ángulo máximo de apertura de piernas al andar**
  Para calcular esta característica angular, representada en radianes, se definen los dos vectores que unen el centro de gravedad del individuo con ambas plantas de sus pies, y se obtiene el ángulo θ que forman ambos vectores. Dicho ángulo se calcula únicamente en el fotograma donde la envergadura del paso es máxima. La Figura 9.a ilustra esta característica.
- **Pendiente de la postura corporal**
  Esta característica angular ilustra la inclinación corporal una persona al andar (véase la Figura 9.b). Se define como el ángulo que forma el eje de

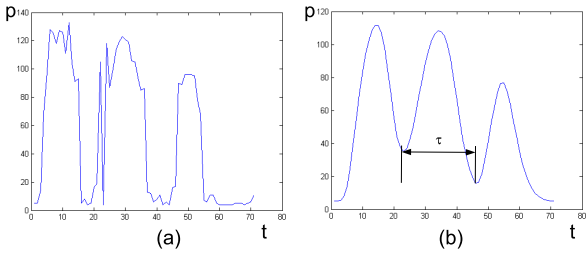


Fig. 8. (a) Variación de la distancia entre los pies en píxeles p en el eje de ordenadas a lo largo de los fotogramas (tiempo t en el eje de abcisas) y (b) suavizado de la curva de distancia (Nota: los valores de las distancias representadas son previos a la normalización).

ordenadas con el vector que une el centro de gravedad CG con el punto más alto de la cabeza EC en la silueta del individuo. Este ángulo permanece más o menos constante durante los fotogramas de la secuencia de vídeo y caracteriza bien a cada persona, ya que ésta ha acostumbrado a su cuerpo a repartir las masas de una determinada forma para que se mantenga el equilibrio al andar. Este ángulo Φ, en radianes, se ha normalizado dividiéndolo por su módulo.

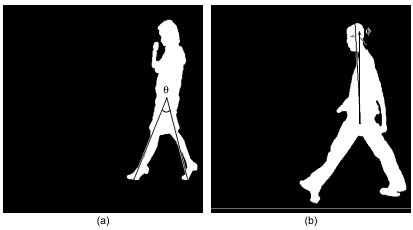


Fig. 9. (a) Ángulo máximo formado por ambos pies con el centro de gravedad corporal y (b) Pendiente corporal de un individuo.

– **Variación máxima (en vertical) del centro de gravedad**
Como al caminar la cadera realiza un movimiento ondulatorio, el centro de gravedad del individuo también lo hará. Se ha calculado la variación en el eje de ordenadas de la posición del punto CG durante todos los fotogramas de cada secuencia de vídeo (ver Figura 10). El valor máximo a de esta variación, dada en píxeles, se ha normalizado de manera similar a la envergadura del paso (es decir, dividiendo por la altura en pixeles de la silueta del individuo en el primer fotograma), obteniéndose un resultado en el intervalo [0..1].

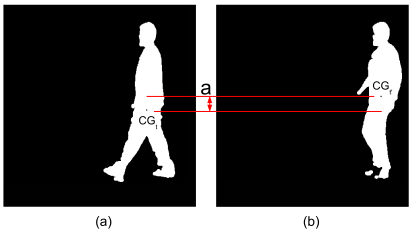


Fig. 10. Amplitud del centro de gravedad.

6 Resultados de clasificación

Nuestra experimentación se ha realizado sobre un conjunto de seis individuos (cinco hombres y una mujer), grabándose una única secuencia de vídeo por persona. Los valores calculados para cada una de las cinco características discriminantes, descritas en la sección previa, y para cada uno de los seis individuos participantes en nuestros experimentos, aparecen en la Tabla 1.

Puede observarse que los valores de las características permiten claramente diferenciar (o clasificar) a los individuos. Por ello, se ha usado un sencillo clasificador euclídeo basado en distancias entre clases con el fin de discriminar a los sujetos. Para nuestra base de datos se ha conseguido reconocer al 100% de los individuos.

Table 1. Cabecera de la tabla.

Individuo	**MaxEnv** (l)	**Periodo** (τ)	**MaxAng** (θ)	**AmpCG** (a)	**Pendiente** (Φ)
1	0.327	19	0.7413	0.057	-0.3333
2	0.330	14	0.7298	0.037	0.3333
3	0.358	13	0.7759	0.041	0.4999
4	0.451	16	1.0057	0.049	0.0832
5	0.353	17	0.7139	0.057	0.4999
6	0.412	14	0.8918	0.062	0.1110

7 Conclusiones y trabajo futuro

Este trabajo presenta un prototipo inicial de sistema de reconocimiento biométrico basado en la postura y el movimiento corporal al caminar (*gait*). En general, el rendimiento de sistemas biométricos reales basados en *gait* está por debajo de lo requerido para su uso como modalidad biométrica [4]. Ello nos ha motivado para comenzar a explorar esta técnica de identificación personal. A pesar de la relativa simplicidad del prototipo desarrollado, consideramos que las características

extraídas en este estudio son fundamentales (y suficientes) para distinguir a dos personas caminando. También se ha observado que un mismo sujeto en condiciones anímicas distintas puede no resultar identificable como él/ella mismo/a. Para llegar a esta conclusión de forma experimental pedimos a un individuo que repitiese su toma pero simulando estar en un estado anímico diferente. Esta repetición produjo valores claramente diferenciados para las características discriminantes consideradas en el estudio.

Un trabajo futuro es ampliar la experimentación para un conjunto mucho mayor de individuos y grabando varias secuencias de vídeo por cada individuo. También es interesante el uso de bases de datos de *gait* más estandarizadas y extensas como por ejemplo la *CASIA Gait Database* [6], donde además las secuencias de vídeo corresponden a sujetos caminando en el exterior (en nuestro caso, todas las secuencias usadas han sido grabadas en interior). Ello permitiría comparar nuestros resultados con los de otros autores sobre los mismos conjuntos de datos. Además, se requiere ampliar el número de características extraídas así como estudiar su capacidad discriminante en la etapa de clasificación, donde se podrían usar otros clasificadores más complejos que requieren de un proceso de entrenamiento previo (como, por ejemplo, los SVM).

Agradecimientos

Este trabajo ha sido subvencionado por el proyecto de investigación TIN2005-08943-C02-02 del Ministerio de Educación y Ciencia de España.

References

1. Nixon, M.S., Tan, T.N., Chellappa, R.: Human Identification based on Gait. International series on Biometrics, Springer (2006)
2. Niyogi, S.A., Adelson, E.H.: Analyzing Gait with Spatiotemporal Surfaces, Proc. IEEE Workshop on Nonrigid and Articulated Motion (1994) 64-69
3. Sarkar, S. el al.: The human id gait challenge problem: data sets, performance, and analysis. IEEE Trans. Pattern Analysis and Machine Intelligence 27 (2005) 162-176.
4. Boyd, J.E., Little, J.J.: Biometric Gait Recognition. LNCS 3161, Springer (2005) 19-42
5. Boulgouris, N.V., Chi, Z.X.: Human gait recognition based on matching of body components. Pattern Recognition 40 (2007) 17631770
6. Yu, S., Tan, D., Tan, T.: A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition. Proc. of the 18th International Conference on Pattern Recognition (ICPR06). Hong Kong, China (2006)
7. Kale, A., et al.: Identification of Humans using Gait, IEEE Trans. Image Processing (2004) 1163-1173

Classification Accuracy improvements by the use of simultaneous biometric measurements: Hand palm recognition.

Artzai Picón[1], Alberto Isasi[1], Aritz Villodas[1]
[1] **TECNALIA-Infotech**
ROBOTIKER-Tecnalia, Parque Tecnológico, Edificio 202. E-48170 Zamudio (Bizkaia) (SPAIN)
Tel. +34 94 600 22 66, Fax. +34 94 600 22 99

{apicon, aisasi, avillodas}@robotiker.es

Abstract. Biometric systems can achieve good results on their own. Sometimes, the single use of one biometric feature does not suit properly the application requirements because of the fact that the security level has to be extremely high. To achieve this requirement, it is possible to use a multimodal biometric system. This implies that the user must pass through a set of biometric devices to login in each one. This process could be sometimes overwhelming for the users and, for this reason, a new multimodal hand-palm-print identification device is proposed. Within this approach, the acquisition of the different biometric measures is done at the same time.

Keywords: Biometrics, Hand recognition, Machine vision, Pattern recognition, Security, Multimodal Biometrics.

1 Introduction

Biometrics is the technology which can identify and obtain human features using physical characteristics or behavior. These technologies can make a relationship between a person and his pattern in a safe and non-transferable way.

The verification of the person's identity based on the measurements of the palmprint is presented as a safe and very economic method allowing its implementation in lots of Biometric applications.

There are two working modes in a biometric system: Authentication and identification. Authentication consists of verifying whether the person is who he pretends to be (1vs1). Identification consists of obtaining the identity of a person searching in a database so the customer has not to tell the system who he is (1vsN).

To measure the performance of a biometric system False Acceptance Rate and False Rejection Rate terms are used [1][2]. FAR (False Acceptance Rate): it is defined as the point per cent of the users who are accepted by the system when should not have been accepted. FRR (False Rejection Rate): it is defined as the point per cent of the users who are rejected by the system when should have been accepted. EER (Equal Error Rate): it is defined as the threshold in which FAR and FRR are equal so a true user have the same possibilities of being rejected as a fake user of being accepted.

Biometric measurement is a high reliable technique. However, this confidence can change depending on what is being measured. This reliability can range from the security given by face or fingerprint recognition, to the high-confidence given by iris or retina based systems.

Therefore, in some cases, the use of a single biometric feature is not capable to reach the desired security level. To face this, the combination of different measures from different sources (biometrical or not) can really strengthen the security. For example, the user can combine a fingerprint system with a PIN number to get two different inputs or maybe, a high security authentication device can be designed merging an iris recognition system with a password given by voice. The more inputs the system has, the higher the system security will be against authentication errors.

One of the main inconveniences regarding multimodal biometric systems is the fact that the user needs to be measured by all the biometric devices comprising the system. For example, combining an iris and fingerprint system the user has to be captured by the two systems putting his finger in the fingerprint device and, after that, showing his eye to the iris device in order to provide the system the needed biometrics for the recognition.

By this reason, it would be worthy to be able to design a multimodal device, diminishing both the discomfort and the user's rejection caused by this kind of systems.

The proposed way to reduce this discomfort is based on engineering a methodology that could perform multibiometric recognition using a single capture of the biometric features.

This approach has been tested in a proposed prototype which is able to extract different biometric features of the user from his hand. This system obtains the "palm print" through a single capture from an off-the-self scanner. Proposed algorithms extract on one hand, the geometry of the hand, and, on the other hand information about the texture of the image.

In this way, the system copes with biometric processing performing a dual recognition but having only to perform a single capture step. This approach achieves dual systems' robustness but keeping the one-device traditional simplicity.

2 Developed Prototype

The prototype developed consists of a software system which takes charge of the processing and the management of the data obtained from each user using a database, and of a hardware module constituted by a PC Scanner and a compatible PC.

This prototype allows camera or scanner image acquisition, user registration, information centralization in a database accessible from several ID checkpoints.

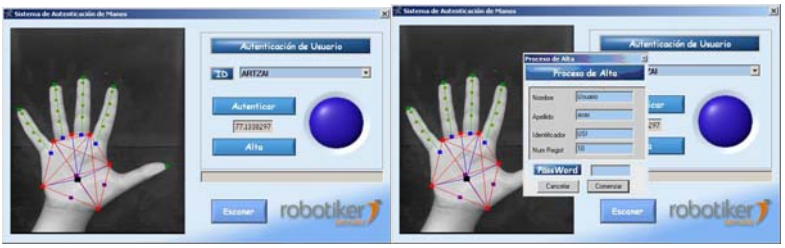


Fig.1 Developed prototype picture.

3. Feature extraction

The main problem in a hand-based recognition system lies in doing the correct election of the features which are going to be used to classify a person. This algorithm extracts a set of characteristics independent to noise and to hand placement.

To obtain these characteristics this method is applied

- Hand image acquisition.
- Hand isolation.
- Detection of points of interest.
- Geometrical characteristics extraction.
- Texture characteristics location.

a. Geometric features extraction

A colour image is directly obtained from the scanner.

Fig. 1. Acquired Image

The method consists of obtaining the hand edges to allow the detection of finger intersections and finger ends through the hand shape curvature.

Fig. 2. Hand Obtained Borders

The hand borders are extracted using classical machine vision algorithms [4]. To get the hand curvature, the whole border array is examined and outer product is used to determine the curvature of each point in the edge image (1).

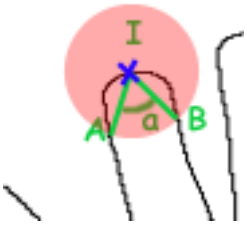


Fig.3. Curvature estimation

$$Curvature = Sin(\alpha) = \frac{\mathbf{IA} \wedge \mathbf{IB}}{|\mathbf{IA}| \cdot |\mathbf{IB}|} \quad (1)$$

After calculating the curvature values in the array it is observed that the finger ends are corresponded to the valleys in the curvature graph and the fingers intersection are corresponded to the peaks of the graph (Fig. 4).

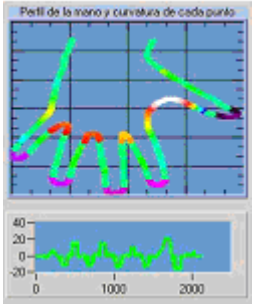


Fig. 4. Curvature graph

Searching for maximum and minimum points in the curvature graph allows extracting these points which are called “main points” (Fig.5).

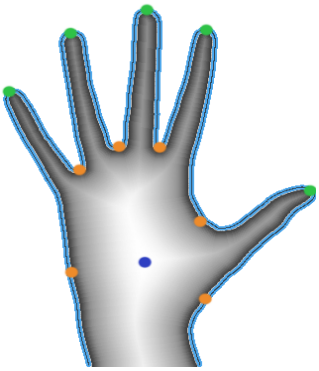


Fig. 5. Main points extracted from the captured image

Using these points, and some characteristics related to the distance of a point inside the hand to the closest border point, we are able to create the feature vector of a hand image.

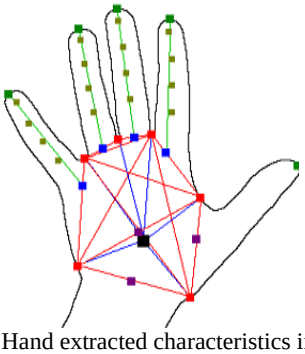


Fig. 6. Hand extracted characteristics image

Finally 48 characteristics are extracted from a hand and used to create a hand feature vector.

$$\mathbf{X} = \{X_1, X_2, \ldots, X_M\}^T \quad (2)$$

b. Texture characteristic extraction

As well as the previous characteristics, a texture feature extraction is needed to perform a texture based classification.

For this purpose, a trapezoidal zone is extracted using the main points coordinates and a transformation is applied to convert this in a rectangle (Fig. 7). The hand texture in the rectangle will be used as new features for hand based classification. In order to

reduce the amount of information contained in the texture, a PCA (Principal component analysis) based reduction method is chosen.

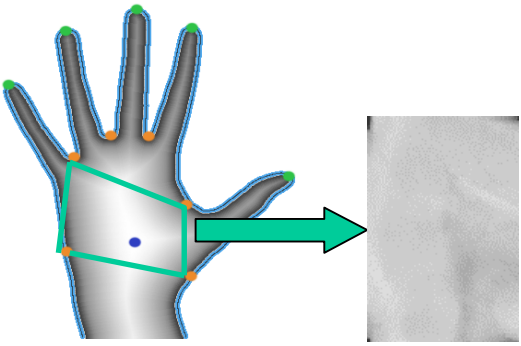


Fig. 7. Hand texture extraction

Based on this method, each texture image is described as a vector, and the mean of these textures is used to generate the mean based PCA vectors as described in [8] for face recognition. This way of generating the PCAs diminishes spectacularly the intraclass differences as low noisy information is not captured by the PCA based on mean vectors.

4. Classifiers

The two types of features extracted are classified by the use of statistical gaussian classifiers. In the case of geometrical features, this is done directly. However, the high dimensionality of the data makes necessary the use of dimension reduction schemes to avoid the Hughes phenomenon. A simple combination of classifiers is proposed for dual classification.

a. Hand geometry classification

Let **X** (2) be an m components feature vector of the geometric distances obtained from one person's hand. And let n be the number of hand images of that person. Assuming a Gaussian distribution of the geometric features, a Gaussian model can be created for each of the classes creating what is known as user's *template*.

The mean vector and the covariance matrix are calculated using each class samples to estimate the gaussian properties of that class.

$$\boldsymbol{\mu}_{ML} = \sum_{n=1}^{N} (\mathbf{x}_n) \tag{3}$$

$$\boldsymbol{\Sigma}_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{ML})(\mathbf{x}_n - \boldsymbol{\mu}_{ML})^T \tag{4}$$

Using that information, a gaussian model of each class is created as shown in (5).

$$\mathrm{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma) = \frac{1}{(2\cdot\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}e^{\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}} \tag{5}$$

After modeling each class, any incoming vector will be checked for classification against all the modeled classes and its belonging probability will be calculated for each of them measuring its distance to each of the classes (6).

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{6}$$

The Δ^2 value indicates the similarity of the present vector with one template of a person. Δ^2 is smaller if the feature vector has been obtained from a hand of the same person the *template* belongs. This allows the use of that value as a measurement of distance or probability.

b. Hand texture classification

Due to its high dimensionality, palm print texture information's dimensionality is reduced using the widely known Karhunen-Loève transform also known as PCA [5, 6, 7]. PCA is defined as the orthogonal projection of the data onto a lower dimensional subspace where the variance of the projected data is maximized.

After transformation, the feature vector is represented by the **X** vector in the reduced dimensionality space defined by the PCA eigenvectors. This representation avoids the curse of dimensionality as well as decorrelates and compresses the information contained in the spectrum allowing the creation of simpler classifiers and smaller training sets.

After transformation, feature vectors are classified in the same way as defined for geometric features.

c. Combined classification

The simple method of voting is used for classification. In this way a user has to be classified correctly by the 2 classifiers to be authenticated.

This increases in an additive way the probability of FRR, but diminishes the probability of FAR in a multiplicative way.

$$P_{ACCEPTED} = P_{ACCEPTED\ Geometry} \cdot P_{ACCEPTED\ Texture} \quad (7)$$

$$P_{REJECTEDD} = P_{REJECTED\ Geometry} \cdot P_{REJECTED\ Texture} \quad (8)$$

5. Experimental Results

Authentication tests were accomplished using 155 hand images from 14 different persons, using 5 images from each person to create the 14 *templates* and the rest of them were used to test the system.

To determine the classifier performance, each test image was crossed against the 14 *templates* observing the distances produced (3) whether they belong to the same person or not.

a. Hand geometric recognition

The obtained performance of the biometric system is shown in Table.1 and in Fig.8.

FAR with no FRR	FRR with no FAR	EER
0.6 %	14%	2.25%

Table. 1. Geometric system accuracy.

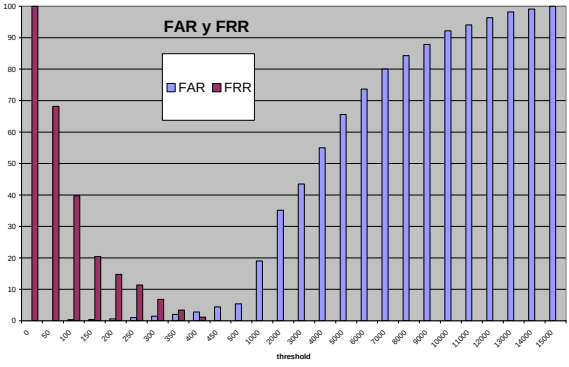


Fig. 8. Geometric system FAR FRR table.

b. Hand texture recognition

The previous classifiers do not use the hand texture information to make their decision. The use of Mahalanobis classifier is proposed to accomplish this.

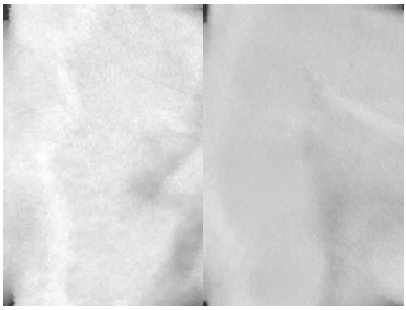


Fig. 9 Hand textures from two different persons

35 hand textures were used against the previous calculated classifiers (as described in [5] for face recognition systems) obtaining a very low error rates.

Mahalanobis distance was used as threshold to accept or reject a person and the results are shown in Table.6 and in figure.11.

FAR with no FRR	FRR with no FAR	EER
0 %	4.16%	0.47%

Table. 2. Texture system accuracy.

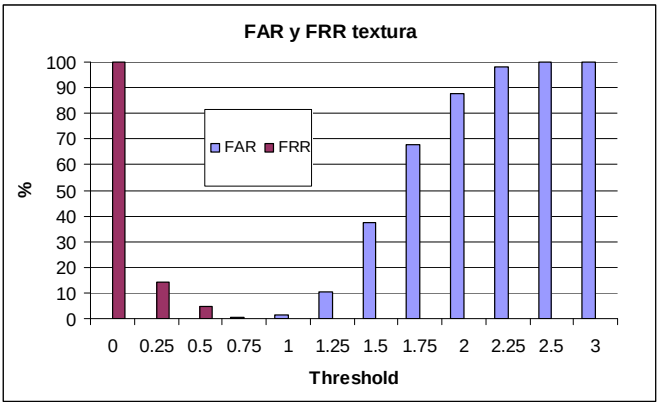


Figure 9. Texture FAR-FRR results table

c. Dual classification

This classification can be tuned by changing the threshold of the two previous classifiers. The idea is to tune each of the individual classifier to its EER so the combined one will be working in the optimal individual parameterization. Taking this into account, the following results were obtained:

Feature	EER
GEOMETRY	2.25%
TEXTURE	0.47 %

Table.3. Individual systems' accuracy

The results obtained by the combined system were so good, being quite close to the theoretical expected results.

FAR	FRR
0,01%	2.81%

Table 4. Dual system's accuracy

6. Result Discussion

The results obtained in geometric features based classification are very promising, similar or better than the ones obtained by other groups as is presented in Bulatov et al [4].

- Jain et.al. FAR of 2% and a false rejection rate (FRR) of 15%.
- Jain and Duta FAR of 2% and FRR 3:5%.
- Raul Sanchez-Reillo et. al. error rates below 10% in verification
- Bulatov et al. FAR 1% y FRR 3 %.

Our results using the geometric features obtain a FAR of 0.6% and a FRR of 14% and changing the threshold sensibility we obtain a result of FAR 2% and FRR 3% as shown in Fig.8.

Hand texture based classification has obtained really amazing results, better than ones obtained using geometrical features.

The results using texture features obtain a FAR of 0% and a FRR of 4.16%, and changing the sensibility of the threshold we get a FAR of 0.47% and a FRR of 0% (as shown in Fig.11) being these results better than other groups obtained ones.

However, the system can obtain a better behavior merging information from different sources, working as a multimodal biometric system.

The combination of the two biometric measures using the simple method of voting (7, 8) obtains even better results and a more robust system against intrusions because of the fact that the source data comes from 2 different biometric characteristics.

What is more, using the combined method, the system achieves a very good performance (0.01% FAR with a 2.81% FRR). It implies a great security level and it is done using just one biometric device performing two different measures at the same time.

7. Future Work

To validate the result it is planned to repeat the tests using a huge number of images and persons.

The texture and geometric features will be combined to get a better performance in identification and authentication tasks.

Due to the good results obtained with texture features, it is planned to extend the present texture area to other hand areas such as fingers.

In the future it would be possible to add, in the same measure acquired now, a new biometric information source: fingerprints.

Taking advantage of the image obtained through the palmprint scanner, user's fingerprints can be also processed to introduce the biometric recognition of the 5 fingerprints of the user into the system.

Normally, it will be not necessary to include all the fingerprints into the user profile; it might be possible to select the ones that will be used to build the user profile.

This selection would be done depending on the desired security necessities. Using the five user fingerprints there will be seven different biometric information sources increasing system security a great deal.

Although fingerprint can be easily observed at good quality in the acquired images, it has to be determined whether this quality is enough to allow fingerprints' minutiae location.

7. Conclusions

Hand texture based classification offers a very good result in authentication, better than the one geometrical methods offer. The combination of both increased the system performance considerably while users' acceptance of the prototype remains quite good, and no complaints have been detected.

According to the results, it can be appreciated that the palm print texture recognition is a bit more reliable than geometric one. However, merging both of them, a much more reliable authentication system is built.

One of the main advantages of this approach is that a dual authentication is performed using a single capture. In this way, the user does not have to use different authentication devices for biometric features extraction (iris, voice, fingerprint...)

As a conclusion, proper design can allow to obtain the same benefits as multimodal biometric devices using a simple device and without overwhelming the users by taking multiple measures in several devices for authentication.

8. References:

1. J. A. Gutierrez et al, *"Estado del Arte: El Reconocimiento Biométrico."* Robotiker, 2002.

2. V. Espinosa Duró, *"Evaluación de sistemas de reconocimiento biométrico"*, Departamento de Electrónica y Automática. Escuela Universitaria Politécnica de Mataró. 2001.

3. R. Sanchez-Reillo, et al, *"Biometric identification through hand geometry measurements"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(10):1168–1171, 2000.

4. Y. Bulatovet et al, *"Hand recognition using geometric classifiers"*, DIMACS Workshop on Computational Geometry, Rutgers University, Piscataway, 2002.

5. Pedro Mª Iriondo, *"Reconocimiento automático de caras basado en el análisis de características físicas y la expresión"*, Tesis Doctoral, Escuela superior de ingeniería de Bilbao, 2000.

6. C. M. Bishop, *"Pattern recognition and machine learning"*, Springer, 2006, ISBN-10: 0-387-31073-8.

7. Anil K. Jain et al. *"Statistical Pattern Recognition: A Review"*, IEEE Transactions on Pattern Analysis and Machine Intelligence vol 22, Nº1 January 2000.

Reconocimiento biométrico entrenando y testeando con diferentes bases de datos de caras

Joan Fàbregas, Marcos Faundez-Zanuy

Escola Universitària Politècnica de Mataró (Adscrita a la UPC)
08303 MATARO (BARCELONA), Spain
fabregas@eupmt.es ,faundez@eupmt.es
http://www.eupmt.es/veu

Resumen. El reconocimiento biométrico presenta un amplio conjunto de fuentes de variabilidad. Si bien algunas bases de datos públicas contemplan diversas variabilidades, se trata de condiciones de laboratorio y por tanto pertenecen a un entorno idealizado. En este artículo presentamos una serie de experimentos en los que los modelos de una determinada característica biométrica se han obtenido con una base de datos y el test se realiza con otra diferente. Para poder llevar a cabo este tipo de experimentos no se calcula un modelo por persona, sino un clasificador con dos clases: iguales y diferentes. Por tanto, durante el test basta con presentar al clasificador la imagen de la identidad declarada y la imagen de test para que el sistema decida si ambas medidas biométricas pertenecen a la misma persona o no. De esta forma, el clasificador puede tratar muestras pertenecientes a personas no usadas durante el entrenamiento. La realización de tests con datos biométricos de personas extraídos de bases de datos distintas de la de test puede ser considerado una prueba más alejada de las condiciones ideales de laboratorio y por tanto, son más realistas de cara a probar la bondad de un sistema en un funcionamiento más real.

1 Introducción

La mayor parte de estudios en reconocimiento biométrico de personas contienen una parte experimental basada en una única base de datos. En el mejor de los casos, proporcionan resultados con dos o tres bases de datos, con la finalidad de demostrar que el algoritmo propuesto puede funcionar satisfactoriamente en diversos escenarios y que no está sobre-entrenado (adaptado a una base de datos en concreto).

En este artículo presentamos una serie de experimentos usando varias bases de datos de forma conjunta. La sistemática consiste en entrenar con una base de datos y testear con otra que contiene personas distintas. Para poder llevar a cabo este tipo de experiencias no entrenaremos un modelo por persona, sino que usaremos un único clasificador con dos clases: iguales (genuino) y diferentes (impostor). Por tanto, el clasificador se entrena para decidir si dos muestras de entrada (una de “entrenamiento” o plantilla modelo y otra de test) pertenecen al mismo usuario.

Si bien en principio este tipo de clasificador está adaptado a problemáticas de verificación (comparaciones 1:1) se puede extender fácilmente a aplicaciones de identifi-

cación (comparaciones 1:N) mediante la verificación contra todos los usuarios presentes en la base de datos. Por tanto, el problema de identificación se convierte en N verificaciones consecutivas.

1.1 La problemática de los sistemas biométricos

Los sistemas de seguridad biométrica [1] ofrecen un buen conjunto de ventajas frente a los sistemas clásicos (passwords, llaves, etc.) Sin embargo, existen una serie de problemáticas no resueltas todavía. En [2] dividimos estos problemas en cuatro categorías principales. En este artículo, presentamos un Nuevo mecanismo que puede aliviar dos de estos problemas:

a) Precisión: Cómo representar de forma precisa y eficiente los patrones biométricos.
b) Escala: Cómo medir de forma repetible y distintiva patrones biométricos de una población amplia.

La mayor parte de la investigación biométrica comienza con la adquisición de una base de datos biométrica o con la utilización de una base de datos disponible, como las descritas en [3]. Esto supone una primera aproximación válida. Solventa el problema de disponer de una cantidad elevada de voluntarios para testear el sistema cada vez que se modifican los parámetros del algoritmo, pero presenta inconvenientes. Algunos de ellos han sido expuestos de forma irónica, por ejemplo, por Naggy en "Candide's practical principles of experimental pattern recognition" [4]. Una regla, que ciertamente tiene que ser evitada por los investigadores honestos, es la siguiente:

a) Teorema: Existe un conjunto de datos para los cuales un algoritmo candidato es superior a cualquier otro algoritmo rival. Este conjunto puede ser construido omitiendo de la base de datos aquellas muestras que son incorrectamente clasificadas por el algoritmo candidato.
b) Precaución "Casey": Nunca pongas tus datos experimentales disponibles a otros investigadores; alguien puede encontrar una solución obvia que tu fuiste incapaz de hallar.

Por consiguiente, la disponibilidad de la base de datos es importante de cara a validar un algoritmo dado, permitir la comparación entre diferentes algoritmos, así como el desarrollo de algoritmos. Este tipo de evaluaciones se conocen como evaluación de tecnología [16]. El objetivo de la evaluación de tecnología es comparar diferentes algoritmos de forma que el test de todos los algoritmos se lleva a cabo sobre una base de datos estandarizada, recogida mediante un sensor "universal". El test se lleva a cabo mediante procesado offline. Dado que la base de datos es fija, los resultados serán repetibles.

Algunos aspectos importantes de la base de datos son:

a) Número de usuarios (Un número elevado de usuarios permite estudiar la capacidad discriminativa de un determinado rasgo biométrico).
b) Número de sesiones de grabación (varias sesiones realizadas en diferentes días permiten estudiar la variabilidad inter-sesión).
c) Número de muestras diferentes por sesión (varias adquisiciones por sesión permiten estudiar la variabilidad intra-sesión).

Una ventaja principal de la disponibilidad de bases de datos es que las condiciones de los experimentos suelen estar fijadas, de forma que se evitan algunos de los errores principales del diseño de sistemas [5]:

a) "Testing on the training set": las puntuaciones del test se obtienen usando los datos de entrenamiento, lo cual es una situación óptima y nada realista.
b) "Overtraining": La base de datos es usada de forma intensiva para optimizar el comportamiento. Este problema se puede identificar cuando un algoritmo concreto proporciona un comportamiento excepcionalmente bueno sobre una base de datos, pero dichas prestaciones no se mantienen al cambiar la base de datos.

Las bases de datos incluyen materiales diferentes para entrenar y testear con la finalidad de evitar el primer problema. Adicionalmente, la disponibilidad de varias bases de datos ayuda a testear el algoritmo sobre nuevos datos y, por tanto, a comprobar si los algoritmos desarrollados por un determinado laboratorio son generalizables (se mantienen al cambiar de base de datos). Por consiguiente, se solventa el segundo problema.

En [13,página 161] se encuentra una afirmación interesante, en el contexto de reconocimiento de firmas on-line [17]: "for any given database, perhaps a composite of multiple individual databases, we can always fine tune a signature verification system to provide the best overall error trade-off curve for that database –for the three databases here, I was able to bring my overall equal-error rate down to about 2.5%- but we must always ask ourselves, does this fine tune make common sense in the real world? If the fine tuning does not make common sense, it is in all likelihood exploiting a peculiarity of the database. Then, if we do plan to introduce the system into the market place, we are better off without the fine tuning." Nosotros hemos tomado en especial consideración esta observación, y para llevarla plenamente a cabo:

- No hemos realizado ningún ajuste fino que, aunque hubiera mejorado los resultados, hubiera proporcionado tasas de error poco o nada realistas.
- Hemos ido un paso más allá: hemos entrenado y ajustado el sistema con una base de datos y hemos realizado el test con dos bases distintas que contienen diferentes usuarios, zooms, panorámicos, rotaciones, dispositivos de adquisición y, en definitiva, diferentes casuísticas.

Usualmente, en los sistemas clásicos de reconocimiento de patrones, existe un número limitado de clases y una cantidad muy elevada de muestras de entrenamiento. Por ejemplo, en el sistema de reconocimiento de dígitos manuscritos en el servicio postal de los Estados Unidos descrito en [6] únicamente hay 10 clases (dígitos) y miles de muestras por clase. En biometría la situación es justamente la contraria, puesto que normalmente tenemos un número elevado de clases (personas) de las que únicamente se toman de tres a cinco muestras durante el entrenamiento [7]. En estas condiciones no hay suficiente número de muestras para entrenar un modelo demasiado sofisticado para cada persona.

En este artículo presentamos:

a) Una aproximación nueva [19] capaz de gestionar un número elevado de clases con pocas muestras por clase.
b) Algunos experimentos de reconocimiento de caras entrenando con una base de datos y testeando con otra diferente.

2 Estrategia de entrenamiento para un número pequeño de muestras de entrenamiento.

En general, el reconocimiento de patrones se puede llevar a cado desde dos puntos de vista distintos [8]:

a) Generativo (también llamado informativo): El clasificador aprende las densidades de probabilidad de la clase, examina la probabilidad de cada clase para producir las características medidas y les asigna la clase más probable. Dado que cada densidad de clase se considera a parte de las otras, el modelo para cada clase es relativamente simple de entrenar. En el caso biométrico, corresponde a un modelo por persona. Únicamente se usan muestras pertenecientes a la persona. En este caso, el principal problema es el pequeño número de muestras disponibles por usuario. Por ejemplo, es típico un valor de 5 fotografías de entrenamiento en un sistema de reconocimiento de caras. Algunos ejemplos incluyen el Linear Discriminant Analysis (LDA) y los Hidden Markov Models (HMM).
b) Discriminativo: El clasificador no modela las densidades de las características de la clase, sino que modela los límites de la clase o las probabilidades de pertenencia a la clase directamente. En el caso biométrico, corresponde a entrenar el clasificador para que aprenda a diferenciar a un usuario de los restantes. Esto significa que el algoritmo requiere muestras de un usuario dado pero también necesita muestras de los otros. En esta aproximación el número de muestras para entrenar un modelo es más elevado, pero la mayor parte de las muestras son inhibitorias (comparativamente existe un número pequeño de muestras pertenecientes a un usuario dado, respecto al número de muestras de los otros usuarios). Estos modelos son más difíciles de entrenar y a menudo suponen el uso de algoritmos complejos. Algunos ejemplo son las redes neuronales y los support vector machines.

En el primer caso (modelos generativos), cuando se quiere añadir un nuevo usuario, basta calcular su modelo asociado. En el segundo caso se debe reestrenar el sistema completo, lo cual requiere tiempo y esto puede ser un serio inconveniente en aplicaciones funcionando en tiempo real. Especialmente cuando las altas y bajas de usuarios son frecuentes. La tabla 1 resume las principales características de las dos aproximaciones [8].

Tabla 1: Comparación entre las aproximaciones generativa y discriminativa al reconocimiento de patrones.

	Generativo	Discriminativo
Presunciones del modelo	Densidades de clase	Fronteras de las clases (funciones discriminantes)
Estimación de parámetros	"fácil"	"Complicado"
Ventajas	Más eficiente si el modelo es correcto.	Más flexible y robusto, puesto que realiza pocas presunciones.
Inconvenientes	Bias si el modelo es incorrecto.	También puede tener bias. Ignora información de la distribución subyacente.

Con la finalidad de evitar estos inconvenientes presentamos un método alternativo denominado 'dispersion matcher' (ajuste por dispersión), que es especialmente útil en los sistemas biométricos. Entrenaremos un único clasificador para resolver la dicotomía: ¿pertenecen estos dos vectores de características a la misma persona? De esta forma resolvemos el problema relativo al número de muestras por clase. Dado que no entrenamos el clasificador con todos los individuos presentes en la base de datos, será capaz de clasificar en un entorno de mundo abierto ("open world"). De hecho, el sistema biométrico, a diferencia de los algoritmos clásicos generativos y discriminativos, no aprende ningún modelo específico para el usuario, y presenta una mayor capacidad de generalización.

Cuando el usuario desea ser autenticado por el sistema, simplemente tiene que presentar su muestra biométrica y el 'dispersion matcher' compara la muestra biométrica a autenticar con las muestras usadas como referencia para dicha persona (adquiridas durante el proceso de enrolamiento). El usuario será aceptado si la fusión de puntuaciones [15] obtenida en cada comparación es mayor que un umbral predefinido. Por ejemplo se puede usar como método de fusión la media de las puntuaciones obtenidas al contrastar la muestra de test con cada una de las muestras de entrenamiento.

Al medir cualquier característica fisiológica de una persona, como por ejemplo la longitud de un dedo, el proceso está sujeto a errores y no siempre se obtiene el mismo resultado. La estadística nos dice que si repetimos la medida diversas veces los valores estarán distribuidos según una distribución gaussiana normal $\mathcal{N}(x \mid \mu_i, \sigma_i^2)$, la cual está caracterizada por la media (μ_i) y la variancia (σ_i^2) de las medidas. En muchas situaciones, la distribución de la media de la característica fisiológica sobre el total de la población es también otra distribución normal, $\mathcal{N}(x \mid \mu_p, \sigma_p^2)$ caracterizada por la media (μ_p) y la variancia (σ_p^2) de la población. El 'dispersion matcher' está basado en el hecho de que la variancia σ_i^2 siempre es menor que σ_p^2. Al plantear la diferencia entre dos muestras de esta medida fisiológica, su valor será normalmente menor cuando se trate de muestras de la misma persona que cuando se trate de personas distintas. Esto se puede representar esquemáticamente en la figura 1, con dos gaussianas: una para las diferencias correspondientes a pares de muestras genui-

nas y otra para pares correspondientes a diferencias de muestras entre impostores y genuinos.

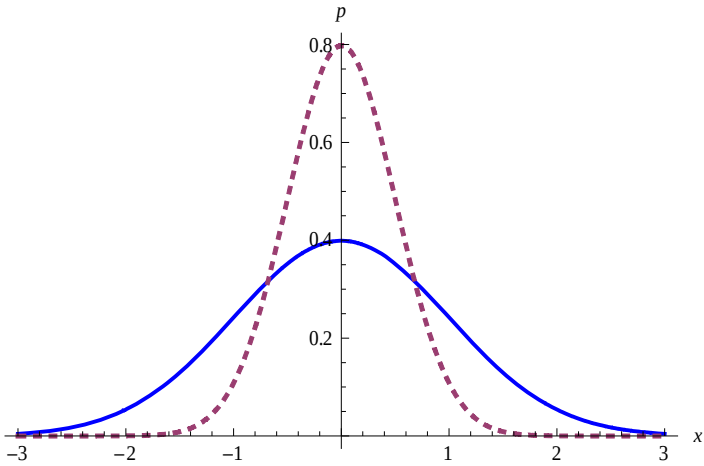


Fig. 1: Ejemplo de dos gaussianas. La punteada corresponde a diferencias de muestras genuinas ($\mathcal{N}(x \mid 0, 0.25)$ en este ejemplo), y la continua a diferencias entre pares de muestras genuinas e impostoras ($\mathcal{N}(x \mid 0,1)$ en este ejemplo).

Para resolver la dicotomía planteada anteriormente (iguales versus diferentes) usaremos un clasificador discriminante cuadrático (Quadratic Discriminant Classifier, QDC) [9], dado que en la práctica las distribuciones de cada característica fisiológica tienen forma de campana y presentan correlaciones lineales.

Para comprender la diferencia entre un clasificador dicotómico y las otras aproximaciones clásicas al reconocimiento de patrones analizaremos un ejemplo sencillo. La base de datos ORL [10] contiene 40 usuarios y 10 capturas por usuario. Una partición típica consiste en usar cinco imágenes por persona para entrenar y las otras cinco restantes para testear. La tabla 2 muestra los datos disponibles para cada tipo de clasificador en esta situación. Se puede observar que la diferencia entre las distintas estrategias radica en el número disponible de muestras para entrenar, mientras que el número de muestras de test es el mismo en todos los casos. Por lo tanto, la significancia estadística de los resultados experimentales es la misma para todos ellos.

Obsérvese que en los sistemas generativo y discriminativo cada clase suele equivaler a una persona, mientras que en un clasificador dicotómico sólo hay dos clases, iguales y diferentes, que se corresponden con el resultado de la verificación biométrica aceptar y rechazar.

Tabla 2: Ejemplo de muestras de entrenamiento y test para una base de datos con *n* individuos y *s* muestras por individuo, usando la mitad de las muestras para entrenar y la otra mitad para testear.

Estrategia	Muestras por clase para entrenar	Muestras por persona para testear	
		genuino	impostor
Generativo	$\frac{s}{2}$ muestras genuinas	$\frac{s}{2}$	$(n-1)\frac{s}{2}$
Discriminativo	$\frac{s}{2}$ muestras genuinas $(n-1)\frac{s}{2}$ impostores	$\frac{s}{2}$	$(n-1)\frac{s}{2}$
Clasificador dicotómico	$n\frac{s(s-1)}{2}$ pares genuino-genuino $n(n-1)\left(\frac{s}{2}\right)^2$ pares genuino-impostor	$\frac{s}{2}$	$(n-1)\frac{s}{2}$

Ciertamente, existen dependencias lineales entre los distintos pares de entrenamiento considerados en el caso del clasificador dicotómico. Sin embargo, dependiendo del tipo de clasificador, podemos sacar partido de estas muestras redundantes (por ejemplo al entrenar una red neuronal, esto supone una alternativa a la estrategia habitual de añadir ruido a los patrones de entrenamiento para incrementar su número) [20]. Para el caso particular del 'dispersion matcher' considerado en este artículo se calculan las matrices de covarianza, que equivalen a considerar los $n\left(\frac{s}{2}-1\right)$ pares genuino-genuino independientes y los $n\frac{s}{2}-1$ pares genuino-impostor independientes.

3. Resultados experimentales de reconocimiento de caras

Una forma de testear un algoritmo concreto de reconocimiento de caras en condiciones difíciles, alejadas de la situación ideal de condiciones de laboratorio, consiste en usar simultáneamente varias bases de datos. Entrenaremos el sistema con una base de datos y haremos el test con otra diferente, que ha sido adquirida en un entorno distinto, con usuarios distintos, etc.

El clasificador se entrena para comprobar si dos muestras de entrada pertenecen a la misma clase (usuario genuino) o no (impostor). Por tanto, al tratar de verificar una persona cuyos datos biométricos no han sido usados para entrenar el sistema, únicamente será necesario introducir la siguiente información dentro del clasificador:

- La muestra o muestras adquiridas durante el proceso de enrolamiento y que pertenezcan a la identidad declarada por el usuario. Estas muestras estarán almacenadas en la base de datos.
- La muestra de test de entrada, que acaba de ser adquirida conjuntamente con la identidad declarada por el usuario.

Será irrelevante si la persona en concreto fue usada o no para entrenar el clasificador, ya que no se asigna ningún modelo a cada persona. Este es el punto clave del algoritmo propuesto que permitirá obtener una mejora más significativa respecto a los modelos clásicos generativo y discriminativo.

3.1 Base de datos

Hemos usado las bases de datos ORL [10], AR [11] y JAFFE [12]. Las figuras 2, 3 y 4 muestran capturas de un usuario de cada base de datos. Las características principales de estas bases de datos son:

a) ORL: 10 capturas diferentes de 40 personas. Para algunos sujetos, las imágenes fueron adquiridas en tiempos diferentes, variando la iluminación, expresión facial (ojos abiertos /cerrados) y detalles faciales (con y sin gafas). Ver la figura 2.
b) AR: 126 individuos, 26 imágenes de cara individuo, tomadas en dos sesiones diferentes, variando la iluminación y la expresión facial. Se han usado 6 de las 26 imágenes, excluyendo las sobre-expuestas y aquellas con oclusiones parciales (gafas, bufandas). Dado que 9 individuos no estaban completos sólo se han usado 117. Ver la figura 3.
c) JAFFE: Contiene imágenes de expresión facial de 10 mujeres japonesas (6 emociones diferentes más la expresión neutra). Ver la figura 4. Las expresiones faciales corresponden a las 6 emociones primarias, o emociones básicas. Existen varias capturas para cada persona y emoción (2 o 3 imágenes).

Fig. 2: Imágenes de muestra del primer usuario de la base de datos ORL.

Fig. 3: Imágenes de muestra de la primera persona de la base de datos AR.

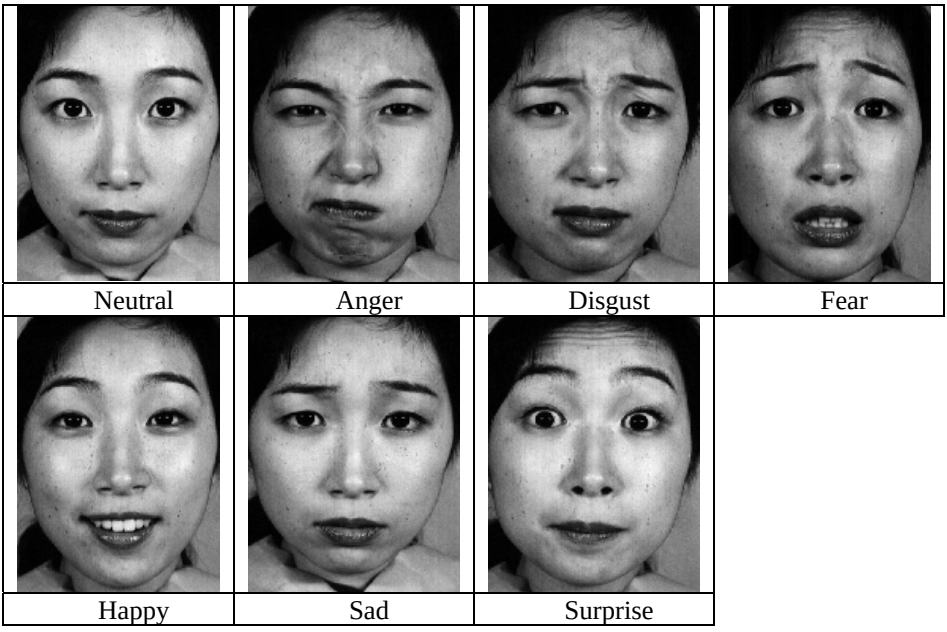


Fig. 4: Imágenes de muestra de la primera persona de la base de datos JAFFE (Japanese Female Facial Expression).

3.2 Experimentos

En primer lugar se realiza una extracción de características, a partir de las imágenes, basada en la DCT. Puede encontrarse en nuestros trabajos previos [14] y [18].

La tabla 3 presenta los resultados experimentales con diferentes combinaciones de entrenamiento y test. Se puede observar que los mejores resultados se obtienen cuando los experimentos se realizan sobre la misma base de datos. Además, los resultados obtenidos al entrenar y testear con bases distintas también son suficientemente satisfactorios, y la degradación de las tasas de reconocimiento es pequeña. Por otra parte, las tasas de reconocimiento son competitivas respecto al estado del arte en sistemas de reconocimiento de caras. Hemos usado tres muestras de referencia por cara y el método de fusión max{·} [15]. Para la base de datos JAFFE se han escogido tres fotos de cara neutra como muestras de enrolamiento, de forma que los tests se realizan en condiciones "difíciles" (existe una expresión facial marcadamente distinta de la usada como plantilla de referencia o modelo de esa persona).

En las aplicaciones de verificación se ha evaluado el mínimo de la función de detección de coste (minimum Detection Cost Function, minDCF) [1], que es una medida parecida al Equal Error Rate. Por consiguiente, cuanto menor sea este valor, mejores son las prestaciones del sistema.

Tabla 3: Minimum detection cost function (%) para algunas bases de datos. A=Anger, D=disgust, F=Fear, H=Happy, Sa=Sad, Su=Surprise. Al entrenar con la base de datos JAFFE se han usado todas las fotografías. Dado que no es realista, no se presentan los resultados de testear con el conjunto de entrenamiento.

	Testing database							
Training database	ORL	AR	JAFFE					
			A	D	F	H	Sa	Su
ORL	2.30	3.08	20.7	17.78	11.5	6.85	5.93	18.15
AR	6.82	3.42	12.0	12.96	8.52	6.48	7.96	16.67
JAFFE	16.5	5.84	--	--	--	--	--	--

Es importante destacar que el clasificador dicotómico propuesto puede gestionar un conjunto importante de situaciones en las cuales los sistemas clásicos experimentan problemas. Se trata, principalmente, de los siguientes aspectos:

a) Puede decidir si dos fotografías (imágenes de modelo y test) pertenecen a la misma persona. No es necesario que la información de dicha persona participara en el proceso de entrenar el sistema.
b) No es necesario reentrenar el sistema al añadir un usuario nuevo. Esto ha sido experimentalmente comprobado a partir del buen comportamiento mostrado al verificar personas que pertenecen a una base de datos no usada durante el entrenamiento del clasificador. Únicamente en algunas expresiones faciales existe un incremento en las tasas de error (principalmente enfado y disgusto).

4. Conclusiones

En este artículo hemos realizado una serie de experimentos consistentes en entrenar y testear un clasificador biométrico con diferentes usuarios extraídos de bases de datos

diferentes. Esto ha sido posible porque no se calcula un modelo por persona sino que se simplifica el problema de clasificación a un problema con únicamente dos clases (genuinos e impostores), de forma que se entrena un "clasificador universal" que nos dice si dos muestras pertenecen a la misma persona o no. Los resultados experimentales son similares a los sistemas reconocedores de caras del estado del arte, e incluso funciona satisfactoriamente cuando existen expresiones faciales no presentes en el ni en el entrenamiento del clasificador ni en las muestras enroladas. Consideramos que nuestra propuesta, al poder gestionar problemáticas de gran variabilidad, es adecuada para entornos de aplicaciones reales.

AGRADECIMIENTOS

Este trabajo ha sido financiado por FEDER y MEC, TEC2006-13141-C03-02/TCM

Referencias

1. M. Faundez Zanuy "Biometric security technology" IEEE Aerospace and Electronic Systems Magazine, Vol.21 nº 6, pp.15-26, June 2006.
2. M. Faundez-Zanuy "Biometric recognition: why not massively adopted yet?" IEEE Aerospace and Electronic Systems Magazine. Vol.20 nº 8, pp.25-28, August 2005.
3. M. Faundez-Zanuy, J. Fierrez-Aguilar, J. Ortega-Garcia y Joaquin Gonzalez-Rodriguez "Multimodal biometric databases: an overview". IEEE Aerospace and electronic systems magazine. Vol. 21 nº 9, pp. 29-37, August 2006.
4. Nagy G. "Candide's practical principles of experimental pattern recognition". IEEE Trans. On Pattern Analysis and Machine Intelligence Vol. 5 No. 2, pp.199-200, March 1983.
5. Bolle R. M., Ratha N. K., Pankanti S., "Performance evaluation in 1:1 Biometric engines". Springer Verlag LNCS 3338, pp.27-46 S. Z. Li et al. (Eds.) Sinobiometrics 2004.
6. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer, 2001
7. R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior, Guide to Biometrics, Springer, 2004
8. Y.D. Rubinstein, T. Hastie, Discriminative vs Informative Learning, Knowledge Discovery and Data Mining, 1997, pp. 49-53.
9. R.O. Duda, P.E. Hart, D.G. Strork, Pattern Classification, Second Edition, Wiley-Interscience, 2001.
10. F. Samaria & A. Harter "Parameterization of a stochastic model for human face identification". 2nd IEEE Workshop on Applications of Computer Vision December 1994, Sarasota (Florida).
11. A. M. Martinez "Recognizing Imprecisely Localized, Partially Occluded, and Expression Variant Faces from a Single Sample per Class IEEE Transaction On Pattern Analysis and Machine Intelligence, Vol.24, N.6, pp. 748-763, June 2002
12. Lyons M., Akamatsu S., Kamachi M., and Gyoba J. "Coding Facial Expressions with Gabor Wavelets". In Third IEEE International Conference on Automatic Face and gesture.
13. Biometrics, personal identification in networked society. Edited by Anil K. Jain, Ruud Bolle and Sharath Pankanti. Kluwer academic publishers 1999.

14. Marcos Faundez-Zanuy, Josep Roure-Alcobe, Virginia Espinosa-Duró, Juan Antonio Ortega “An efficient face verification method in a transformed domain” Pattern recognition letters. Vol.28/7 May 2007 pp.854-858.
15. Faundez-Zanuy M. “Data fusion in biometrics”. IEEE Aerospace and Electronic Systems Magazine. Vol. 20 nº 1, pp.34-38, January 2005
16. Mansfield A. J., Wayman J. L., “Best Practices in Testing and Reporting Performance of Biometric Devices”. Version 2.01. National Physical Laboratory Report CMSC 14/02. August 2002.
17. Faundez-Zanuy M. “Signature recognition state-of-the-art”. IEEE Aerospace and Electronic Systems Magazine. Vol.20 nº 7, pp 28-32, July 2005.
18. Faundez-Zanuy M, Fabregas J., “On the relevance of facial expressions for biometric recognition” Aceptado para su publicación en ICTAI 2007, Patras, LNCS Springer.
19. Fabregas J., Faundez-Zanuy M. “Biometric dispersion matcher” Pattern Recognition. doi:10.1016/j.patcog.2008.04.020
20. Faundez-Zanuy M., “On the usefulness of almost-redundant information for pattern recognition”. Nonlinear Speech Modeling, LNAI 3445, pp. 357-364. Springer 2005.

Sistema de autenticación biométrica sin contacto basado en la geometría de la mano para entornos operacionales

Aythami Morales, Miguel A. Ferrer, Carlos M. Travieso, Jesus B. Alonso

Centro Tecnológico para la Innovación en Comunicaciones (CeTIC)- Departamento de Señales y Comunicaciones - Universidad de Las Palmas de Gran Canaria, Las Palmas 35017, Spain,
amorales@gi.ulpgc.es, mferrer@dsc.ulpgc.es
WWW home page: http://www.cetic.eu

Resumen En este artículo se presenta un sistema de autenticación biométrico sin contacto basado en la geometría de la mano. El sistema está desarrollado para su utilización en entornos operacionales, con condiciones ambientales y de fondo no controladas. Se generó una base de datos formada por más de 4000 imágenes adquiridas durante 3 meses a partir de la cual validar el sistema.

1. Introducción

La biométrica juega un papel cada vez más importante en los sistemas de autentificación y de identificación. El reconocimiento biométrico permite la identificación del individuo basándose en las características físicas o de comportamiento del usuario. Se han desarrollado muchas tecnologías: huella dactilar, iris, cara, voz, firma, geometría de la mano entre otras. Este último método se basa en un estudio de la forma de la mano y presenta algunas ventajas respecto a otras tecnologías. En primer lugar, respecto al dispositivo de captura, a diferencia de por ejemplo el iris, bastará con un dispositivo de captura de bajo coste tipo webcam o sensor CCD. La resolución necesaria es considerablemente menor a la necesaria en sistemas basados en huella dactilar. Por otra parte, los sistemas basados en geometría de la mano presentan una alta aceptabilidad por parte de los usuarios [1].

La mayor parte de los sistemas biométricos basados en la mano requieren de pegs o superficies de contacto. Esto causa problemas relacionados sobre todo con la higiene y la necesidad de limpiar el dispositivo para su correcto funcionamiento. En este artículo proponemos un sistema biométrico sin contacto basado en la geometría de la mano.

El sistema está compuesto por una TabletPC, una cámara y un sistema de iluminación. Los usuarios deben situar la mano en el espacio libre delante de la cámara. En este tipo de sistemas existen dos principales problemas que afrontar: problemas de segmentación asociados a un entorno operacional y las distorsiones de proyección asociadas a la ausencia de contacto.

Trabajando en entornos operacionales y condiciones de iluminación no controladas, la segmentación no es una tarea trivial. Se trabaja con fondos heterogéneos donde técnicas basadas en el color o la forma tienen dificil aplicación. El sistema debe ser robusto ante cambios de iluminación, ya sea natural o artificial. Por último, la carga computacional debe ser baja para que el sistema funcione en tiempo real.

En este artículo se propone el uso de iluminación infrarroja para solucionar los aspectos relacionados con la segmentación y la robustez del sistema ante cambios ambientales. Una plantilla mostrada por pantalla servirá para guiar al usuario a colocar correctamente la mano y así reducir las distorsiones asociadas a los cambios de proyeccción.

Así pues, en la siguiente sección se hará un repaso del estado del arte de los sistemas biométricos basados en mano. Posteriormente se expondrá nuestra propuesta: se abordará la segmentación en la sección 3, la extracción de características geométricas será descrita en la sección 4. La sección 5 describirá el proceso de verificación utilizado y la sección 6 mostrará los resultados experimentales obtenidos. El artículo se cierra con conclusiones, agradecimientos y las referencias.

2. Estado del arte

Tradicionalmente, los sistemas biométricos basados en la geometría de la mano se fundamentan en el análisis de la forma de la mano. La forma se caracteriza por medidas geométricas, por el contorno de la mano o ambos. Las medidas geométricas incluyen medidas de longitudes y de anchuras de los dedos, el grueso de los dedos y de la palma, y las anchuras de la palma entre otras. El contorno de la mano está formado por el límite de la mano entera o por los límites de los dedos. En trabajos de investigación recientes, Tantachun [2] representa un patrón de la mano por un eigenhand obtenido del análisis de componentes principales (PCA) o una malla construida a partir de puntos característicos. Existen diferentes técnicas propuestas para obtener y representar matemáticamente estas características [3].

De forma intuitiva, determinadas medidas geométricas de algunas regiones particulares de la mano se pueden utilizar para caracterizarla. Las regiones utilizadas deben ser las mismas para cada mano. Esto requiere que la mano se sitúe siempre de una forma similar. El correcto posicionamiento se consigue normalmente gracias a la utilización de pegs o topes sobre una superficie de apoyo en la que se sitúa la mano. Jain [4] desarrolló un sistema con las características comentados. Se utilizaron cinco topes para dirigir la colocación de la mano del usuario. Se capturaban imágenes del dorso y del lateral de la mano. Utilizó varias medidas geométricas, incluyendo anchos, largos, y gruesos de los dedos, además de anchos de la palma en diferentes regiones. Con 16 medidas geométricas, consiguió una tasa de error (EER) del 6 %. Sanchez-Reillo [5], [6] utilizó seis topes en su sistema basado en geometría de la mano. Tomaron 25 medidas geométricas de la mano de cada usuario, incluyendo ancho de los dedos y la palma, los

gruesos, desviaciones de los dedos, y ángulos obtenidos a partir de las puntas y los valles de los dedos. El error (EER) obtenido fue inferior al 3 %.

Los topes proporcionan ciertas garantías en cuanto a la obtención de las medidas, pero también presentan algunos inconvenientes:

- Los topes pueden deformar la forma de la mano. Por lo que la fiabilidad de las medidas geométricas basadas en el tamaño o contorno decae. Esta deformación de la forma de la mano afecta sobre todo a la variabilidad intra-clase, que da lugar al falso rechazo [7].
- Los topes agregan complejidad al dispositivo. Los supervisores del sistema y los usuarios deben estar bien entrenados para cooperar con el sistema. Esto aumenta la responsabilidad de los usuarios, degradando así la confiabilidad del sistema.
- Los dispositivos basados en el contacto cada vez sufren mayor rechazo debido fundamentalmente a cuestiones relacionadas con la higiene y la salud pública.

Tras los sistemas basados en topes comenzaron a aparecer sistemas libres de ellos, donde la el usuario situa la mano libremente. Los sensores CCD y las cámaras dieron paso a los escaners. Ofreciendo una mayor resolución y unas condiciones de iluminación más constantes. La eliminación de los topes otorga cierta libertad del movimiento. Para solucionar los problemas asociados a la libertad de posición se utilizan puntos característicos de la mano como puntas y valles para normalizar las imágenes. Wong y Shi [7] propusieron un sistema de autentificación libres de topes basado en la geometría de la mano. Se midieron diversas características geométricas de la palma y los dedos. El índice de aceptación genuino del sistema se situó en 88.9 % y el índice de aceptación falso en 2.2 % con 30 características de la mano. Bulatov [8] midió 30 distancias geométricas de la mano. Como características, se añadieron círculos que caracterizaban la forma de la mano. Los radios, los perímetros, y las áreas de los círculos, junto con las longitudes y los anchos de los dedos, fueron medidos. Alcanzaron una FAR del 1 % y un FRR del 3 %. Boreki [9] y Heshemi [10] midieron las longitudes y los anchos de cada dedo individualmente. Boreki utilizó curvaturas a lo largo del contorno de la mano. Sus resultados se asemejaron a los de Bulatov, con una FAR de 0.8 % y un FRR de 3.8 % a partir de una base de datos formada por 360 imágenes de 80 usuarios.

Los sistemas sin topes dieron paso a los sistemas sin contacto. En los que la mano se situaba en un espacio libre delante del sensor que captura las imágenes. Haeger [11] en su sistema, adquirió las imágenes de la mano en un espacio libre. El centro de figura de una mano dividida en segmentos fue detectado. Utilizando 124 medidas geométricas de los dedos, alcanzó una tasa de falsa aceptación de 45.7 % y una de falso rechazo 8.6 %. Las bajas prestaciones se debieron principalmente a los problemas asociados a la distorsión de la proyección.

Otros trabajos de investigación utilizaron diferentes formas de parametrizar la mano. Garrison [12] desarrolló un sistema de autentificación sin contacto. Se utilizó la transformada PCA para caracterizar a los usuarios. Este método decae mucho ante cambios de posición debido a la distorsión de la perspectiva

en la forma de la mano. Doi y Yamanaka [13] utilizaron una cámara infrarroja CCD para capturar las imágenes. Crearon una malla a partir de entre 20 a 30 puntos característicos extraídos de los pliegues principales de los dedos y de la palma. Utilizaron la desviación de la media cuadrática (rms) para medir la distancia entre las mallas. Al igual que los sistemas vistos con anterioridad, el método dependía mucho de la correcta posición de la mano del usuario. Zheng [3] presentó en su trabajo una serie de descriptores basados en relaciones geométricas entre puntos característicos del interior de la mano. Su sistema parece solucionar en gran medida los problemas asociados a la distorsión proyectiva. Un EER de 0 % sobre 52 imágenes de 23 usuarios diferentes capturadas en un entorno controlado fue el resultado obtenido.

Aunque ya existen diferentes estudios acerca de sistemas sin contacto, no se tiene constancia de su uso en entornos operacionales. Los sistemas presentados hasta el momento se basan en estudios realizados en laboratorios bajo condiciones controladas. Aspectos como la segmentación con fondos altamente heterogéneos o problemas asociados a la robustez ante cambios de iluminación han sido poco tratados. Estos son los aspectos en los que se centra el sistema propuesto en este artículo.

3. Segmentación

El dispositivo de captura usado en el sistema propuesto es una webcam. En términos de resolución existen mejores opciones. Se escogió este dispositivo debido a su bajo coste y a la no necesidad de imágenes de alta resolución.

El sistema trabaja en tiempo real con secuencias video y es necesaria una segmentación rápida. Dado que el sistema es colaborativo, se utiliza una segmentación basada en objeto [13]. La mano del usuario se presupone siempre como el objeto en primer plano. Un sistema de iluminación infrarroja compuesto por 16 diodos proporciona el suficiente contraste para segmentar. En entornos operacionales, con condiciones de iluminación y de fondo no controladas, la robustez de la segmentación se convierte en una tarea crucial.

Se probaron diferentes técnicas de segmentación. Las técnicas más comunes son los métodos basados en detección de piel [14]. La detección de piel no es lo suficientemente robusta en entornos operacionales. Condiciones de iluminación variables, y fondos complejos con superficies y objetos con colores semejantes a la piel causan problemas graves. En un principio, se utilizó un sistema de iluminación basado en una lámpara de 60W emitiendo en la gama de luz visible. Se capturaron 10 imágenes de 20 usuarios diferentes. Después de un estudio de las manos capturadas, se observó que el rendimiento del sistema decaía en las siguientes situaciones:

- cuando incidía de forma directa luz en la lente de la cámara.
- cuando se trabajaba con fondo altamente no uniforme.
- cuando existía más de un objeto en primer plano.
- cuando existía un objeto reflectivo en el fondo.

La disminución del rendimiento del sistema se debía fundamentalmente a problemas en la segmentación, siendo imposible extraer la mano en ciertas condiciones. El rendimiento de este tipo de sistemas depende en gran medida de la etapa de segmentación.

Con una iluminación correcta, el problema de la segmentación puede ser solucionado. En la figura 1.a podemos ver un ejemplo de una imagen capturada en un entorno no controlado. La segmentación del fondo es una tarea complicada con esta clase de entornos. En La figura 1.b podemos ver los problemas que aparecen al utilizar métodos de segmentación basados en la detección de piel. Para solucionar los problemas asociados a la segmentación, se propone el uso de iluminación en la banda de infrarrojos. El sistema de iluminación con luz visible fue substituido por un sistema de iluminación infrarroja. Se extrajo el filtro de infrarrojos de la webcam (filtro habitual en todas las webcams comerciales) y se sustituyó por un filtro de visible. En los siguientes párrafos se darán más detalles de esta adaptación. Para obtener una adquisición adecuada de la imagen, se deben tener en cuenta algunas opciones de adquisición de la webcam: el brillo a su valor mínimo con un valor máximo de contraste, ganancia baja y exposición mínima para conseguir un gran contraste del objeto en primer plano. Podemos ver un ejemplo en la figura 1.c La figura 1.d muestra como disminuyen las interferencias de objetos similares cercanos.

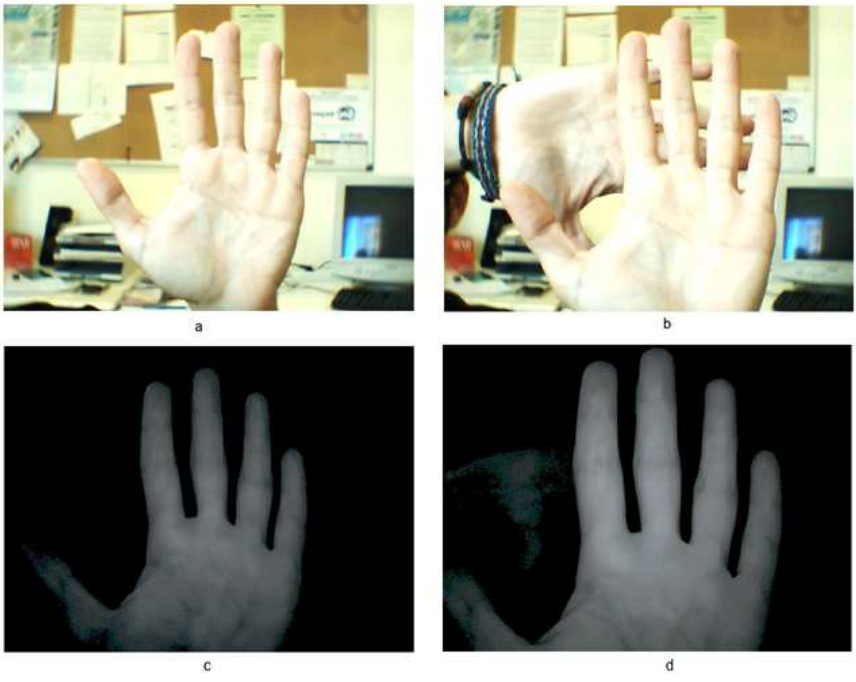


Figura 1. Segmentacion con iluminación infrarroja

El sistema de iluminación está formado por diodos GaAs (CQY 99) con una emisión media en 850 nanómetros y un ancho de banda espectral medio de 40 nanómetros. Los diodos fueron colocados en forma invertida de U con la webcam situada en medio, figura 2.

El número de diodos fue reducido gradualmente y la corriente fue aumentada hasta encontrar la mejor relación entre consumo, número de diodos y la iluminación correcta de la mano. El número final de diodos es 16 y la corriente es de 30 mA por cada par de diodos. El sistema se alimenta por usb, con 5V y 500mA de capacidad.

La webcam fue modificada para adaptarla a las emisiones infrarrojas: el filtro infrarrojo fue extraido y se agregaron dos filtros en cascada. Los filtros utilizados son los Kodak No 87 FS4-518 y No 87c FS4-519, sin transmitancia entre los 400-700 nanómetros.

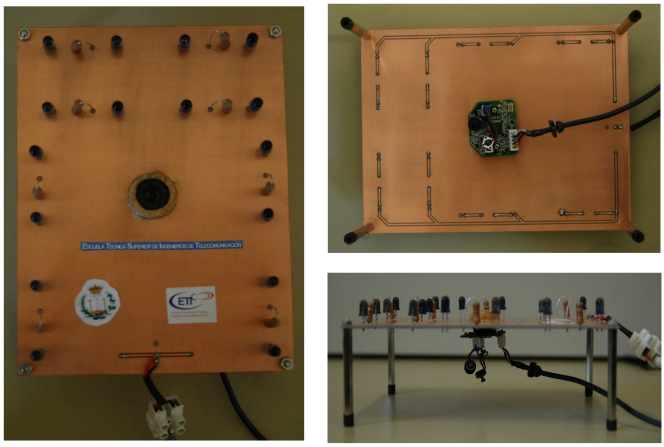


Figura 2. Capturas del sistema

El usuario coloca la palma de la mano libremente en el espacio 3D delante de la cámara. No se utiliza superficie de apoyo. Se utiliza una plantilla mostrada por pantalla para guiar al usuario y reducir las distorsiones de proyección asociadas a la ausencia de superficie de contacto, ver figura 3.

Una vez se obtiene la imagen infrarroja, la segmentación es simple. Se utiliza un filtrado paso bajo para resaltar la mano frente al fondo. Se utiliza una ventana de Hamming de dos dimensiones. Las frecuencias del corte son $w_1 = \pi$ para el filtro paso todo y $w_2 = 0{,}5$ para el filtro paso bajo. Este filtrado agudiza el contraste entre la mano y el fondo.

La imagen filtrada se normaliza en amplitud y es binarizada por el método Otsu [15]. Este método escoge el umbral minimizando la variación intraclases de los pixeles blancos y negros.

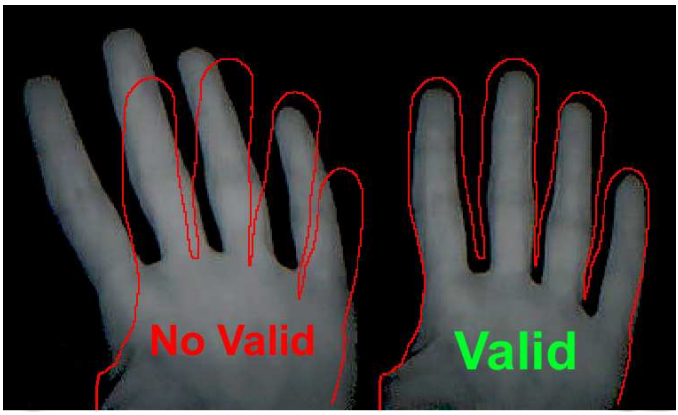


Figura 3. Plantilla para el correcto posicionamiento

4. Parametrización

El contorno de la mano se obtiene de la imagen en blanco y negro. Para localizar las puntas y los valles entre los dedos se convierten las coordenadas cartesianos del contorno de la mano a coordenadas polares (radio y ángulo) que consideran como origen de las coordenadas el centro de la base de la mano. Los valores máximos del radio establecen las puntas de los dedos y los mínimos indican los valles entre los dedos. Posteriormente se refinan los puntos aplicando un estudio del gradiante para cada dedo [16].

Localizados los puntos a partir de los cuales caracterizar los dedos, se pueden obtener las características geométricas de cada uno de ellos. Se utilizan solamente las características geométricas de los dedos: índice, corazón y anular, véase figura 4. La razón de desechar los dedos pulgar y meñique se debe a problemas de iluminación de la región para el caso del meñique y la alta variabilidad de las medidas del dedo pulgar.

Obtener el vector de características geométricas es inmediato una vez se conocen las puntas y los valles de la mano. Cada dedo se caracteriza como un triángulo. La punta del dedo se considera la cima del triángulo y los valles forman la base. Se obtienen alrededor de 25 medidas para cada uno de los dedos. Se desecha el primer 20 % del dedo para evitar problemas asociados a la presencia de anillos. Por lo tanto, cada usuario es caracterizado por 75 medidas.

5. Verificación

La base datos a partir de la cual se ha realizado la verificación del sistema consta de 927 accesos obtenidos durante 3 meses. Los accesos se dividen en tres clases: 564 accesos de usuarios originales; 174 accesos de intrusos; 189 accesos de usuarios originales que introducen mal su número de identificación.

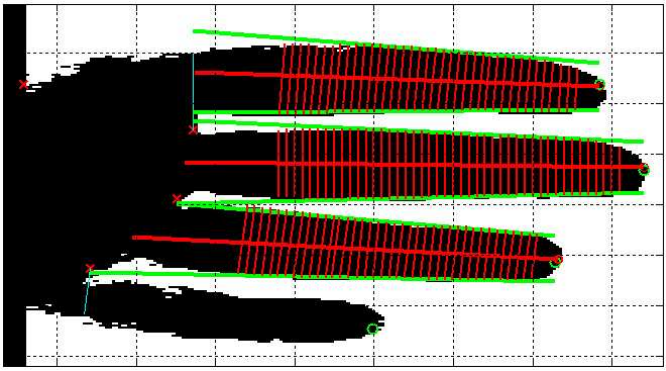


Figura 4. Parametrización de los dedos índice, corazón y anular

La base de datos consta de más de 4000 imágenes. Tomando un máximo de 8 imágenes por acceso para resolver la autenticación. Se solicitan dos imágenes identificadas positivamente para considerar al usuario autenticado.

Se ha utilizado tecnología de Máquina de Vector Soporte (SVM) para la clasificación. El software utilizado para entrenar y para clasificar es SVMlight [17]. Para autenticar que una mano de entrada pertenece a la identidad demandada, se calcula la distancia de la mano respecto al hiperplano separador del modelo SVM generado para esa identidad. Si la distancia es mayor que un umbral en al menos dos imágenes, se acepta la identidad. Si tras 8 imágenes capturadas no se consiguen estas dos autenticaciones positivas, el usuario se considera impostor.

6. Experimentos

Se utilizaron cuatro imágenes de la mano derecha de cada usuario para el entrenamiento. Se situó el sistema en un laboratorio de acceso restringido y se realizaron accesos durante 3 meses. Se capturaron más de 4000 imágenes de 57 usuarios diferentes.

Se diferencian tres tipos de accesos:

- usuarios originales que acceden correctamente al sistema: son usuarios dados de alta en el sistema que acceden de forma normal con su número de identificación.
- usuarios intrusos: son usuarios no dados de alta en el sistema que intentan suplantar la identidad de usuarios dados de alta.
- usuarios originales que se equivocan de identificador: son usuarios dados de alta en el sistema que se intentan hacer pasar por otros usuarios dados de alta en el sistema.

El rendimiento del sistema se medirá partir de la Tasa de Falsa Aceptación (FAR), Tasa de Falso Rechazo (FRR), Tasa de Error Común (EER). En la figura 5 se pueden observar los resultados obtenidos.

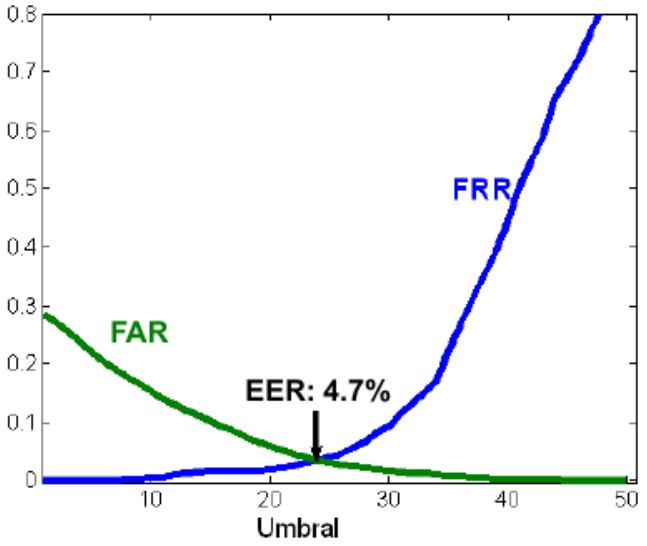


Figura 5. Resultados obtenidos

Además de las tasas mostradas, se debe destacar la de Fail To Enrollment, que alcanza un porcentaje del 9.4 %. Esta tasa se debe fundamentalmente a la dificultad de los usuarios en un principio para hacer coincidir la mano con la plantilla mostrada en pantalla. Este porcentaje se reduce a medida que el usuario adquiere experiencia en el uso del sistema.

7. Conclusiones

En este artículo se ha presentado un prototipo de sistema de autenticación sin contacto basado en la geometría de la mano. Una base de datos de más de 4000 imágenes, capturadas durante tres meses se ha utilizado para validar el sistema. Los resultados arrojan un EER del 4.7 %, lo que anima a seguir con esta línea. Con el problema de la segmentación prácticamente solucionado, se debe mejorar los algoritmos de parametrización. Se busca reducir al mínimo los efectos de la distorsión debida a la proyección sin dificultar la accesibilidad al sistema.

Agradecimientos

Este trabajo ha sido realizado gracias a la ayudas del Gobierno español en el proyecto TEC2006-13141-C03-01/TCM y TSI-020100-2008-279.

Referencias

[1] A.K. Jain, R. Bolle, and S. Pankanti: Biometrics: Personal Identification in Networked Society. Kluwer Academic Publishers. (2001).

[2] Tantachun, S.; Pintavirooj, C.; Lertprasart, P.; Bunluechokchai, S: Biometrics with Eigen-Hand. Electronics and Applications 2006 1ST IEEE Conference on 24-26 May 2006 Page(s):1 - 4.

[3] Zheng, G.; Wang, C.-J.; Boult, T. E.: Application of Projective Invariants in Hand Geometry Biometrics, Information Forensics and Security, IEEE Transactions on, Volume 2, Issue 4, Dec. 2007 Page(s):758 - 768.

[4] A. Jain, A. Ross, and S. Pankanti: A prototype hand geometry-based verification system, in Proc. 2nd Int. Conf. Audio- and Video-Based Biometric Person Authentication, Mar. 1999, pp. 166-171.

[5] R. Sanchez-Reillo, C. Sanchez-Avila, and A. Gonzalez- Marcos: Biometric identification through hand geometry measurements, IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 10, pp. 1168-1171, Oct. 2000.

[6] R. Sanchez-Reillo: Hand geometry pattern recognition through Gaussian mixture modelling, in Proc. 15th Int. Conf. Pattern Recognition, 2000, vol. II, pp. 941-944.

[7] A.Wong and P. Shi: Peg-free hand geometry recognition using hierarchical geometry and shape matching, in Proc. IAPRWorkshop on Machine Vision Applications, Nara, Japan, Dec. 2002, pp. 281-284.

[8] Y. Bulatov, S. Jambawalikar, P. Kumar, and S. Sethia: Hand recognition using geometric classifiers, in Proc. 1st Int. Conf. Biometric Authentication, Hong Kong, China, Jul. 2004, pp. 753-759.

[9] G. Boreki and A. Zimmer: Hand recognition using geometric classifiers, in Proc. 1st Int. Conf. Biometric Authentication, Hong Kong, China, Jul. 2004, pp. 753-759.

[10] J. Hashemi and E. Fatemizadeh: Biometric identification through hand geometry, in Proc. Int. Conf. Comput. Tool, 2005, vol. 2, pp. 1011-1014.

[11] S. Haeger: South Florida. Tampa, FL, Dec. 2003.

[12] K. Garrison, A. Sorin, Z. Liu, and S. Sarkar: Hand Biometrics From Image at a Distance Dept. Compu. Sci. Eng., Univ. South Florida, Tampa, Tech. Rep.: USF-Nov-2001-Palm, 2001.

[13] Lijie Liu; Guoliang Fan: Combined key-frame extraction and object-based video segmentation, Circuits and Systems for Video Technology, IEEE Transactions on, Volume 15, Issue 7, July 2005 Page(s):869 - 884.

[14] Ruiz-del-Solar, J.; Verschae, R.: Skin detection using neighbourhood information, Automatic Face and Gesture Recognition 2004 Proceedings. Sixth IEEE International Conference on 17-19 May 2004 Page(s):463 - 468.

[15] Image Processing Toolbox^{TM} 6.0 for MATLAB The Language of Technical Computing,The MathWok Inc., Natick, MA., 1997.

[16] Ferrer, M.A.; Morales, A.; Travieso, C.M.; Alonso, J.B.: Low Cost Multimodal Biometric identification System Based on Hand Geometry, Palm and Finger Print Texture", Security Technology, 2007 41st Annual IEEE International Carnahan Conference on 8-11 Oct. 2007 Page(s):52 - 58.

[17] T. Joachims: Making large-Scale SVM Leaming Practical. Advances in Kemel Methods Support Vector Leaming, B. Scholkopf and C. Burges and A. Smola (ed.), MITPress, 1999.

Desarrollo de un Sistema que Integra Componentes Biométricos Acoplado a un Esquema Transaccional Bancario Aplicando Reconocimiento de Huella Digital y Captura de Rostro

Juan Francisco Fuentes Tamayo
Biometrika S.A. Tecnología Innovadora
Av Shyris n32-14 y Diego de Almagro, Edificio Torrenova 5to Piso Quito-Ecuador

{ ffuentes@biometrika.ec }

http://www.biometrika.ec

Resumen: el presente documento contiene una solución práctica de un aplicativo el cual se encuentra actualmente en funcionamiento y permite a las instituciones bancarias reducir los fraudes utilizando el reconocimiento de huella dactilar y la captura del rostro del cliente.

1. Caso de estudio.-

Institución bancaria que posee 95 oficinas y más de 450 cajeros automáticos.

2. Objetivos.-

2.1 Toma fotografías de aquellas personas que utilizan sus servicios tanto en ventanilla como en cajeros automáticos, con la finalidad, en caso de algún reclamo o robo, poder identificar a la persona que realizo mencionada transacción.

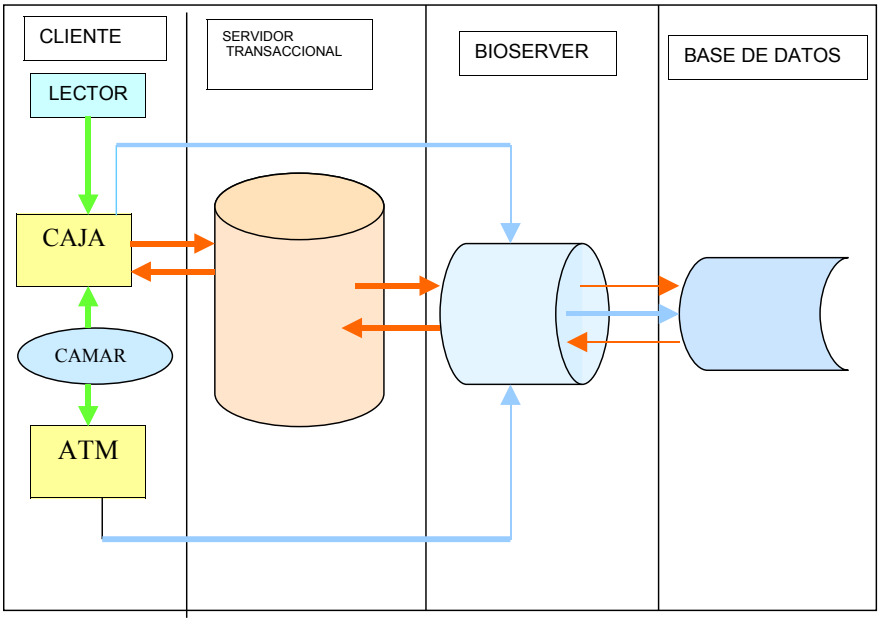
2.2 La identificación de los clientes que realizan transacciones de retiro de dinero y la verificación de la identidad para el bloqueo de transacciones para aquellas personas que han cometido un delito.

3. Desarrollo.-

Mencionados los puntos principales del problema, es necesario recalcar, que la solución debe tener un impacto mínimo en cuanto se refiere a la infraestructura informática del banco.

Es prudente que la mayoría de la solución sea centralizada. Además debe integrase al servidor de la institución bancaria sin grandes impactos tecnológicos, de tal manera, que las transacciones en especial de ventanilla no se vean afectadas en los tiempos de respuestas.

La solución que se encuentra instalada para resolver los puntos referentes a biometría informática y fotografía, es la siguiente



Como se puede apreciar en el diagrama, existen cuatro actores en esta solución:

3.1 Cliente

Esta capa tiene que ver con todo lo que es cajas, atm, cámaras y lectores de huella digital.

Tiene por objetivo captar y despachar transacciones de atención al público.
Las flechas de color verde indican la conexión física entre las cámaras de videos y lectores a las cajas de ATM.

En este nivel se encuentran las librerías de integración biométrica, encargadas de la lectura de huella digital y el programa para toma de fotografía y reconocimiento de rostro.

3.2 Servidor transaccional

Aquí se encuentra todo el sistema central del banco.
Las flechas de color rojo indican, todas las transacciones que el servidor transaccional, recibe para procesarlas.

3.3 Bioserver

Esta tercera capa es el monitor biométrico y el servidor de rostros.

Aquí se encuentran las reglas de negocio, que reconocen la huella digital y el rostro captado en oficinas y ventanillas.

Las flechas azules, indican la integración con el servidor transaccional y la comunicación directa con la capa del cliente, así como también la integración con la base de datos.

3.4 Base de Datos

Está es la última capa que se encuentra todas las transacciones efectuadas por la caja y ATM. Las transacciones contienen datos relevantes encriptados.

Como se puede apreciar, la solución consta de varias capas de las cuales 3 son centralizadas, facilitando el mantenimiento y la ejecución de un plan de contingencia en caso de cualquier falla.

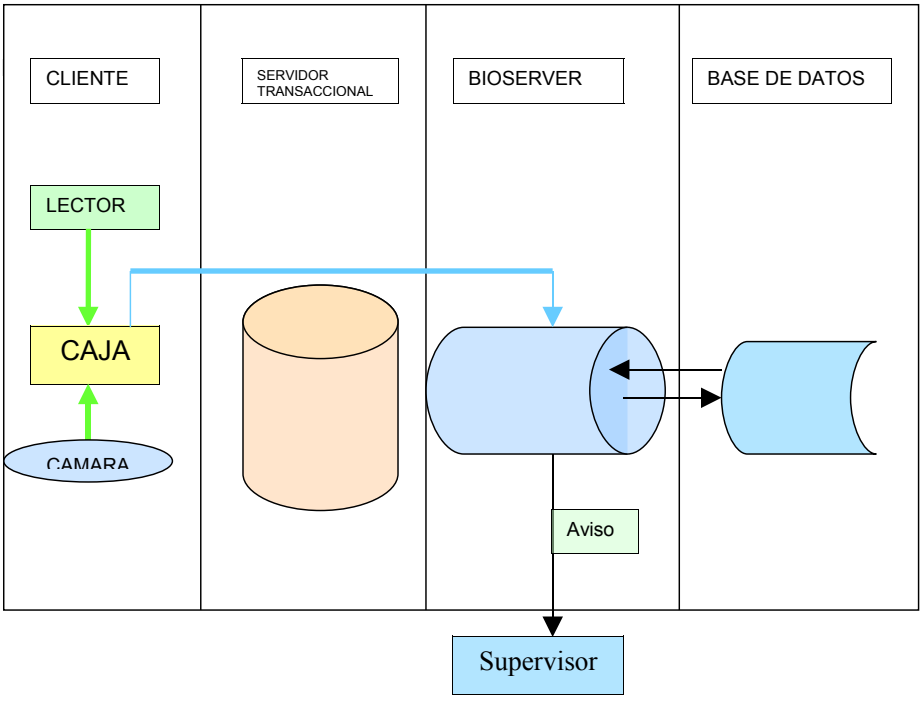
---

* Nótese que Bioserver es marca registrada exclusiva de Biométrika S.A.

3.5 Alarmas

Se encuentra implementado un sistema de alarmas, con el propósito de prevenir el mal funcionamiento de dispositivos con la cámara de video o cualquier anomalía a nivel de cliente, servidor y base de datos.

Este sistema funciona de la siguiente manera:



3.6 Análisis de la capa del cliente

El cliente por ser una capa no centralizada, puede tener varias complicaciones. Examinemos que contiene el cliente.

a) Conexión del lector Biométrico
b) Conexión de la cámara de video
c) Librerías de integración de lectura y manejo del dispositivo biométrico
d) Aplicación de reconocimiento de rostros

3.6.1 Conexión del lector biométrico-

En este punto pueden presentarse algunos casos:

1) Fallas en el lector.- la caja tiene la opción de deshabitar la utilización de los lectores.

2) Fallas en el puerto USB.- si existen fallas en el puerto USB, el software emite una alerta al cajero indicando que no se encuentra el lector conectado, y no se puede continuar la transacción.

3.6.2 Conexión de la cámara de video

Se pueden presentar los siguientes casos:

1) **Fallas en la cámara.-** si la cámara se daña, el proveedor de la cámara debe cambiar por otra, y garantizar el correcto funcionamiento de la misma

2) **Fallas en el puerto USB.-** si no se detecta la cámara, aparece un mensaje en caja el mismo que denuncia que existe problema con el reconocimiento de la cámara

Los problemas de huellas digitales que se han presentado en el universo de clientes que se acercan a realizar una transacción, se los clasificó en tres categorías:

Personas que trabajan en el sector urbano, es decir aquellas personas que realizan trabajos los cuales no inciden en la deformación de las huellas dactilares.

Personas que trabajan en el sector rural, es decir aquellas personas que realizan trabajos los cuales tienden con el paso del tiempo a deformaciones de las huellas que posteriormente no pueden hacer un match con la huella almacenada en la base de datos (ver figura 1).

figura 1

Personas discapacitadas, las cuales no tienen extremidades superiores.

Cada persona ubicada en estos grupos, se tuvieron que someter a una capacitación diferente, ya que el uso de un sistema biométrico no solo es tecnológico, sino cultural y de actitud.

Cabe señalar que se obtuvo en un inicio solamente con el uso de herramientas tecnológicas sin ninguna metodología, un 40% de desaciertos, el cual fue reducido a un 2% con la utilización de procesos metodológicos a nivel del sistema y a nivel de capacitación dentro del grupo de personas que trabajan en el sector rural.