UNIVERSIDAD AUTÓNOMA DE MADRID

ESCUELA POLITÉCNICA SUPERIOR

PROYECTO FIN DE CARRERA

RECONOCIMIENTO DE ESCRITURA OFF-LINE A PARTIR DE CARACTERÍSTICAS DE EMISIÓN ALOGRÁFICA

Ingeniería de Telecomunicación

Rubén Fernández de Sevilla García

Abril 2012

RECONOCIMIENTO DE ESCRITURA OFF-LINE A PARTIR DE CARACTERÍSTICAS DE EMISIÓN ALOGRÁFICA

AUTOR: Rubén Fernández de Sevilla García
TUTOR: Fernando Alonso Fernández

ATVS Grupo de Reconocimiento Biométrico
(http://atvs.ii.uam.es)
Dpto. de Ingeniería Informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid

Resumen

En este proyecto se desarrolla, implementa y evalúa un sistema de identificación de escritor basado en características alográficas. Este sistema opera sobre caracteres aislados, considerando que cada persona utiliza un determinado número de formas para cada uno de ellos. La base de datos utilizada para evaluar el sistema está formada por muestras reales de caracteres de 30 escritores. Dichas muestras fueron tomadas en un entorno forense real, y segmentadas y etiquetadas manualmente por un experto forense, por lo que se dispone de la información de clase de carácter de cada una de ellas. En total, consideramos 62 clases alfanuméricas (10 números y 52 letras, incluyendo minúsculas y mayúsculas). El sistema realiza identificación *off-line*, es decir, sobre muestras de escritura ya escritas y escaneadas, y es independiente del contenido del texto.

En el sistema de identificación de escritor implementado la identidad de cada escritor se modela a partir de una función de densidad de probabilidad. Para obtener dicha función es necesario contar con un catálogo de alógrafos que recoja la variedad de formas de caracteres que podemos encontrar en una muestra de escritura. La generación del catálogo de alógrafos se ha realizado utilizando técnicas de agrupamiento (*clustering*). En concreto, el algoritmo utilizado ha sido el conocido como *k-means*, mediante el cual se realiza una cuantificación vectorial de los datos deseados, obteniendo una serie de centroides, que conforman nuestro catálogo (*codebook*) de alógrafos. Para obtener los catálogos se ha utilizado la base de datos de caracteres CEDAR.

En el proyecto se estudia la influencia que el catálogo de alógrafos tiene sobre el rendimiento total del sistema. Para poder llevar a cabo dicho análisis se han determinado dos escenarios de evaluación distintos. En el primero de ellos se ha generado un único catálogo global, que no hace uso de la información de clase dada por el etiquetado previo de cada carácter. Por otra parte, en el segundo de los escenarios sí se hace uso de la información de clase de cada carácter, generándose un subcatálogo por cada una de las 62 clases alfanuméricas contempladas, siendo en este caso, por tanto, 62 funciones de densidad de probabilidad las que caracterizan a cada escritor. Al comparar el rendimiento del sistema entre ambos escenarios se puede determinar si el etiquetado manual de caracteres es beneficioso o no.

En los dos escenarios evaluados también se ha analizado la influencia que posee el tamaño del catálogo de alógrafos, es decir, el número de centroides del mismo, en el rendimiento del sistema. Y en el caso del escenario basado en subcatálogos locales se han realizado pruebas utilizando subconjuntos de los 62 caracteres totales, para poder analizar la posibilidad de utilizar un reducido número de caracteres, reduciendo así el tiempo del segmentado y etiquetado manuales.

Por último, se ha evaluado cómo afecta el tamaño de la lista objetivo (Top N), así como el rendimiento de la fusión del sistema de alógrafos desarrollado en este proyecto con un sistema basado en características de gradiente disponible en el grupo ATVS.

Palabras clave

Biometría, alógrafos, función de densidad de probabilidad, identificación de escritor, identificación forense, k-means, sistemas biométricos.

Abstract

In the present project, a writer identification system based on allographic features has been studied, implemented and evaluated. The system operates on isolated characters, considering that each writer uses a reduced number of shapes for each one of them. The database used to evaluate the system contains character samples of real forensic cases from 30 different writers. These samples were acquired from real forensic cases, and were segmented and labeled by a forensic expert, therefore information about the character class of every character is provided. The total number of alphanumeric classes considered is 62 (10 numbers and 52 letters, including lowercase and uppercase letters). The system is text-independent and works on offline mode, i.e. scanned handwriting images are used for identification.

In the writer identification system implemented, the identity of each writer is modeled by a probability density function. To obtain that function, it is required to have a common allographs codebook which provides a common shape space and captures the individual shape usage preference of the writer. The codebook has been generated by means of clustering techniques. In particular, the clustering technique used has been k-means, a vector quantization method applied to obtain a certain number of clusters, which compose the common allographs codebook. For the codebook generation, the CEDAR database has been used.

In this project we have analyzed the influence of the codebook in the system performance. In order to carry out that analysis, two different scenarios have been proposed. In the first one of them a global codebook that does not use the character class information, given by the previous manual labeling, has been generated. On the other hand, in the second scenario a local character-based codebook is generated. This local codebook is composed of 62 "sub-codebooks", one per each character. In this case, we exploit the class information given by the character segmentation and labeling carried out by the forensic expert. Each writer is here modeled by 62 probability density functions. A comparison between the system performance in both scenarios has been made in order to determine if the manual character labeling is worthy.

In both scenarios we have also studied the influence of the codebook size (i.e. the number of clusters in the codebook) in the system performance. Besides, in the local character-based scenario experiments have been done to evaluate the combination of the 62 alphanumeric characters, examining therefore the possibility of using a reduced subset of the 62 characters and reducing the time spent on manual segmentation and labeling.

Finally, we have evaluated the performance of the system depending on the hit list size (Top N), as well as the fusion of the allograph system developed in the present project and a gradient-based system available at the ATVS biometric recognition group.

Key words

Biometrics, allographs, probability density function, writer identification, forensic identification, k-means, biometric systems.

Agradecimientos

A mi ponente Javier Ortega, por confiar en mí al darme la oportunidad de formar parte del ATVS.

A mi tutor Fernando Alonso, por su inestimable ayuda a la hora de realizar este Proyecto.

A todos los miembros del ATVS, que de una forma u otra siempre estuvieron dispuestos a ayudarme con una sonrisa en la cara.

A mis compañeros de la EPS, y en especial a Chema, Javier, Pablo, Roberto, Rubén y Santi, por hacer tan llevaderos estos años de universidad.

A mi familia, mis padres y mi hermano Sergio, por apoyarme en todo momento y servirme siempre como referentes y guías.

A Rebeca, porque sin ella nada sería lo mismo.

Rubén Fernández de Sevilla García
Abril 2012

Índice General

Índice de Figuras

Índice de Tablas

Glosario de acrónimos

- ADN: Ácido Desoxirribonucleico
- CEDAR: Center of Excellence for Document Analysis and Recognition
- DET: Detection Error Trade-Off
- EER: Equal Error Rate
- FAR: False Acceptance Rate
- FDP: Función de Densidad de Probabilidad
- FRR: False Rejection Rate
- ICFHR: International Conference on Frontiers in Handwriting Recognition
- JRBP: Jornadas de Reconocimiento Biométrico de Personas
- LLR: Log-Likelihood Ratio

1

Introducción

1.1 Motivación

En los últimos años la proliferación del uso de las tecnologías de la información, unido a una creciente preocupación en aspectos de seguridad, ha permitido que se desarrollen nuevas técnicas de identificación automática de personas en base a sus rasgos biométricos.

Estos rasgos biométricos pueden clasificarse en rasgos biométricos fisiológicos (basados en medidas de características físicas) y rasgos biométricos de comportamiento [1]. Entre los primeros, se encuentran el iris, el ADN, la huella dactilar, etc. Estos rasgos se caracterizan por una reducida variabilidad a lo largo del tiempo, aunque por el contrario su adquisición es a menudo más invasiva. Por otra parte, los rasgos biométricos de comportamiento, como la voz, la firma o la escritura, suelen ofrecer tasas de identificación menores, pero son menos invasivos y su aceptación en la sociedad es mayor.

En este proyecto el rasgo biométrico utilizado es la escritura, que es considerada algo individual, como muestra el alto grado de aceptación social y legal de las firmas como un medio de validación de la identidad, algo también apoyado por estudios experimentales [2]. Los sistemas de autentificación de personas basados en imágenes escaneadas de escritura (reconocimiento de escritor *off-line*) tienen su base en considerar que la variación del estilo de escritura entre diferentes escritores es mayor que la variación intrínseca de cada escritor de manera aislada, como podemos ver en la Figura 1.1.

Estos sistemas han suscitado en los últimos años un gran interés por su aplicación en el ámbito forense o en el análisis de documentos históricos, con el objetivo de determinar la identidad del escritor [3]. Esto último también es una importante tarea de aplicación en el ámbito forense, existiendo numerosos casos en juicios a lo largo de los años en los que ha sido utilizada la evidencia provista por el análisis de estos documentos, con el objetivo de determinar la autoría de los mismos. [4].

Figura 1.1. Comparación entre caracteres escritos por tres escritores distintos

El reconocimiento de escritor tiene como objetivo determinar si dos documentos escritos, referidos comúnmente en el ámbito forense como documento dubitado (cuya autoría se desconoce) y documento indubitado (del cual se conoce su autor), han sido escritos por la misma persona o no. Técnicas basadas en la visión artificial y el reconocimiento de patrones han sido a menudo aplicadas a este problema para dar soporte a los expertos forenses [5, 6].

El escenario forense presenta algunas dificultades debido a sus particulares características de [7]: reducido número de muestras escritas, variabilidad del estilo de escritura, lápiz o tipo de papel, presencia de patrones de ruido, etc. o no disponibilidad de información *on-line* (dinámica). Como consecuencia de ello, este dominio de aplicación aún se basa fuertemente en la interacción del experto humano. El uso de sistemas de reconocimiento semi-automáticos es muy útil para, dada una muestra de escritura dubitada, obtener una lista reducida de posibles candidatos que se encuentran en una base de datos de identidades conocidas, haciendo más fácil el posterior cotejo del experto forense [6, 7].

1.2 Objetivos y enfoque

En los últimos años se han definido diversos algoritmos de identificación de escritor que se basan en diferentes grupos de características [8]. En este proyecto se presenta un sistema que hace uso de características del nivel alográfico, que se basa en discriminar escritores mediante la codificación de sus alógrafos más utilizados, en base a su probabilidad de ocurrencia en una muestra escrita. Algunos trabajos previos que se enmarcan también en el nivel alográfico hacen uso de imágenes de componentes conectadas [9] o de contornos [10, 11] usando segmentación automática.

La segmentación automática perfecta de caracteres individuales aún es un problema sin resolver [7], mientras que los componentes conectados compuestos por varios caracteres o varias sílabas sí pueden segmentarse de forma automática fácilmente, y también permiten capturar detalles de la forma de los alógrafos utilizados por el escritor [12]. El sistema propuesto en este proyecto, sin embargo, utiliza caracteres individuales segmentados de forma manual por un experto forense, que a la vez asigna cada carácter a una de las 62 clases alfanuméricas: dígitos ("0"-"9"), letras minúsculas ("a"-"z") y mayúsculas ("A"-"Z"). El hecho de utilizar esta configuración se deriva de que es la configuración usada por el grupo forense participante en este trabajo. Para cada individuo, se escanea el documento autentificado y después se aplica una herramienta de software para realizar la segmentación de caracteres. Esta segmentación se hace manualmente por el experto forense, que realiza la selección del carácter mediante el ratón del ordenador y etiqueta la muestra correspondiente de acuerdo a las 62 clases mencionadas.

En este trabajo se ha adaptado el algoritmo de reconocimiento basado en características alográficas de [12] para adaptarlo a esta configuración. Adicionalmente, el sistema se evalúa utilizando una base de datos creada a partir de documentos forenses reales (confiscados a criminales reales o autentificados en presencia de un agente de la policía), lo que es una diferencia importante en comparación con los experimentos de otros trabajos, en los que las muestras de escritura son obtenidas con la colaboración de voluntarios y bajo condiciones controladas [13].

El sistema es evaluado en modo identificación, donde cada individuo se identifica por una búsqueda entre todos los integrantes de la base de datos (búsqueda uno a muchos). Como resultado, se devuelve una clasificación ordenada de candidatos. Idealmente, la primera posición (Top 1) debería corresponder con la identidad correcta del individuo, pero se puede considerar un tamaño de lista más grande (p.ej. Top 10) para incrementar las posibilidades de encontrar la identidad correcta. La identificación es un componente crítico en aplicaciones forenses y criminales, donde el objetivo es comprobar si la persona es quien él/ella (implícita o explícitamente) niega ser [14].

Otro de los objetivos que se persiguen con este proyecto es analizar la influencia que el catálogo de alógrafos tiene sobre la tasa de acierto del sistema de identificación. En concreto, analizaremos el impacto que tiene el tamaño de dicho catálogo, así como las diferencias entre usar un único catálogo global que no tenga en cuenta los diversos caracteres o, por el contrario, un catálogo propio para cada uno de los caracteres. Del mismo modo, se analizará cómo influye el número de caracteres a utilizar, y si es viable utilizar un subconjunto de los mismos sin que se vea mermada la capacidad de identificación del sistema.

Como sistema de referencia respecto al sistema implementado en este proyecto, realizaremos también experimentos sobre un sistema de identificación de escritor basado en características de gradiente, desarrollado por el grupo ATVS.

1.3 Contribuciones del proyecto

Las contribuciones de este Proyecto Fin de Carrera al grupo de reconocimiento biométrico ATVS y a la comunidad científica pueden resumirse en los siguientes aspectos:

- Estudio del estado del arte en biometría, sistemas biométricos y medidas de rendimiento, prestando especial detalle a los sistemas de reconocimiento de escritor.

- Implementación y desarrollo de un sistema de reconocimiento de escritor basado en características alográficas.

- Evaluación del sistema de reconocimiento de escritor desarrollado, y de otro sistema de reconocimiento de escritor disponible en el grupo utilizando una base de datos con muestras forenses reales, a diferencia de la mayoría de trabajos existentes, donde las bases de datos utilizadas para evaluar los sistemas están compuestas por muestras tomadas en condiciones controladas y con escritores colaborativos.

- Análisis del impacto del tamaño de los catálogos de alógrafos utilizados en este tipo de sistemas en el rendimiento del mismo.

- Comparativa del rendimiento de los sistemas de escritor basados en características alográficas utilizando un catálogo global o utilizando un catálogo para cada uno de los caracteres.

- Adecuación del sistema desarrollado al reconocimiento de firma, participando con dicho sistema, junto con otros sistemas del grupo ATVS, en la competición sobre verificación de firma offline "*4NSigComp2010 – Forensic Signature Verification Competition*" [15], que se organizó en el marco de la *International Conference in Frontiers of Handwriting Recognition – ICFHR2010*. (Ver Anexo C)

El sistema de reconocimiento de escritor desarrollado en este proyecto ha dado también lugar a dos artículos científicos aceptados como artículos de congreso con revisión científica, tanto de carácter internacional como nacional. En concreto, el artículo titulado "*Forensic Writer Identificacion Using Allographic Features*" [16] fue aceptado en el congreso internacional ICFHR 2010, y el artículo con título "*Identificación Forense de Escritor Usando Características de Emisión Alográfica*" [17] fue aceptado en las V Jornadas de Reconocimiento Biométrico de Personas (JRBP 10). Ambos artículos se incluyen al final de la memoria, en el Anexo D.

1.4 Organización de la memoria

La memoria de este proyecto está compuesta por los siguientes capítulos:

1. Introducción

Este capítulo describe la motivación para la realización de este proyecto, así como los objetivos a alcanzar y el enfoque utilizado para ello. También se realiza un pequeño resumen acerca de las contribuciones de este Proyecto Fin de Carrera

2. Introducción a la biometría

En este capítulo se realiza una introducción a la biometría, en la que se indican características generales de los rasgos biométricos, así como un breve apunte sobre los más utilizados, y sobre el uso de la biometría en ámbitos forenses. En la última parte del capítulo, se habla de los sistemas biométricos: arquitectura, modos de operación, medidas de rendimiento, y multimodalidad biométrica.

3. Reconocimiento biométrico de escritor

Tras la introducción genérica a la biometría del capítulo anterior, en este capítulo se analizan aspectos relacionados con el reconocimiento de escritor, junto con un resumen de trabajos previos en esta categoría

4. Sistemas de reconocimiento de escritor

En el capítulo 4 se describe por una parte el sistema basado en características de gradiente, disponible en el grupo ATVS, y sobre el que se han realizado algunos cambios con el objetivo de mejorar su rendimiento. Por otra parte, en este mismo capítulo describimos el nuevo sistema desarrollado, basado en características de emisión alográfica.

5. Experimentos en reconocimiento de escritor

Los resultados de los experimentos realizados en el proyecto se muestran en este cuarto capítulo. En la primera parte del mismo, se describen las dos bases de datos utilizadas (base de datos del grupo forense con el que hemos colaborado, y base de datos CEDAR). A continuación, se expone el protocolo experimental utilizado en los experimentos para los dos escenarios planteados. Por último, se muestran los resultados, tanto del sistema disponible evaluado como del sistema desarrollado, así como los obtenidos con la fusión de ambos sistemas.

6. Conclusiones y trabajo futuro

En el último capítulo se presentan las conclusiones del proyecto, así como el trabajo futuro que podría derivar del mismo.

Al final de la memoria se presenta la bibliografía utilizada, junto a una serie de anexos que incluyen las publicaciones científicas derivadas del presente proyecto y la competición de verificación de firma en la que se participó, así como el presupuesto y el pliego de condiciones.

2

Introducción a la biometría

2.1 Biometría

Actualmente el uso de técnicas de reconocimiento y autenticación biométrica está cobrando gran relevancia, ya que suponen una forma sencilla y segura de identificación de personas. A diferencia de otros métodos, como las claves personales o las tarjetas magnéticas, la principal ventaja de los rasgos biométricos es que éstos no pueden ser olvidados o fácilmente copiados, al estar basados en lo que la persona es y no en lo que la persona posee o conoce.

Podemos definir la biometría como la ciencia utilizada para el reconocimiento de individuos en base a características físicas o de comportamiento. Estas características son llamadas rasgos biométricos.

2.1.1 Características de los rasgos biométricos

Para que una característica del ser humano pueda ser considerada como rasgo biométrico debe reunir las condiciones siguientes [18]:

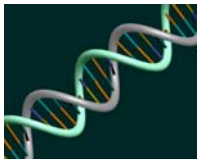
- **Universalidad:** cualquier persona debe poseer dicha característica.
- **Distintividad:** personas distintas deben poseer rasgos distintos.
- **Permanencia**: el rasgo debe ser invariable con el tiempo.
- **Mensurabilidad:** el rasgo debe poder ser caracterizado cuantitativamente.
- **Rendimiento:** el proceso de identificación debe ser preciso
- **Aceptabilidad**: los usuarios deben estar dispuestos a emplear ese rasgo

2.1.2 Rasgos biométricos

En la actualidad se emplean una serie de rasgos biométricos en una gran variedad de aplicaciones. Cada uno de estos rasgos posee una serie de ventajas e inconvenientes que lo hacen más adecuado para unas aplicaciones que para otras. Por lo tanto, la elección de un rasgo biométrico a la hora de diseñar un sistema de reconocimiento biométrico se

realiza en función de sus características y de los requisitos de la aplicación deseada. Podemos dividir los rasgos biométricos en rasgos fisiológicos (características intrínsecas a la naturaleza física de la persona, como el iris o la huella dactilar) y rasgos de comportamiento, que se basan en el modo en que una persona realiza una acción determinada. En este segundo grupo encontramos la escritura, el rasgo biométrico utilizado en el proyecto.

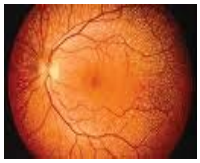
A continuación se presenta una muy breve introducción a algunos de los rasgos biométricos más utilizados, por orden alfabético:



ADN. Excepto en el caso de gemelos monocigóticos, el ADN presenta una distintividad suficiente como para poder englobarlo dentro de la categoría de rasgos biométricos. Su principal ventaja es su alto poder discriminante. Por el contrario, se considera fácil de robar y conlleva problemas de privacidad (puede revelar enfermedades o discapacidades que el usuario no desee que se conozcan). Es utilizado principalmente en aplicaciones forenses de reconocimiento.

Cara. Entre las ventajas del reconocimiento biométrico basado en imágenes de la cara encontramos su alto grado de aceptación y facilidad de captura (simplemente se necesita una fotografía). Por el contrario, la variabilidad con el paso del tiempo se presenta como la principal desventaja.

Dinámica del tecleo. Este rasgo biométrico se basa en la hipótesis de que cada persona tiene una forma característica de teclear. Es un rasgo conductual y, por tanto, muy variable en el tiempo. Principalmente se utiliza para identificación en casos sencillos.

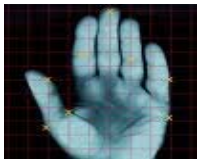


Escáner de retina. La estructura vascular de la retina es considerada como el rasgo biométrico más seguro, por su dificultad para duplicarlo. Este rasgo se supone diferente para cada usuario y para cada ojo. Para su captura requiere alta cooperación por parte del usuario, así como contacto con el sensor, lo cual compromete su aceptabilidad. Al igual que el ADN, puede conllevar problemas de privacidad, al revelar ciertas afecciones médicas.

Escritura. La escritura se engloba dentro de los rasgos biométricos de comportamiento, por lo que es variable a lo largo del tiempo. No tiene tan alto poder discriminante como otros rasgos como el ADN o el iris, pero su captura no es invasiva, por lo que su grado de aceptabilidad es alto.

Firma. La firma es un rasgo biométrico común con una alta aceptación por parte de la sociedad. Al igual que la escritura, es un rasgo comportamental, por lo que está sujeta a cambios a lo largo del tiempo y su captura puede verse afectada por condiciones físicas o emocionales. En la actualidad, con la eclosión de dispositivos móviles como los *Tablet PC* o *smartphones*, el uso de la firma online como medio de identificación está aumentando.

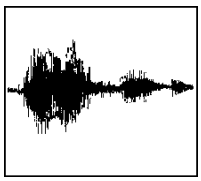
Forma de caminar. Es un rasgo biométrico poco distintivo, pero puede ser suficiente en aplicaciones que requieran un bajo nivel de seguridad. Es variable a lo largo del tiempo, y para su captura es suficiente con una cámara de vídeo.



Geometría de la mano. Se trata de un rasgo biométrico basado en diversas medidas relacionadas con la geometría de la mano (longitud y anchura de los dedos, distancias entre puntos clave de la mano…). Para la captura basta con situar la mano encima de un escáner o dispositivo similar. Es útil cuando se requiere poco espacio de almacenamiento, pues las imágenes necesitan poco espacio en memoria para ser almacenadas.

Huella dactilar. En la actualidad la huella dactilar es el rasgo biométrico más empleado. Consiste en un patrón de crestas y valles formado en la superficie del dedo y que es capturado al presionar éste contra un sensor. Cuenta entre sus principales ventajas con un alto grado de unicidad y aceptabilidad.

Iris. Es un rasgo con un alto poder de discriminación. Su captura requiere cooperación por parte del usuario, que debe colocar el ojo a una distancia determinada del sensor, por lo que se considera intrusivo. Además requiere ciertas condiciones de iluminación para que la imagen capturada posea mayor calidad.



Voz. A pesar de su menor distintividad y de la facilidad para ser imitada, la voz es un rasgo biométrico fácil de obtener y ampliamente aceptado. Este rasgo es una combinación de características físicas y de conducta. Éstas últimas son variables a lo largo del tiempo, al verse afectadas por factores como la edad o el estado de ánimo.

Además de los rasgos descritos anteriormente, existen otros rasgos biométricos menos estudiados, como la forma de la oreja, las venas de la mano o el olor, cuyo uso se está extendiendo.

2.1.3 La biometría en el ámbito forense

La biometría es utilizada en la actualidad en diversos campos donde se requiere la identificación de personas, como por ejemplo en controles de acceso o para identificarse en un ordenador portátil. Sin embargo, aparte de esos usos comerciales y de seguridad, la biometría también puede ser de aplicación en el ámbito forense.

Llamamos ciencia forense a la aplicación de la ciencia o la tecnología en la investigación de actividades criminales y al establecimiento de los hechos o evidencias en un tribunal [19,20]. Dentro de la ciencia forense, la biometría juega un papel muy importante, pues ayuda a determinar la identidad de una persona o a asociar a un individuo con una fuente desconocida. Los científicos forenses han demostrado que los rasgos físicos y de comportamiento pueden informar sobre la identidad de personas implicadas en crímenes [20].

En el ámbito general, la biometría se utiliza en sistemas automáticos de reconocimiento con el objetivo de identificar personas a partir de sus rasgos [14], pero dentro del ámbito forense en la mayoría de los casos la biometría no tiene esa aplicación, sino que se emplea como herramienta de filtrado automático de grandes bases de datos, que permite posteriormente a expertos humanos realizar comparaciones más precisas. [19]

Además, existen diferencias entre la biometría comercial y la biometría forense relacionadas con la calidad de las muestras. En términos generales, la calidad de las

marcas biométricas en biometría forense es mucho menor que en la biometría comercial, por lo que el poder de discriminación disminuye de forma considerable.

Las situaciones en las que, en general, se emplea la biometría forense son [21]:

1. Cuando se necesita identificar a una persona, viva o muerta

2. Cuando en un lugar de interés se encuentran marcas de rasgos biométricos y se desea conocer la identidad de la fuente.

3. Cuando deseamos relacionar dos o más marcas biométricas para saber si pertenecen a la misma fuente.

Entre los rasgos biométricos que se utilizan para la identificación de personas en entornos forenses podemos encontrar:

- Grabaciones de voz
- Huellas dactilares, palmares, o de los pies
- Imágenes y vídeos de individuos
- Notas manuscritas
- Manchas de sangre, saliva u otros fluidos, de los que puede extraerse muestras de ADN

2.2 Sistemas automáticos de reconocimiento

Un sistema biométrico es un reconocedor de patrones cuya operativa consiste principalmente en los siguientes cuatro pasos:

1. Captura de un rasgo biométrico
2. Extracción de características del rasgo biométrico
3. Comparación con patrones almacenados en una base de datos.
4. Decisión sobre la identidad del individuo.

2.2.1 Arquitectura genérica de un sistema biométrico

El esquema general de un sistema de reconocimiento biométrico se muestra en la Figura 2.1. Los módulos marcados en línea discontinua son opcionales.

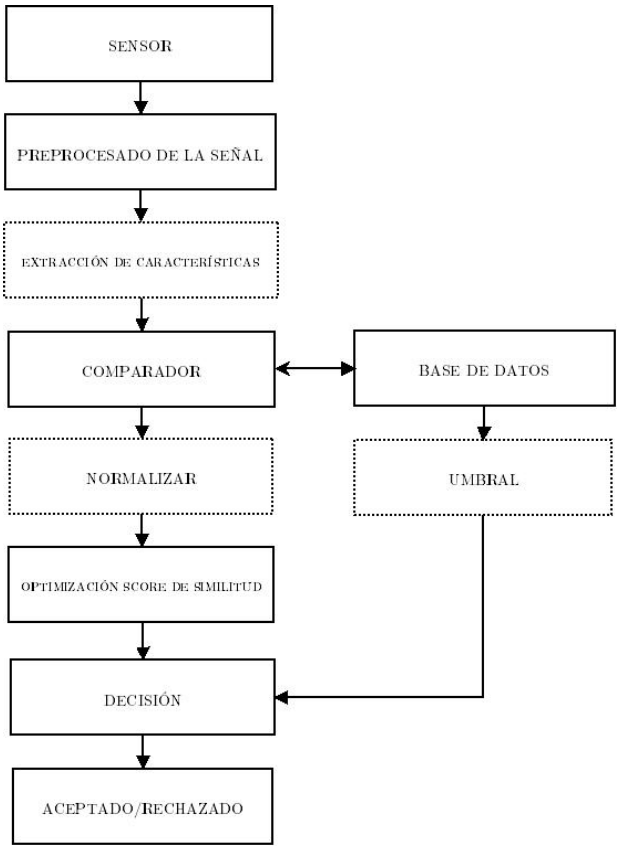


Figura 2.1. Arquitectura genérica de un sistema biométrico.

El **sensor** suele ser el único punto del sistema al que tiene acceso el usuario. Se utiliza para poder digitalizar el rasgo biométrico utilizado. Es un proceso determinante, pues la cantidad y calidad de la información que se obtiene dependen de él. En ocasiones se realiza posteriormente a la captura un **preprocesado de la señal,** para poder adaptarla a las siguientes etapas del sistema y eliminar los posibles ruidos y distorsiones que se hayan producido en la adquisición.

En la fase de **extracción de características** se obtiene la información relevante del rasgo biométrico, eliminando aquella que no resulte útil en el proceso de reconocimiento por no ser suficientemente discriminante.

Una vez extraído el vector de características del rasgo biométrico capturado, éste se **compara** con los vectores ya existentes en la base de datos del sistema. Dicha base de datos puede estar almacenada en un lugar único centralizado o por el contrario cada usuario puede poseer una tarjeta inteligente con el modelo de su identidad. Al comparar dos vectores de características obtenemos una puntuación o *score,* que mide el grado de similitud entre los dos vectores comparados. Esta puntuación puede sufrir un proceso de **normalización** y/o **optimización.**

En base a la puntuación obtenida y al umbral que exista en el sistema, el módulo de **decisión** indicará si los dos patrones comparados pertenecen al mismo usuario o no.

2.2.2 Modos de operación de los sistemas biométricos

Existen tres modos principales de trabajo de un sistema biométrico desde el punto de vista de su funcionamiento:

a) Modo registro. Es el modo utilizado para dar de alta a los usuarios en el sistema. Para ello, se capturan una o varias realizaciones del rasgo biométrico a almacenar, se extraen las características requeridas y se genera la plantilla del usuario que quedará almacenada en la base de datos. En la Figura 2.2 podemos ver el esquema de funcionamiento en este modo.

b) Modo verificación. Cuando un sistema biométrico trabaja en este modo realizará comparaciones 1 vs. 1 entre una plantilla de la base de datos y el rasgo adquirido por el sensor en cada momento. Las entradas del sistema en este modo son dos: la realización del rasgo biométrico a identificar y la solicitud de identidad (esta última determinará con qué plantilla de la base de datos se debe comparar la realización). La salida será "aceptación" o "rechazo", indicando respectivamente si el usuario es o no es quien dice ser. En la Figura 2.3 podemos ver el esquema de funcionamiento de un sistema biométrico en modo verificación.

c) Modo identificación. Este modo se utiliza para clasificar una realización determinada de un rasgo biométrico cuya identidad se desconoce como perteneciente a uno de entre un conjunto de N individuos. En lugar de realizar comparaciones 1 vs. 1 como sucedía en el modo verificación, en este caso se realiza una comparación 1: N entre la realización capturada por el sensor y los vectores de características de los N usuarios almacenados en la base de datos. Este modo de funcionamiento es el más utilizado en el ámbito forense, pues lo que se busca a menudo es averiguar si una muestra obtenida pertenece a algún individuo de la base de datos con la que se trabaja. En la Figura 2.4 podemos ver el esquema de funcionamiento de un sistema biométrico en modo identificación.

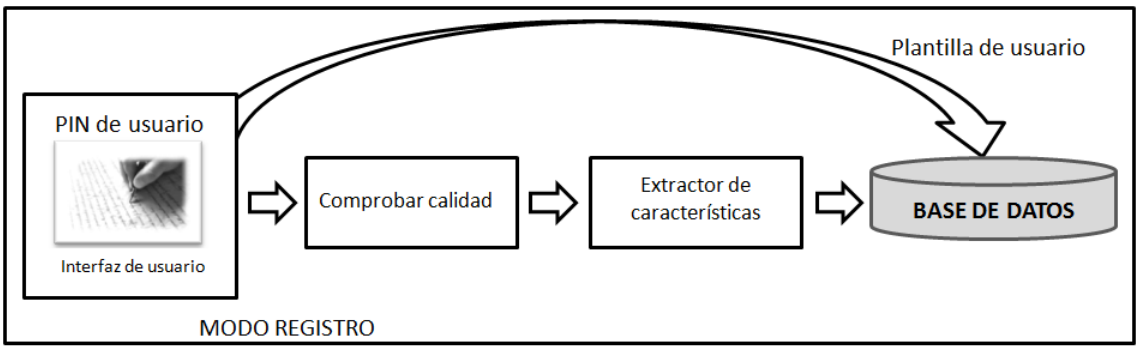


Figura 2.2. Esquema de funcionamiento de un sistema biométrico en modo registro

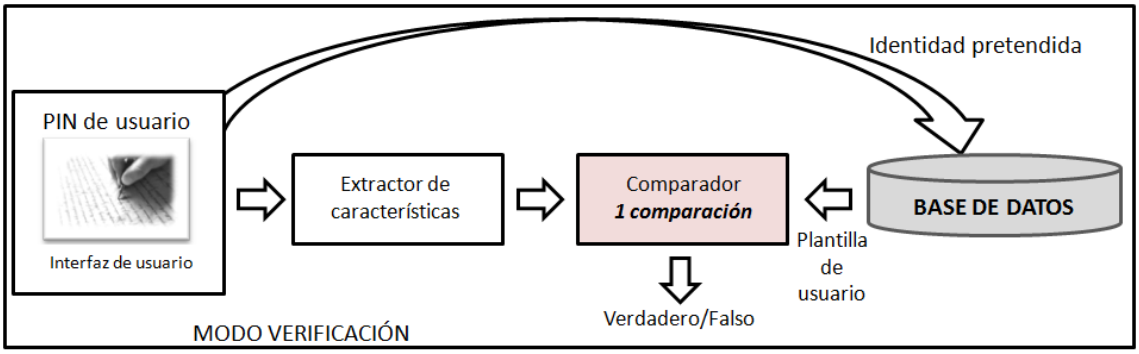


Figura 2.3. Esquema de funcionamiento de un sistema biométrico en modo verificación

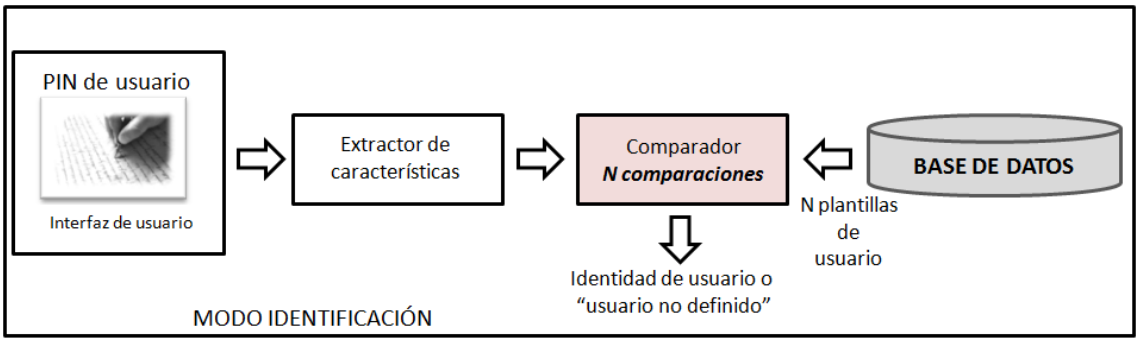


Figura 2.4. Esquema de funcionamiento de un sistema biométrico en modo identificación

2.2.3 Rendimiento de los sistemas biométricos

Al trabajar con sistemas biométricos, es necesario disponer de procedimientos que permitan medir la fiabilidad de dichos sistemas, para de esta manera poder evaluar nuevas mejoras o desarrollos, y para poder comparar el rendimiento de dos sistemas biométricos distintos.

Como hemos visto al analizar la arquitectura general de un sistema biométrico, existe un bloque comparador en el que se mide la similitud entre la muestra de entrada y una muestra de la base de datos. La salida de este bloque es una puntuación o *score*, que nos permite cuantificar el parecido entre ambas muestras. En la medida en que las muestras sean más parecidas, mayor será la puntuación que devuelva el comparador. A partir de esta puntuación y del umbral de trabajo del sistema, se decidirá si la muestra de entrada y la muestra de la base de datos pertenecen o no al mismo usuario. Por lo tanto, el umbral marca la decisión del sistema.

Fijado un umbral, todas las puntuaciones que sean superiores a dicho valor, serán consideradas por el sistema como usuarios aceptados, mientras que las puntuaciones menores pertenecerán a usuarios impostores, es decir, que no se encuentran en la base de datos. En un sistema biométrico ideal, la distribución de puntuaciones de usuarios y de impostores no se solaparía, por lo que el error del sistema sería nulo. Sin embargo, en un sistema real existen errores que provocan que usuarios reales sean rechazados (FR – Falso Rechazo // FRR – *False Rejection Rate*), y que usuarios impostores sean aceptados (FA – Falsa aceptación // FAR – *False Acceptance Rate*).

En la Figura 2.5 observamos una posible distribución de puntuaciones de usuarios y de impostores para un sistema biométrico determinado. Se observa que existe una región donde hay solapamiento de ambas distribuciones. En concreto, el área bajo la curva de impostores que queda por encima del umbral será la probabilidad de que un impostor sea aceptado (Tasa de Falsa Aceptación) y el área bajo la curva de usuarios que queda por debajo del umbral será la probabilidad de que un usuario se rechazado (Tasa de Falso Rechazo).

En función del umbral, la FAR y la FRR varían. El punto en que ambas son iguales se conoce con el nombre de EER (*Equal Error Rate*) y se suele utilizar para comparar el rendimiento entre sistemas biométricos.

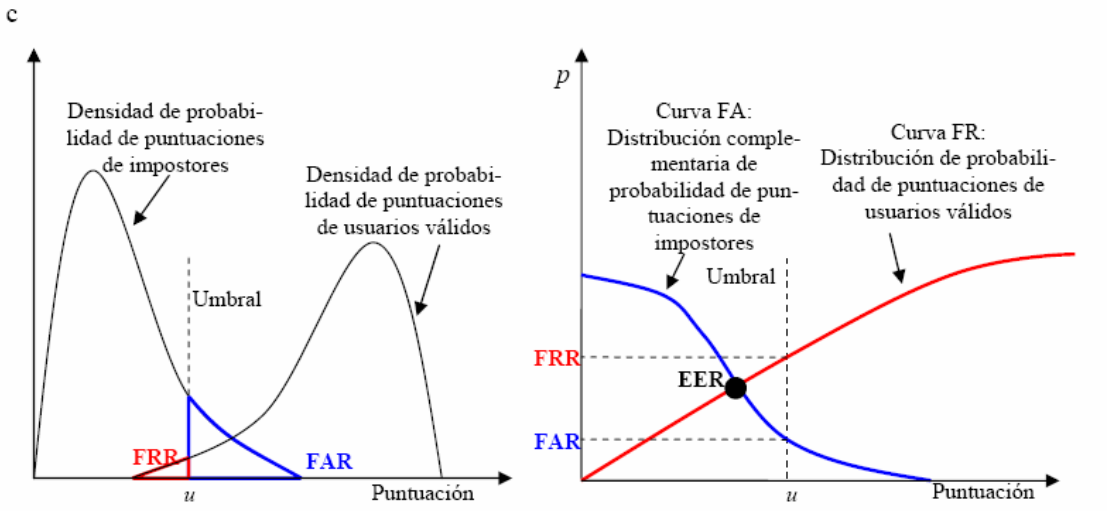


Figura 2.5. Densidades y distribuciones de probabilidad de usuarios e impostores

Otra representación habitual del rendimiento de un sistema de reconocimiento biométrico son las curvas DET (*Detection Error Tradeoff*) [22], que enfrentan la FRR y la FAR, permitiendo por tanto observar los diversos puntos de trabajo del sistema, como se puede ver en la Figura 2.6.

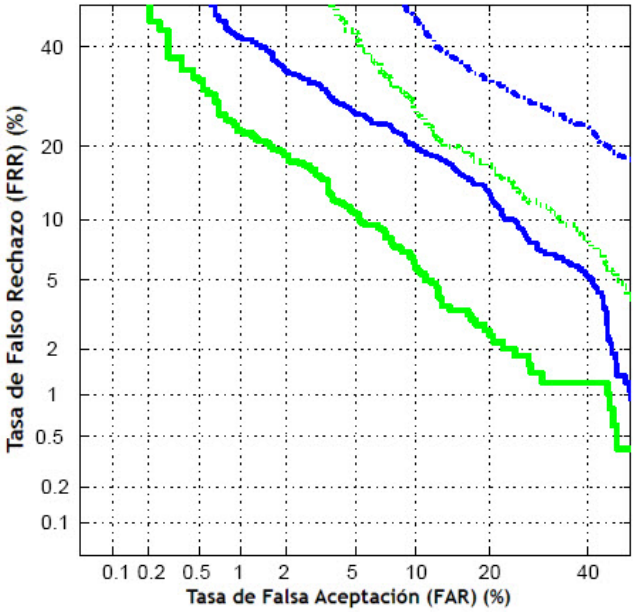


Figura 2.6. Ejemplo de curvas DET

Lo visto hasta ahora en este apartado hace referencia a la evaluación del rendimiento de sistemas en modo verificación, en los cuales la salida del sistema puede ser sólo o aceptación o rechazo, y en los cuales existe un umbral que marca el punto de trabajo del sistema. Sin embargo, en los sistemas en modo identificación, como en el caso del sistema desarrollado en el presente proyecto, no existe un umbral ya que se compara la muestra de entrada con todas las existentes en la base de datos, obteniendo a la salida una clasificación de identidades probables. En este modo de funcionamiento, por lo tanto, los sistemas se suelen caracterizar por la tasa de acierto en función del "Top N", es decir, la probabilidad de que la identidad correcta aparezca en las primeras N posiciones de la lista ordenada de salida.

En este caso, puede ser interesante estimar la probabilidad de acierto para un Top N fijo, o por otra parte, estimar para qué Top N se alcanza cierta probabilidad de acierto. En el primero de los casos, a mayor probabilidad de acierto para un Top N fijado, el rendimiento del sistema será mayor. En el segundo, el mejor rendimiento vendrá dado por un menor Top N respecto a una probabilidad fija de acierto.

2.2.4 Multimodalidad biométrica

Algunos de los problemas introducidos por los sistemas biométricos unimodales pueden ser resueltos utilizando múltiples evidencias de la identidad de la persona. Estos sistemas reciben el nombre de sistemas biométricos multimodales y utilizan más de una característica fisiológica o de comportamiento para el reconocimiento. El principal beneficio de la multimodalidad biométrica es el aumento de la precisión y de la seguridad frente a ataques basados en suplantación de identidad.

Los sistemas multimodales utilizan la multiplicidad en alguno de los siguientes escenarios. En la Figura 2.7 se muestra un ejemplo de ello.

1. Múltiples sensores. El mismo rasgo biométrico es capturado utilizando diferentes sensores, y la información capturada por cada uno de ellos es después combinada. Por ejemplo, en un sistema de reconocimiento basado en huella dactilar, se pueden capturar las huellas mediante sensores ópticos, de estado sólido, etc. [23]

2. Múltiples rasgos biométricos. El reconocimiento está basado en múltiples rasgos biométricos de un mismo individuo, como por ejemplo la escritura y la huella dactilar. La independencia entre cada fuente mejora la robustez del sistema.

3. Múltiples instancias del mismo rasgo. Los seres humanos poseen algunas características biométricas "repetidas", como los ojos, los dedos o las manos. El reconocimiento puede realizarse en este caso combinando la información de dos o más dedos, dos o más iris, etc.

4. Múltiples capturas del mismo rasgo. La combinación de la información adquirida por múltiples capturas del mismo rasgo biométrico puede resultar en una más completa y precisa descripción del individuo. Varias muestras de voz o varias imágenes del iris son un ejemplo de este tipo de multiplicidad.

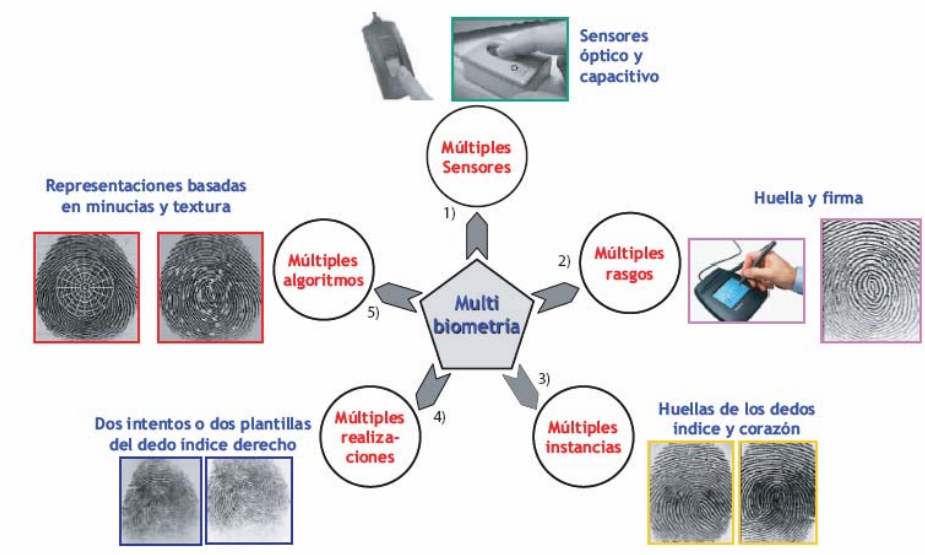


Figura 2.7. Sistema multimodal basado en el uso de múltiples rasgos biométricos

5. Múltiples representaciones/comparaciones del mismo rasgo. Un único rasgo biométrico puede ser analizado mediante diferentes extractores de características, todos ellos trabajando con los mismos datos de entrada pero

procesándolos de formas diferentes. Por lo tanto, cada uno de ellos genera su propio vector de características, que pueden ser comparados separadamente y posteriormente las puntuaciones obtenidas por los distintos clasificadores pueden ser combinadas para tomar una decisión.

Además, un sistema multimodal puede operar en tres modos diferentes a la hora de combinar la información de las diversas fuentes:

1. Modo serie. La salida del análisis de una fuente biométrica es utilizada como entrada para la siguiente, reduciendo el número de identidades posibles en cada paso. Esto implica que no todos los rasgos deben necesariamente ser adquiridos de forma simultánea y que una decisión puede ser tomada incluso antes de que todas las características hayan sido capturadas, si se ha alcanzado suficiente certeza sobre la identidad buscada.

2. Modo paralelo. La decisión se basa en la información de todas las características biométricas capturadas. Esto implica que todas las capturas deben ser realizadas antes de que el proceso concluya. Este modo es muy robusto, especialmente en sistemas basados en identificación.

3. Modo jerárquico. Clasificadores individuales son combinados en una estructura tipo árbol. Este esquema combina las ventajas de los dos modos anteriores, y supone una gran mejora cuando el número de clasificadores es grande.

Por último, los sistemas biométricos multimodales pueden combinar la información que obtienen de múltiples sistemas a diferentes niveles:

- A nivel de extracción de características (Figura 2.8), combinando los diferentes vectores de características obtenidos.

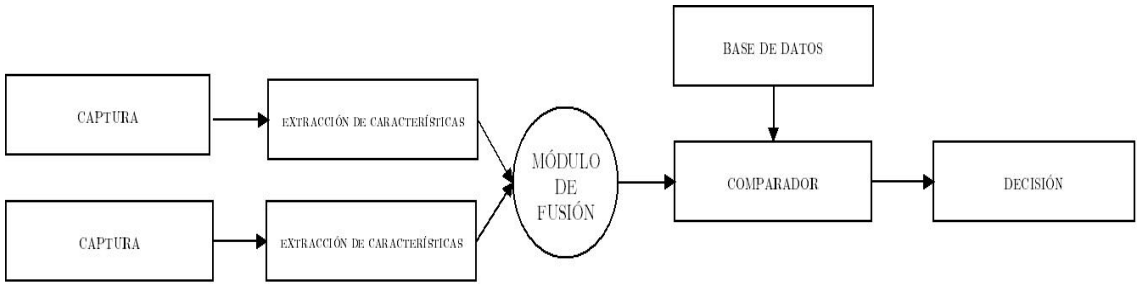


Figura 2.8. Esquema de sistema multimodal que combina la información a nivel de extracción de características

- A nivel de *score* (Figura 2.9), donde las diferentes puntuaciones obtenidas en los sistemas unimodales se combinan, por ejemplo, utilizando la media.

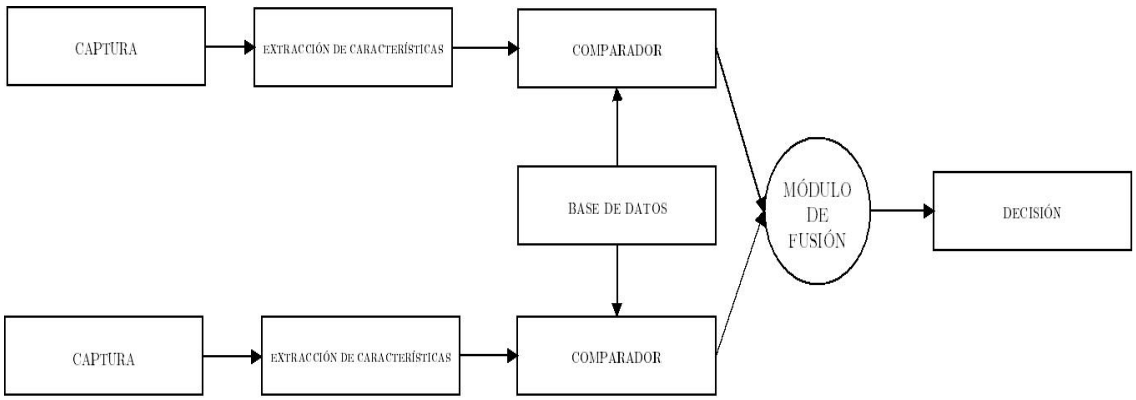


Figura 2.9. Esquema de sistema multimodal que combina la información a nivel de score

- A nivel de decisión (Figura 2.10), combinando las diferentes decisiones de aceptado/rechazado obtenidas en cada sistema utilizando, por ejemplo, la regla de la mayoría.

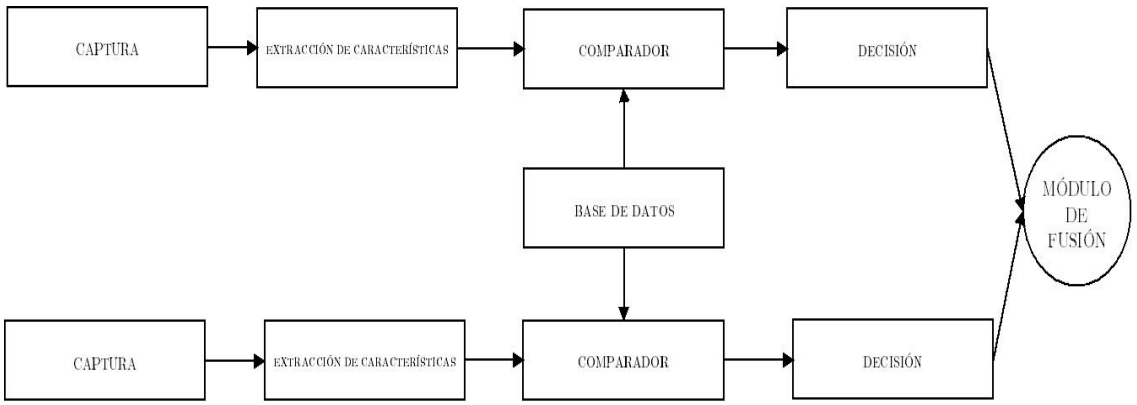


Figura 2.10. Esquema de sistema multimodal que combina la información a nivel de decisión

3

Reconocimiento biométrico de escritor

Los rasgos biométricos, como se comentó en el capítulo anterior, se clasifican en dos categorías: rasgos biométricos fisiológicos, que identifican a un usuario basándose en la medida de una característica física del cuerpo humano, y rasgos biométricos de comportamiento o conducta que emplean características individuales del comportamiento de un individuo para su identificación. La identificación de escritor pertenece a esta segunda categoría. En este capítulo analizamos los aspectos más importantes de este tipo de sistemas.

3.1 Reconocimiento de escritor vs. Reconocimiento de escritura

El objetivo del reconocimiento de escritura es buscar representaciones que sean capaces de eliminar variaciones entre distintas escrituras para poder clasificar la forma de los caracteres y palabras de una manera robusta. Es decir, lo que se busca identificar es qué se ha escrito, y no quién lo ha escrito. Por el contrario, en el reconocimiento de escritor lo que se busca es realzar dichas variaciones, características de cada escritor, con el objetivo de poder identificar quién ha escrito el texto, en lugar de identificar qué se ha escrito. En resumen, el reconocimiento de escritura consiste en obtener generalizaciones, eliminando las variaciones, mientras que el reconocimiento de escritor consiste en realzar las variaciones específicas de cada estilo de escritura individual.

El reconocimiento de escritor ha cobrado importancia en los últimos años, principalmente debido a sus aplicaciones en el ámbito forense y en el análisis de documentos históricos, pese a que históricamente ha sido el reconocimiento de escritura el que ha tenido más peso en el área de análisis de la escritura [24]

3.2 Verificación de escritor vs. Identificación de escritor

Aplicando lo visto en el capítulo anterior respecto a los diferentes modos de operación de un sistema biométrico:

- En un sistema de verificación de escritor se produce una comparación uno a uno para decidir si dos muestras tienen o no el mismo autor, donde por lo general se sabe la identidad del autor de una de ellas (muestra de referencia). Si el parecido entre las dos muestras elegidas supera un cierto umbral predefinido, se considera que ambas muestras han sido escritas por el mismo autor. Si no lo supera, se considera que ambas muestras pertenecen a autores distintos.

- En un sistema de identificación de escritor se realiza una búsqueda de uno a muchos en una base de datos con muestras de escritura de autores conocidos, devolviéndose una lista de posibles candidatos. Esta lista suele devolverse ordenada de forma que el parecido con la muestra de entrada va disminuyendo.

En la Figura 3.1 se muestra el esquema de funcionamiento genérico de un sistema de verificación de escritor y un sistema de identificación de escritor.

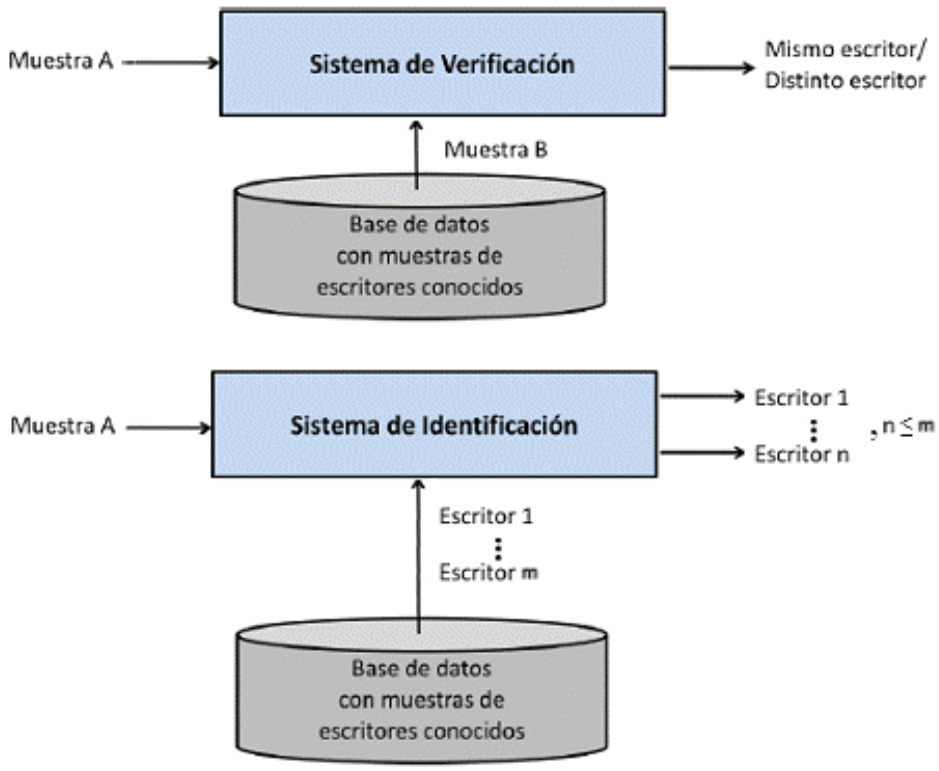


Figura 3.1. Esquemas de funcionamiento general de un sistema de verificación de escritor (arriba) y un sistema de identificación de escritor (abajo)

3.3 Escritura estática (*off-line*) vs. Escritura dinámica (*on-line*)

Otra de las clasificaciones que se pueden realizar en el reconocimiento de escritor, ya sea a través de muestras de escritura o a partir de la firma, es en reconocimiento de escritura estática (*off-line*) y reconocimiento de escritura dinámica (*on-line*). [25]

Esta división se realiza en base al dispositivo de captura y al tipo de información capturada. Los dispositivos *on-line* suelen ser tabletas digitalizadoras sensibles a la presión, que permiten registrar información dinámica sobre la escritura, como la velocidad, ángulo, posición o presión del lápiz, obteniendo por tanto señales dependientes del tiempo que pueden ser utilizadas en el proceso de reconocimiento.

Por otra parte, en el reconocimiento de escritura estática se obtiene una información más limitada, y normalmente se basan en el uso de algún tipo de escáner, que captura una imagen del texto manuscrito. A partir de dicha imagen escaneada, se extraen características de la imagen que permiten discriminar entre escritores. Este proyecto se basa en este último tipo de reconocimiento de escritura.

En la Figura 3.2 podemos ver ejemplos de ambos tipos de sistema, observando la diferencia entre ambos.

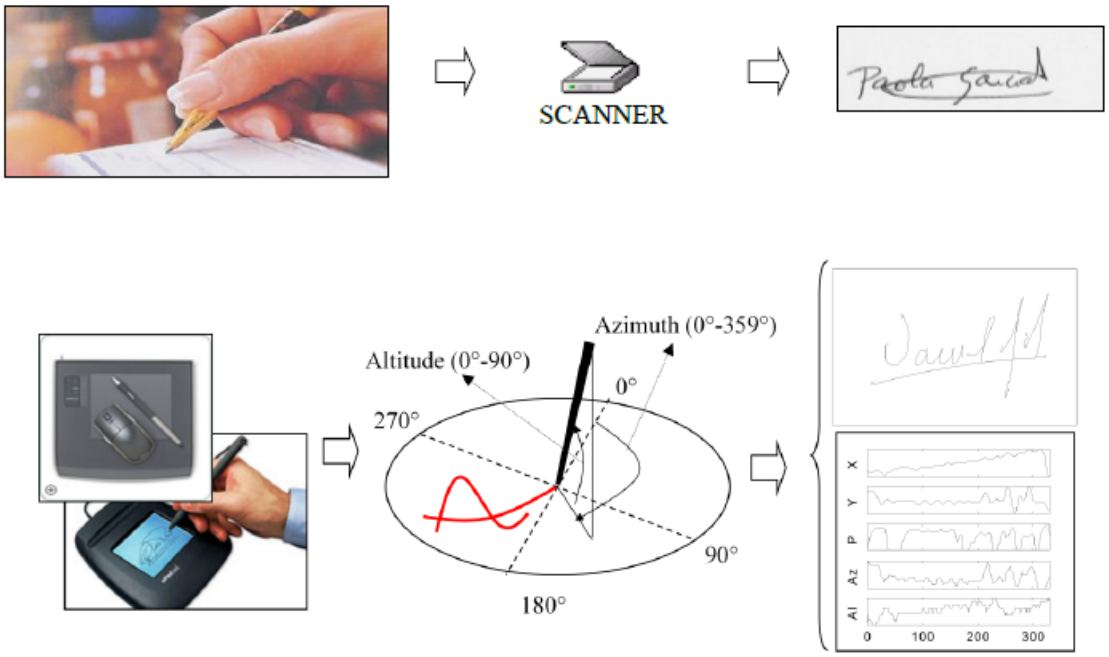


Figura 3.2. Ejemplos de sistemas de adquisición de escritura *off-line* (arriba) y *on-line* (abajo)

3.4 Reconocimiento independiente de texto vs. Reconocimiento dependiente de texto

Dentro del reconocimiento de escritor podemos encontrar dos categorías: reconocimiento independiente de texto y reconocimiento dependiente de texto. [24]

El reconocimiento independiente de texto utiliza características de la imagen entera de un bloque de texto, proporcionando un análisis global de la región escrita. En este tipo de reconocimiento es necesaria una suficiente cantidad de escritura para poder extraer características estables e independientes del texto escrito en las muestras. La ventaja consiste en que la intervención humana y la complejidad se reducen. En sistemas de reconocimiento utilizados en entornos forenses el reconocimiento independiente de texto es el más común, pues las muestras de escritura disponibles en este tipo de entornos normalmente corresponden a muestras intervenidas o requisadas, por lo que no existe ningún control sobre el contenido de las mismas.

Por otra parte, el reconocimiento dependiente de texto sí que tiene en cuenta el contenido semántico de las muestras de escritura, por lo que se requiere una previa segmentación y localización de la información deseada, lo que los hace más complejos. La gran ventaja del reconocimiento dependiente de texto es que permiten alcanzar mayor rendimiento, aún con pequeñas cantidades de muestras de escritura.

3.5 Variabilidad en la escritura

Existen una serie de factores que pueden producir variabilidad en la escritura de un individuo [24]. Podemos ver ejemplos de estos factores de variabilidad en la escritura en la Figura 3.3.

- Transformaciones afines. En este grupo englobamos los cambios en tamaño e inclinación, las rotaciones y las traslaciones. No suponen un gran obstáculo para la identificación de escritor.

- Variabilidad neuro-mecánica. Se deriva principalmente del estado fisiológico del escritor y del contexto local. Como tal, esta variabilidad está más relacionada con el estado del escritor que con su identidad, pero puede obstaculizar de modo importante el proceso de identificación.

- Variabilidad secuencial. Se refiere a la variación en el orden de los trazos que puede existir para un mismo carácter. También es muy dependiente del estado del escritor durante el proceso de escritura, aunque su efecto es más evidente en escritura *on-line*, donde existe información temporal del proceso de escritura.

- Variación alográfica. Hace referencia a la variedad de formas gráficas que existen para un mismo carácter. Este último factor es el utilizado en el sistema de reconocimiento desarrollado en el marco de este proyecto.

Además de estos factores, la edad constituye otra de las principales causas de variabilidad de la escritura. A lo largo de la vida, la escritura de una persona se vuelve más rápida, suave y continua, llegando incluso a verse afectada en las personas mayores, al perder fuerza y destreza en las manos.

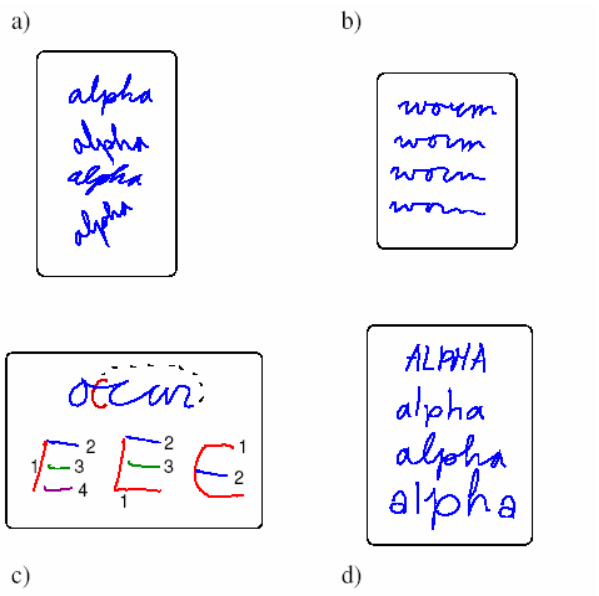


Figura 3.3. Factores de variabilidad en la escritura: transformaciones afines (a), variabilidad neuro-biomecánica (b), variabilidad secuencial (c) y variabilidad alográfica (d)

3.6 Individualidad de la escritura

La escritura puede describirse como un proceso psicomotor jerárquico en el que intervienen la memoria a largo plazo, procesos motores y los músculos y articulaciones. Escribir consiste en rápidos movimientos de los dedos y la mano, sumados a un proceso superpuesto de movimiento horizontal progresivo de la parte baja del brazo. [26]

A medida que se produce la maduración de una persona, su estilo de escritura se va alejando del estándar adquirido en el colegio y progresivamente va adquiriendo rasgos propios que dan individualidad a la escritura. Existen dos tipos de factores que contribuyen a esta individualidad:

- Factores genéticos o biológicos:
    - La configuración de los huesos y articulaciones de la mano y la muñeca
    - Ser zurdo o diestro
    - La fuerza muscular
    - Las propiedades del sistema nervioso central
- Factores miméticos o culturales. Determinados por la educación recibida, el entorno o la observación de la escritura de otras personas. Por ejemplo, el conjunto de alógrafos utilizados por una población de escritores viene fuertemente determinado por los métodos de escritura adquiridos en la escuela, que a su vez dependen de factores como la distribución geográfica o la religión.

3.7 Trabajos previos y algoritmos existentes para reconocimiento de escritor

El objetivo de este apartado es realizar una recopilación de algunos de los diversos algoritmos que han surgido en los últimos años en el ámbito del reconocimiento de escritor. Este análisis estará centrado únicamente en aquellos algoritmos que trabajan en modo *off-line* (estático), por ser ése el modo del sistema desarrollado en el presente proyecto.

En [27] podemos encontrar un resumen general de los primeros trabajos en torno al reconocimiento de escritor.

A partir de dicho resumen, y para realizar este análisis dividiremos los algoritmos restantes en tres bloques, en función del dominio en el que se encuentran las características o representaciones en las que se basa cada uno de ellos: dominio espacial, dominio frecuencial y dominio vectorial.

A. Dominio espacial

Este dominio es el más empleado hasta el momento a la hora de desarrollar sistemas de reconocimiento de escritor. Algunos de los sistemas que utilizan representaciones en el dominio espacial son los siguientes:

- En [28] y [29] encontramos sistemas basados en transformadas de patrón de arco, así como en características *off-line* de caracteres.

- El propuesto en [30], que realiza identificación a partir de medidas de *run-length*, es decir, secuencias continuas de puntos negros en horizontal o vertical.

- El sistema propuesto en [31] se basa en perfiles de proyección horizontal y operadores morfológicos.

- En [32] se realiza identificación de escritor utilizando características basadas en la línea de texto.

- La combinación de macro y micro características se realiza en el sistema descrito en [33]

- El sistema de [2], que emplea un conjunto de 21 características computacionales extraídas a los niveles de documento, párrafo, palabra y carácter. También se incluyen otras características como número de píxeles de tinta, número de contornos externos/internos, inclinación media y longitud de las palabras, de gradiente, estructurales y de concavidad. El sistema de referencia de gradiente usado en este Proyecto está basado en la implementación de [2].

- Los sistemas de [34], que utilizan trasformadas lineales y elementos direccionales sobre caracteres chinos.

- En el sistema presentado en [12] se utilizan características direccionales de los bordes.

- Por último, el sistema de características alográficas presentado también en [12], en el cual se basa la implementación del sistema desarrollado en este Proyecto.

B. Dominio frecuencial

Una alternativa al uso del dominio espacial en los sistemas de reconocimiento biométrico de escritor puede ser la representación frecuencial, que se basa en aplicar la teoría de Fourier, descomponiendo la imagen espacial de la escritura en un espectro de frecuencias que componen dicha imagen. Podemos encontrar un ejemplo de sistema que se basa en este tipo de representación en [35], donde se considera la imagen de escritura como una textura en lugar de cómo una colección de elementos de texto.

C. Dominio vectorial

Los métodos basados en la representación vectorial utilizan *splines* para representar los caracteres. Podemos definir las *splines* como bandas flexibles utilizadas para producir una curva suave a través de un conjunto de puntos prefijados. Aplicadas al reconocimiento de escritor, se definen para la forma de los caracteres un conjunto de

puntos críticos y parámetros, a partir de los que curvas paramétricas pueden aproximar la forma de los mismos. En [36] podemos ver un trabajo que se basa en este concepto.

4

Sistemas de reconocimiento de escritor

4.1 Sistema disponible evaluado (gradiente)

Como sistema de referencia a evaluar hemos utilizado el **sistema basado en características de gradiente** desarrollado en [39] a partir de la descripción de [2].

Las características de gradiente se basan en el cálculo del gradiente (derivada) de la imagen de la letra. Sea un punto (x,y) de la imagen con valor de imagen $f(x,y)$, el vector gradiente en dicho punto se obtendrá a partir de las derivadas parciales de $f(x,y)$ respecto de las componentes x e y (vertical y horizontal).

Para realizar este cálculo, utilizaremos una aproximación del cálculo del gradiente mediante operadores de Sobel [37].

A continuación se describen las diversas etapas del sistema de identificación.

4.1.1 Preprocesado

Para la extracción de las características del gradiente, partimos de las imágenes iniciales en escala de gris. El gradiente se puede obtener tanto de la imagen en escala de gris como de una imagen binarizada. Por tanto, en este sistema se comparará el uso de imágenes escaneadas en escala de gris con el de estas mismas imágenes escaneadas y binarizadas a continuación según el método de Otsu [38].

4.1.2 Extracción de características

Seguidamente se realizará la convolución con dos operadores de Sobel 3x3, que aproximan las derivadas parciales respecto de x e y en la posición del píxel de la imagen. Dado un píxel, se consideran sus píxeles vecinos según la Figura 4.1.

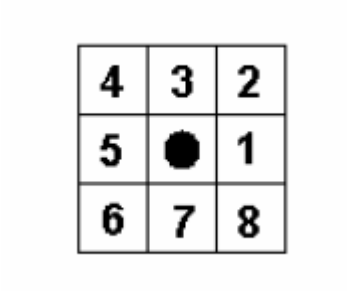


Figura 4.1. Píxeles vecinos a considerar al calcular el gradiente mediante operadores de Sobel

Los operadores de Sobel se consideran un tipo de filtros espaciales que, a diferencia de los filtros espectrales basados en descomposición de frecuencias de Fourier, se utilizan aplicando máscaras de coeficientes sobre un píxel y sus vecinos. Concretamente, los operadores de Sobel son filtros espaciales lineales, pues se aplican mediante una combinación lineal de los valores del píxel y los coeficientes de la máscara espacial.

Sea W la matriz utilizada como máscara espacial a aplicar y Z los valores de los píxeles de la imagen,

$$\mathrm{W} = \begin{bmatrix} w_1 w_2 w_3 \\ w_4 w_5 w_6 \\ w_7 w_8 w_9 \end{bmatrix} \quad \mathrm{Z} = \begin{bmatrix} z_1 z_2 z_3 \\ z_4 z_5 z_6 \\ z_7 z_8 z_9 \end{bmatrix}$$

el resultado de aplicar la máscara W sobre Z será:

$$R = \sum_{i=1}^{9} (w_i z_i)$$

Por tanto, con el objetivo de obtener el gradiente, deberemos aplicar dos máscaras a la imagen, una para la componente x del vector gradiente y otra para la componente y de dicho vector. El gradiente indicará cómo varía la imagen respecto de la dirección vertical y respecto de la dirección horizontal. El sistema de referencia de coordenadas utilizado se muestra en la Figura 4.2.

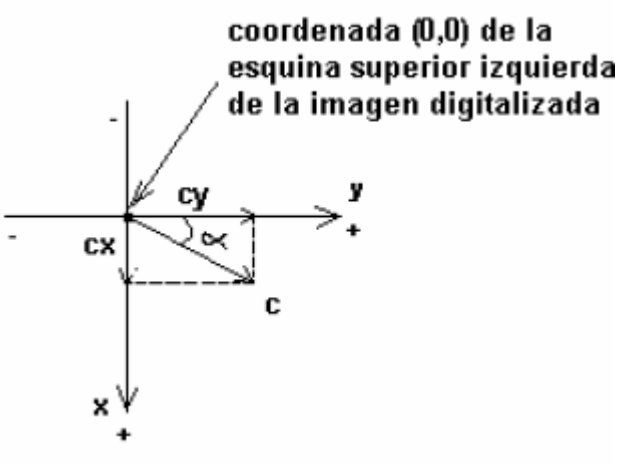


Figura 4.2. Sistema de referencia de coordenadas utilizado al calcular el gradiente de la imagen

La máscara de Sobel a utilizar para obtener la componente x (vertical) del gradiente es:

$$\begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

La máscara de Sobel a utilizar para obtener la componente y (horizontal) del gradiente es:

$$\begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

De esta manera, si el vector gradiente en un punto es $\mathbf{c} = (c_x, c_y)$, podremos expresar sus componentes como

$$c_x = |c| \cdot \text{sen}(\alpha)$$
$$c_y = |c| \cdot \cos(\alpha)$$

y definir la dirección del gradiente con el ángulo α, que vendrá determinado por:

$$\tan(\alpha) = c_x \,/\, c_y$$
$$\alpha = \tan^{-1}(c_x \,/\, c_y)$$

Atendiendo a las máscaras de Sobel aplicadas, las componentes del gradiente vendrán determinadas por los valores de los píxeles de la imagen, según las siguientes fórmulas:

$$c_x = (z7 + 2\ z8 + z9) - (z1 + 2\ z2 + z3)$$
$$c_y = (z3 + 2\ z6 + z9) - (z1 + 2\ z4 + z7)$$

Además, es importante reseñar que debido a que el ángulo α tendrá dos soluciones posibles, ya que la tangente asociada tiene período π radianes, para saber en qué cuadrante estamos y por tanto qué dirección considerar habrá que tener en cuenta los signos de c_x y c_y.

El vector gradiente posee dirección y módulo, pero sólo la dirección es utilizada en el vector de características. Por otra parte, si bien la dirección puede variar entre 0 y 2π radianes, en este sistema de reconocimiento se particionará el espacio de direcciones en 12 regiones posibles, considerándose únicamente los valores múltiplos de $\pi/6$, de acuerdo a la Figura 4.3, y cuantificando el ángulo en un valor entero entre 0 y 11 como se indica en dicha figura.

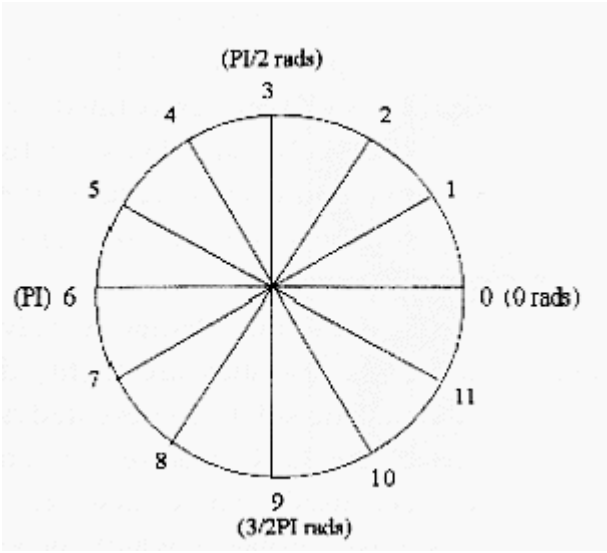


Figura 4.3. Direcciones del vector gradiente normalizadas

4.1.3 Caracterización e identificación de cada usuario

4.1.3.1 Modelado de identidad de usuario

En primer lugar la imagen de la letra se divide en 4x4 celdas como se aprecia en la Figura 4.4. Calcularemos el gradiente en cada una de dichas celdas de forma individual, obteniendo posteriormente un histograma de cuántas veces se da cada región de dirección posible. Obtendremos, por tanto, un histograma de 12 posiciones en cada celda, y un vector de características total de 192 componentes (12 posiciones/celda x 16 celdas).

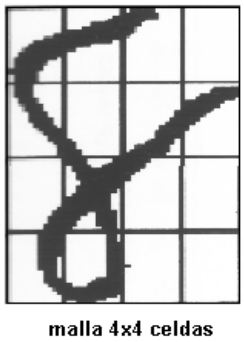


Figura 4.4. Mallado de 4x4 celdas realizado sobre la imagen de cada carácter.

En la Figura 4.5 se muestra un ejemplo del mapa de gradiente de una letra.

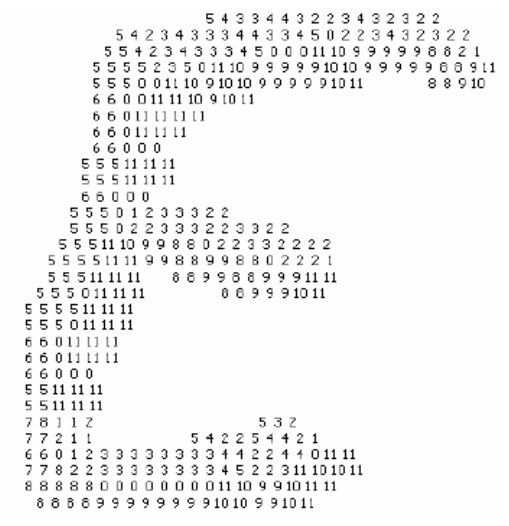


Figura 4.5. Ejemplo de mapa de gradiente de una letra

En este módulo del sistema existen dos posibilidades a la hora de obtener el vector de características:

- Obtener un vector binario. Para ello, sobre el histograma de direcciones de cada celda se calcula un umbral para la binarización, decidiendo a partir de él, si una dirección está presente en la celda (bit = 1) o no lo está (bit = 0). De esta manera obtendremos un vector binario de 192 componentes, como se propone en el sistema descrito en [2].

- Obtener un vector no binario, a partir de la normalización de los histogramas de direcciones de cada celda a una función densidad de probabilidad. De esta forma en el vector de características se modela la probabilidad de cada dirección

dentro de una celda, no únicamente si existe o no esa dirección como ocurre al utilizar vectores binarios.

En la Figura 4.6 se muestra un ejemplo de ambos métodos.

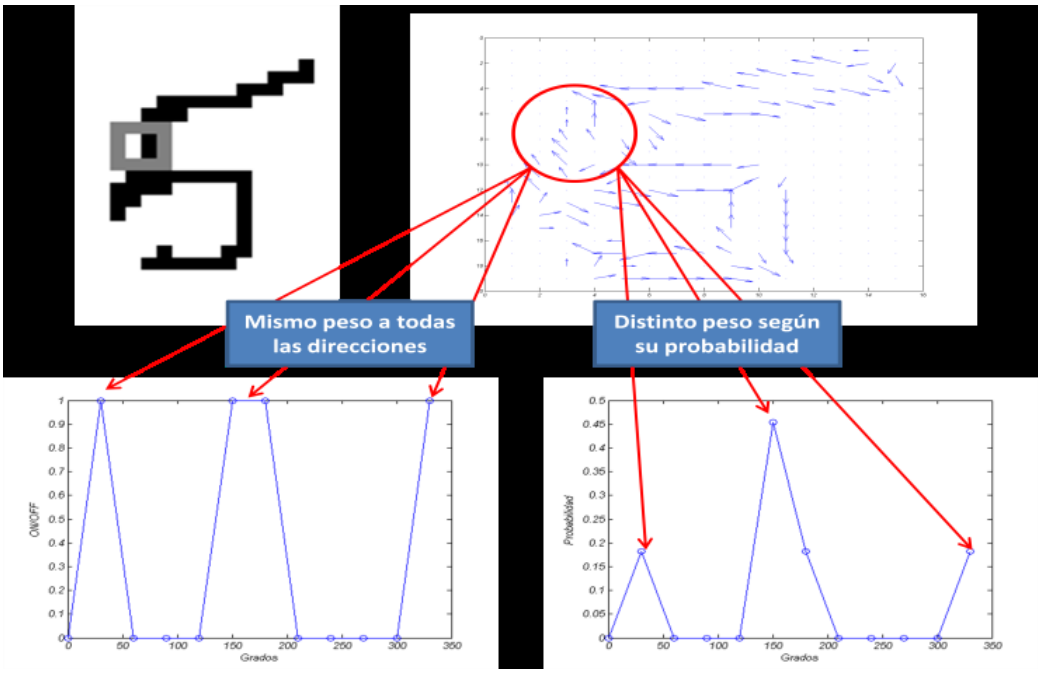


Figura 4.6. Ejemplo de aplicar binarización o no a la hora de obtener el vector de características en una celda de la imagen.

4.1.3.2 Cálculo de distancias

Una vez modelada la identidad de cada usuario mediante su vector de características, necesitamos un método para medir la similitud entre los vectores de características de dos usuarios dados. En base a si el vector de características es o no binario se aplicará una medida de distancia determinada.

Para vectores de características binarios, se utilizará la distancia Euclídea. Siendo a y b dos vectores de características binarios con n componentes, la medida de distancia será:

$$d_{euclidea} = \sqrt{\sum_{i=1}^{n} a_i^2 - b_i^2}$$

En el caso de vectores de características no binarios (función densidad de probabilidad), la medida de distancia a utilizar será la conocida como χ^2, que en [12] demostró tener mejor rendimiento para estos casos.

$$d_{X^2} = \sum_{i=1}^{n} \frac{(a_i - b_i)^2}{a_i + b_i}$$

4.2 Sistema desarrollado en el marco de este PFC

En este proyecto se ha desarrollado e implementado un **sistema de reconocimiento de escritor basado en características alográficas**

El sistema de reconocimiento de escritor implementado en este proyecto es una adaptación del sistema presentado en [12]. En este sistema se considera al escritor como generador estocástico de alógrafos (formas escritas), utilizándose para caracterizar al escritor la función de distribución de probabilidad (FDP) de dichos alógrafos en una muestra de escritura dada.

Como un paso previo al reconocimiento, es necesario contar con un catálogo común de alógrafos que servirá de base a la hora de obtener la FDP de cada escritor. De esta manera, dicho catálogo proporcionará un espacio común de alógrafos y la FDP de cada escritor capturará su preferencia en el uso de estos alógrafos. Para generar dicho catálogo se utilizan técnicas de agrupamiento (*clustering*).

El sistema de identificación de escritor propuesto está formado por tres fases principales:

1) Preprocesado de las muestras de escritura

2) Generación del catálogo de alógrafos (*codebook)*

3) Cálculo de la FDP de cada escritor

En la Figura 4.7 se muestra el modelo de sistema de identificación propuesto.

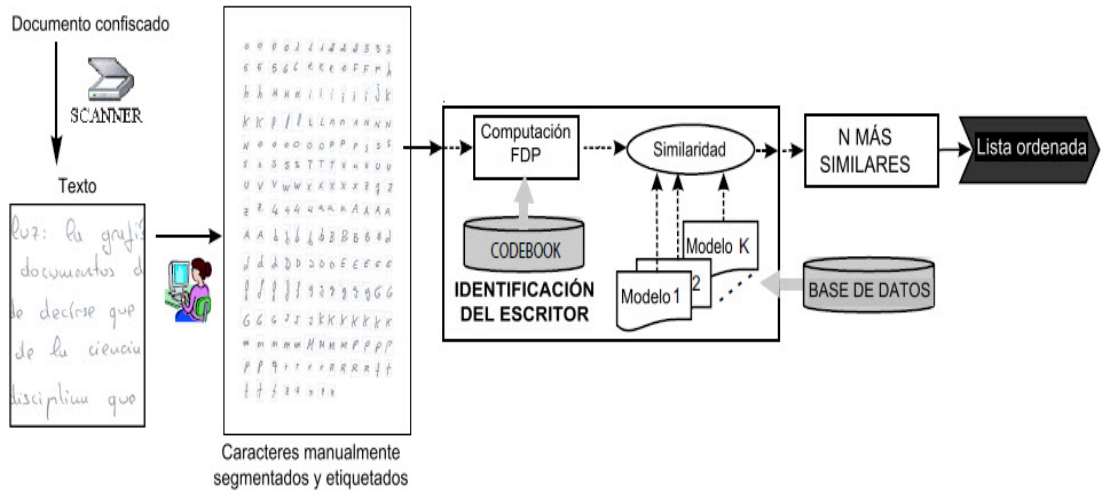


Figura 4.7. Modelo del sistema de identificación forense de escritor basado en características alográficas

4.2.1 Preprocesado

La segmentación automática perfecta de caracteres aún es un problema sin resolver [7], por lo que el sistema propuesto se basa en una segmentación manual previa por parte de un experto forense, que a partir de una muestra de escritura segmenta cada uno de los caracteres contenidos en dicha muestra y les asigna una de las 62 clases alfanuméricas: letras minúsculas ("a"-"z"), letras mayúsculas ("A"-"Z") y dígitos ("0"-"9"). Para ello se hace uso de una herramienta de software disponible en el Laboratorio de Grafística de la Dirección General de la Guardia Civil (DGGC) mediante la cual el experto forense realiza la selección del carácter utilizando el ratón del ordenador y lo etiqueta asignándolo a una de las 62 clases descritas.

En la Figura 4.8 podemos observar un ejemplo de la herramienta de software utilizada para el segmentado y etiquetado manual de las muestras. En la parte izquierda de dicha figura se muestran las 62 clases alfanuméricas utilizadas, y en la parte derecha de la misma diversas muestras de selecciones manuales de caracteres individuales con la herramienta de software. Por otra parte, en la Figura 4.9 podemos observar el conjunto de imágenes individuales de caracteres que resultan para un individuo tras aplicar la herramienta de software mencionada.

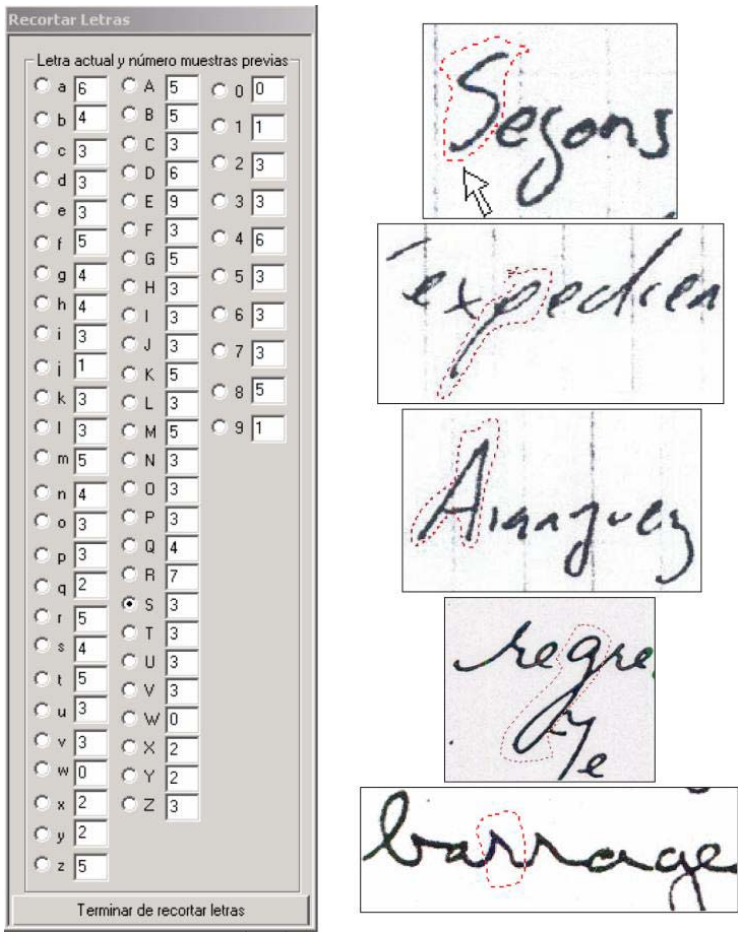


Figura 4.8. Herramienta de software utilizada para el segmentado y etiquetado manual de las muestras.

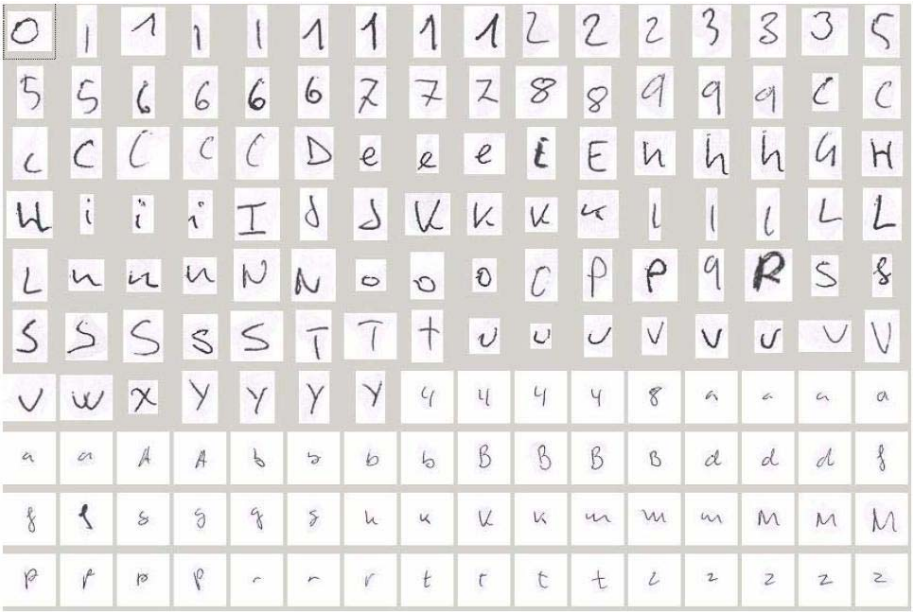


Figura 4.9. Ejemplo de caracteres manualmente segmentados de un individuo tras aplicar la herramienta de software.

En una segunda etapa del preprocesado, y una vez que contamos con imágenes individuales de cada carácter contenido en la muestra escrita del individuo, se procede a la binarización mediante el algoritmo de Otsu [38], a un recorte de los márgenes útiles (aplicando una caja limítrofe) y finalmente a una normalización de tamaño a 32x32, manteniendo la relación de aspecto.

4.2.2 Generación del catálogo de alógrafos

Como se comentó previamente, y como paso previo al reconocimiento, es necesario contar con un catálogo común de alógrafos que utilizaremos de base a la hora de obtener la función densidad de probabilidad característica de cada escritor. Para generar ese catálogo común de formas observables en una muestra de escritura se ha utilizado la base de datos CEDAR [40], que posee caracteres alfanuméricos segmentados y está formada por muestras de un conjunto independiente de escritores que no están incluidos en el material forense.

La base de datos CEDAR (disponible bajo pago en http://www.cedar.buffalo.edu/Databases) contiene imágenes digitalizadas de palabras escritas y códigos postales (300 ppp, 1 bit). Los datos fueron escaneados de sobres reales en una oficina postal de Búfalo, en Estados Unidos, por lo que no existen restricciones en cuanto a estilo, lápiz usado, etc. En este proyecto se hace uso de un conjunto de dígitos y caracteres alfanuméricos aislados. Concretamente, se han utilizado 27.837 caracteres alfanuméricos segmentados de bloques de direcciones postales y 21.837 dígitos segmentados de códigos postales. Al haber sido extraída de texto escrito en cartas postales reales, la distribución de muestras de la base de datos no es uniforme, existiendo para algunos caracteres, como “1”, más de 1.000 muestras, y menos de 10 muestras para otros caracteres, como “j”. Al igual que en el caso de las muestras de la base de datos forense utilizada para las pruebas de identificación, sobre las muestras de esta base de datos también se ha aplicado un preprocesado similar, que consta de la reducción del margen de las imágenes binarias calculando la caja limítrofe de cada una de ellas, y la posterior normalización de tamaño a 32 x 32 píxeles, preservando la relación de aspecto de cada muestra escrita.

En este proyecto se evalúan dos escenarios de cara a la generación del catálogo de alógrafos:

Escenario 1) Un catálogo global que no hace uso de la información de carácter, y que simplemente utiliza como entradas todas las imágenes de caracteres alfanuméricos de la base de datos CEDAR, generándose un catálogo global único.

Escenario 2) Un catálogo local basado en caracteres. Este escenario, por lo tanto, sí hace uso de la información de carácter, que en la base de datos forense se obtiene mediante el etiquetado por parte del experto forense a la hora de realizar la

segmentación manual. Existirán por lo tanto 62 “sub-catálogos”, uno por carácter (52 letras incluyendo minúsculas y mayúsculas, y 10 números).

Tras el preprocesado de las muestras de escritura de la base de datos CEDAR, se aplica un algoritmo de agrupamiento (*clustering*) que tiene como objetivo obtener los catálogos de alógrafos correspondientes a los escenarios descritos. Como técnica de agrupamiento se ha utilizado el algoritmo *k-means* [41], debido a su simplicidad y eficiencia computacional [42]

El algoritmo de agrupamiento *k-means* es uno de los algoritmos de agrupamiento sin supervisión más comúnmente utilizados.

Dicho algoritmo divide un conjunto de datos en $\boldsymbol{k}$ particiones o clases, de manera que la distancia entre un punto $\boldsymbol{x}_i$ de la partición $\boldsymbol{j}$, y el centroide de dicha partición $\boldsymbol{m}_j$ sea minimizada. El error cuadrático que se debe minimizar es, por lo tanto:

$$J = \sum_{j=1}^{k} \sum_{x_i \in C_j} \left\| x_i - m_j \right\|^2$$

El centroide $\boldsymbol{m}_j$ de la clase $\boldsymbol{Cj}$ es, por consiguiente, el mejor representante de las muestras que pertenecen a dicha clase.

El algoritmo comienza particionando los datos en $\boldsymbol{k}$ subconjuntos no vacíos, utilizando alguna heurística o aleatoriamente. Tras este estado inicial, se calcula el centroide de cada partición como el punto medio de la misma y se asigna cada dato a la partición cuyo centroide sea el más próximo. A partir de este punto, en cada siguiente iteración se recalculan de nuevo los centroides de cada partición y se reasignan los datos a cada una de ellas, hasta alcanzar la convergencia, que es obtenida cuando no haya más datos que cambien de partición de una iteración a otra.

Para calcular el centroide más cercano a cada dato se utiliza una función de distancia, que para datos reales suele ser la distancia euclídea.

En resumen, los 4 pasos que deben seguirse para implementar este algoritmo son:

1. Particionar los datos en k subconjuntos no vacíos

2. Calcular los centroides de cada una de las k particiones. El centroide será el punto medio de cada una de ellas.

3. Asignar cada dato a la partición cuyo centroide sea el más cercano

4. Volver al punto 2, y parar cuando no existan más reasignaciones.

El parámetro clave a la hora de generar catálogos mediante técnicas de agrupamiento es el número de centroides que deseamos obtener como salida. En este caso, se han generado catálogos de diferentes tamaños con el objetivo de evaluar el tamaño óptimo para cada escenario (es decir, aquel tamaño que obtenga mejor rendimiento). En el escenario 2, el tamaño máximo posible para cada sub-catálogo depende del número de muestras del carácter en la base de datos CEDAR. Por ejemplo, algunos caracteres como "q" o "j" permiten solamente catálogos de tamaño 2 o 3 (es decir, con 2 o 3 centroides), mientras que otros caracteres como "0" o "A" permiten catálogos de hasta 500 centroides. La Figura 4.10 muestra un ejemplo de un catálogo global de tamaño 100 centroides generado para el escenario 1, y en la Figura 4.11 se muestran algunos de los sub-catálogos generados para el escenario 2.

Figura 4.10. Ejemplo de catálogo global de tamaño 100 centroides generado para el escenario 1.

Figura 4.11. Ejemplo de sub-catálogos locales de varios tamaños generados para el escenario 2.

4.2.3 Cálculo de la FDP y comparación

Una vez preprocesadas las muestras de escritura de la base de datos forense a utilizar en las pruebas de identificación y tras haber generado los diversos catálogos de alógrafos, tal y como se ha descrito en los apartado previos, la siguiente y última etapa del sistema consiste en obtener finalmente la función de densidad de probabilidad que va a caracterizar a cada escritor.

La función densidad de probabilidad modelará la preferencia en el uso de alógrafos de cada uno de los escritores. Para poder calcularla, se construye primero un histograma en el que cada caja representa una muestra del catálogo. Para cada muestra alfanumérica de un escritor, se busca la muestra del catálogo más cercana utilizando la distancia Euclídea y se suma 1 en la caja correspondiente del histograma. De esta manera, por cada escritor obtenemos 1 histograma global (en el caso del escenario 1), o 62 histogramas (uno por carácter, en el caso del escenario 2). Por último, cada histograma se normaliza a una FDP, que será la característica discriminante que modelará la identidad de cada escritor, y que se usará en el reconocimiento.

Para calcular la similaridad entre dos FDPs de dos escritores distintos, se utiliza, al igual que en el sistema de referencia basado en gradiente, la distancia χ^2, que en [12] demostró tener mejor rendimiento para estos casos. Esta distancia, como vimos, se calcula de acuerdo a la siguiente fórmula, siendo a y b los vectores de n componentes a comparar:

$$d_{\chi^2} = \sum_{i=1}^{n} \frac{(a_i - b_i)^2}{a_i + b_i}$$

En el caso del catálogo global (escenario 1) sólo obtendremos una distancia, mientras que al utilizar los 62 sub-catálogos basados en la información de carácter (escenario 2), se obtienen 62 sub-distancias entre dos escritores dados, una por cada canal alfanumérico.

5

Experimentos en reconocimiento de escritor

5.1 Bases de datos

5.1.1 Base de datos forense

La base de datos utilizada para evaluar el sistema es una base de datos forense real formada por documentos originales confiscados o autentificados y que ha sido proporcionada por el laboratorio forense de la Dirección General de la Guardia Civil (DGGC).

Tal y como se describió en el capítulo anterior, los documentos que conforman la base de datos son preprocesados en primer lugar por parte de un experto forense, que segmenta y etiqueta en cada documento cada uno de los caracteres existentes.

En la base de datos existen 9237 muestras (caracteres) de casos forenses reales provenientes de 30 escritores diferentes, con una media de aproximadamente 300 muestras (caracteres) por escritor, que se distribuyen entre un conjunto de entrenamiento y un conjunto de test. Se puede observar en la Figura 5.1 un ejemplo de las muestras de entrenamiento de dos escritores distintos de la base de datos.

Figura 5.1. Muestras de entrenamiento de dos escritores de la base de datos forense

Es importante reseñar que para cada escritor los conjuntos de entrenamiento y test se obtienen de documentos confiscados distintos, es decir, las muestras de escritura fueron capturadas en momentos diferentes.

Al tratarse de documentos escritos reales la distribución de muestras, al igual que ocurre en la base de datos CEDAR utilizada para confeccionar los catálogos de alógrafos, no es uniforme, existiendo mayor cantidad de muestras para algunos caracteres, como 'a' ó 'r', y menor cantidad para otros como 'w' ó 'Q'. En las Figuras 5.2 y 5.3 podemos observar la distribución de muestras por escritor y por carácter de la base de datos forense utilizada.

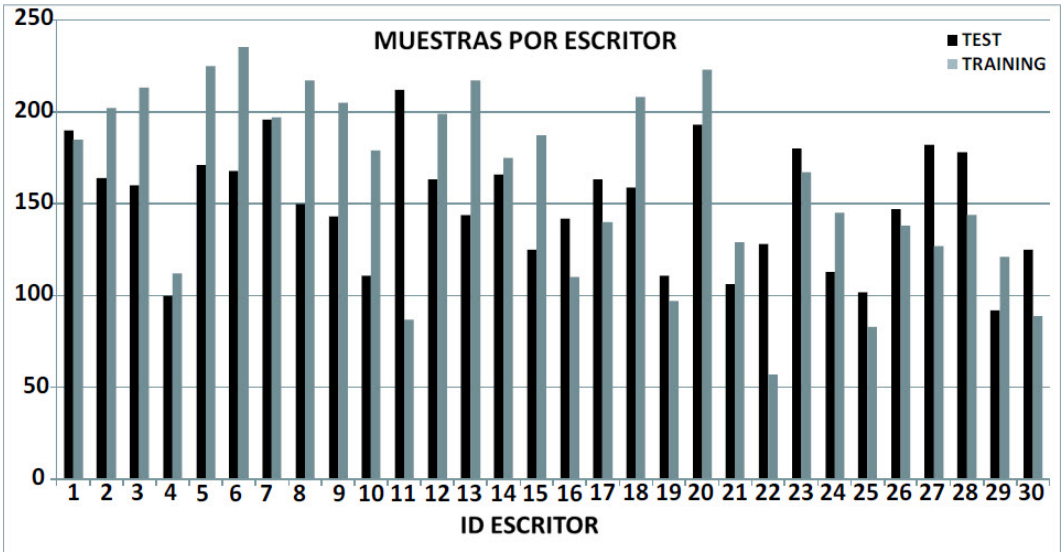


Figura 5.2. Distribución de muestras por escritor de la base de datos forense utilizada

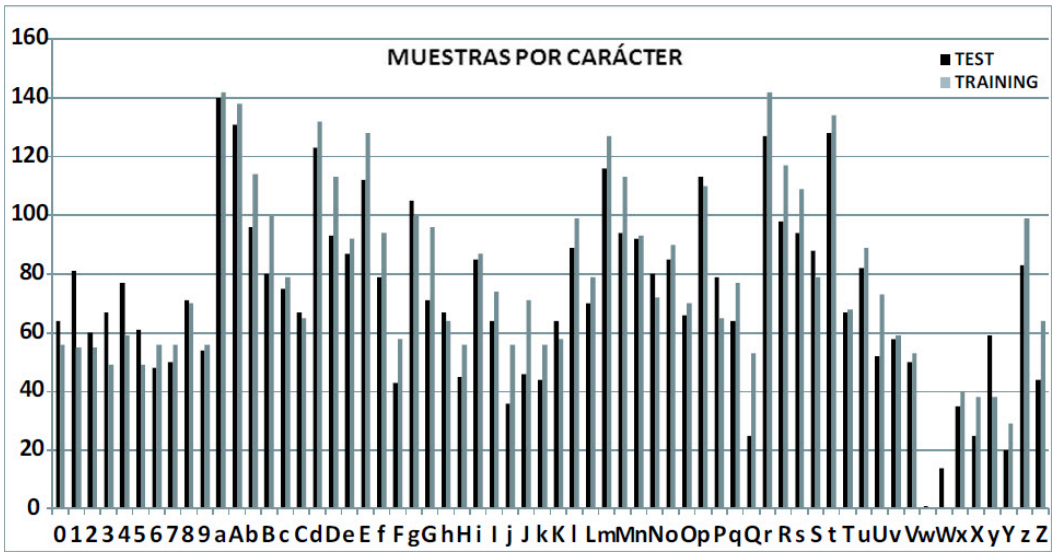


Figura 5.3. Distribución de muestras por carácter de la base de datos forense utilizada

5.1.2 Base de datos CEDAR

La base de datos utilizada para generar el catálogo de alógrafos pertenece al CEDAR (Centro de Excelencia en Reconocimiento y Análisis de Documentos), organismo asociado a la Universidad de Buffalo. [40]

Las imágenes de dicha base de datos se extrajeron de un conjunto de direcciones y códigos postales recogidos en la Oficina Postal de Buffalo, conteniendo muestras de caracteres alfanuméricos y dígitos, segmentados de forma manual a partir de las direcciones y códigos postales, respectivamente.

La base de datos original está separada en dos subconjuntos: entrenamiento y test, con el objetivo de desarrollar y evaluar sistemas de reconocimiento de escritura (en el que el objetivo es identificar lo que está escrito, no la persona que lo ha escrito). Sin embargo, en nuestro caso el objetivo era generar, contando con la mayor cantidad de datos posible, un subconjunto de prototipos orientados al reconocimiento de escritor, por lo que utilizamos todos los caracteres disponibles, sin separar en entrenamiento y test.

En la Tabla 5.1 se indica el número de muestras por carácter que contiene la base de datos y en la Figura 5.4 se muestra esta misma información de forma gráfica:

A	B	C	D	E	F	G
1399	664	639	437	547	319	162
H	I	J	K	L	M	
299	546	83	179	650	647	
N	O	P	Q	R	S	T
1022	1007	575	5	846	911	495
U	V	W	X	Y	Z	
299	228	279	131	298	31	
a	b	c	d	e	f	g
595	92	228	275	843	136	107
h	i	j	k	l	m	
228	873	3	108	753	205	
n	o	p	q	r	s	t
529	917	136	7	507	519	474
u	v	w	x	y	z	
359	147	117	183	151	15	
	0	1	2	3	4	
	913	1296	881	535	670	
	5	6	7	8	9	
	383	516	477	392	446	

Tabla 5.1. Número de muestras por carácter de la base de datos CEDAR

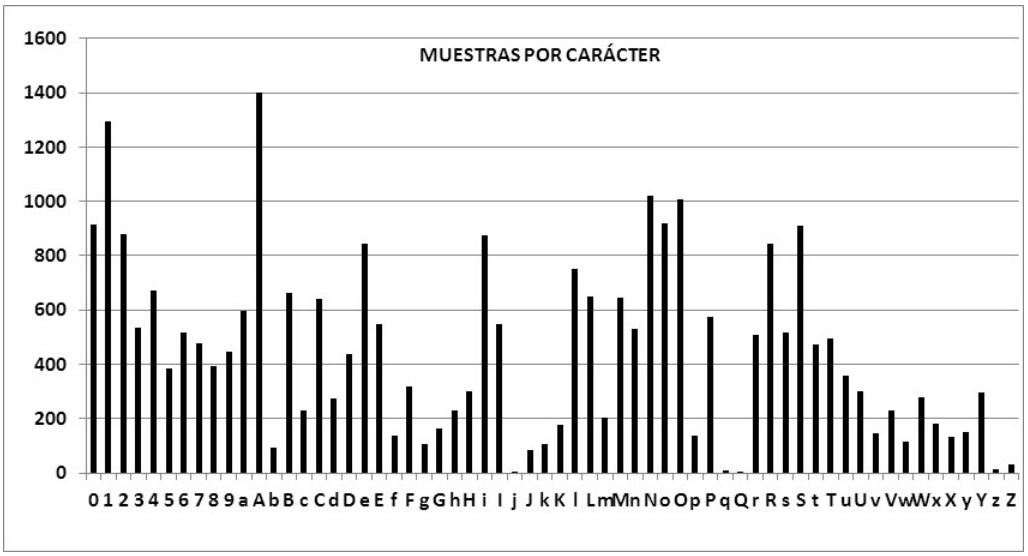


Figura 5.4. Distribución de muestras por carácter de la base de datos CEDAR

Como vemos en la tabla y la figura, al ser una base de datos extraída a partir de cartas postales, la distribución de muestras por carácter no es uniforme, existiendo más muestras de aquellos caracteres que con más frecuencia aparecieron en las cartas utilizadas en la Oficina Postal de Buffalo para la creación de la base de datos.

Las imágenes de la base de datos están escaneadas con una resolución de 300 dpi (puntos por pulgada), y representación de 1 bit (binarizadas).

En la Figura 5.5 observamos ejemplos de muestras de caracteres contenidos en la base de datos CEDAR.

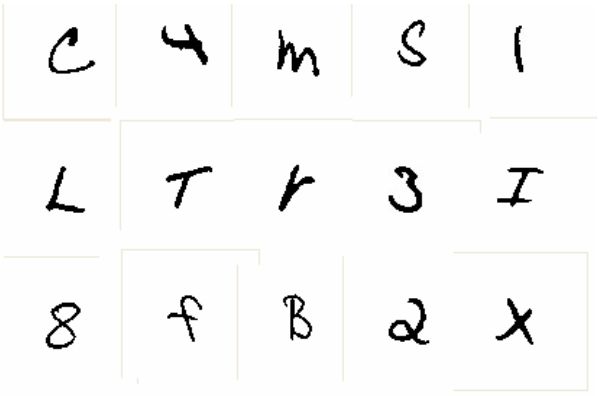


Figura 5.5. Ejemplos de muestras de caracteres contenidos en la base de datos CEDAR

5.2 Protocolo experimental

Los experimentos realizados en el presente proyecto se han realizado utilizando la base de datos forense mencionada en el apartado anterior, y en modo identificación. Empleamos este modo debido a la operativa habitual de trabajo de los sistemas forenses, en los que se suele emplear el reconocimiento negativo, es decir, tratar de identificar a un individuo que no desea ser reconocido.

En identificación biométrica de escritor, y dado un escritor del conjunto de test, el sistema devuelve una lista ordenada con las N identidades más cercanas del conjunto de entrenamiento. En un sistema ideal, la primera identidad de la lista (Top 1) debería ser la identidad correcta, pero en nuestros experimentos consideraremos también tamaños de lista mayores, en concreto desde N = 1 hasta N = 30. El objetivo final, en cualquier caso, será obtener una lista acotada de tamaño N, que permita un posterior cotejo manual por parte del experto forense, reduciendo de este modo el tiempo necesario para identificar a un individuo.

Como vemos en la Figura 5.6, trabajar con nuestro sistema biométrico de reconocimiento de escritor en modo identificación nos va a permitir obtener, dado un texto de identidad desconocida y a partir de la base de datos forense en la que hay almacenados M escritores, una lista ordenada con N identidades probables en orden descendiente de similitud con la muestra de entrada, donde N es menor o igual que M. De esta forma, conseguimos acotar el número de identidades probables, permitiendo que el cotejo manual por parte de un experto forense sea abordable. A medida que aumentamos el tamaño de la lista ordenada de salida, N, el grado de acierto del sistema será mayor, pero también aumentará el tiempo necesario de cotejo manual, por lo que la variable N supone un compromiso entre precisión y tiempo de proceso.

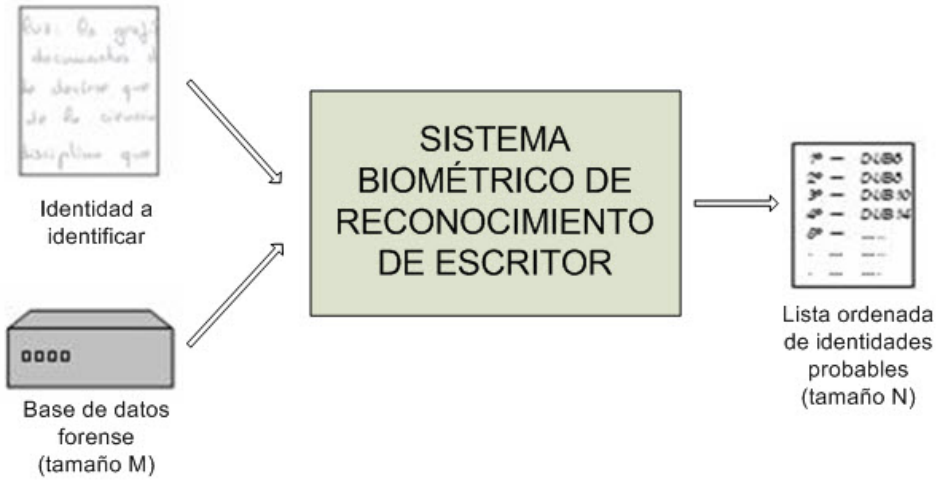


Figura 5.6. Esquema de funcionamiento del sistema de identificación de escritor

Como se indicó en el capítulo en el que se describe el sistema implementado, los experimentos se llevarán a cabo siguiendo dos escenarios distintos: en el escenario 1 se usará un único catálogo global, donde no existe diferenciación de clases/caracteres, mientras que en el escenario 2 se hará uso de un catálogo individual para cada carácter. El objetivo, por tanto, será comparar ambas pruebas para cuantificar la mejora que supone el segmentado y etiquetado manual de las muestras de escritura.

5.2.1 Escenario 1. Catálogo global

En este escenario existe un único catálogo que será obtenido, como se describió en el capítulo anterior, mediante el algoritmo *k-means*, aplicándolo sobre todos los caracteres de la base de datos CEDAR sin tener en cuenta la información de carácter.

Cada usuario estará caracterizado por una única función de densidad de probabilidad, que modelará la preferencia en el uso de los alógrafos del catálogo global por parte del usuario.

En la base de datos forense con la que se han realizado los experimentos existen 30 usuarios, por lo que en este escenario será necesario calcular 30 x 30 = 900 distancias para poder evaluar el rendimiento del sistema. En concreto, para cada usuario se obtendrá la distancia entre la función de densidad de probabilidad de su conjunto de entrenamiento y la función de densidad de probabilidad del conjunto de test de los 30 usuarios. Estas distancias obtenidas resultarán en una lista ordenada de identidades probables para cada uno de los 30 usuarios.

Adicionalmente, se ha evaluado el impacto que tiene el tamaño del catálogo de alógrafos en el rendimiento del sistema. Así, se han realizado pruebas en este escenario para tamaños del catálogo global de alógrafos comprendidos entre 1 y 1000.

5.2.2 Escenario 2. Sub-catálogos por carácter

En este caso sí hacemos uso de la información de carácter de cada muestra. Contamos con 62 catálogos, uno para cada carácter, por lo que cada usuario vendrá caracterizado por 62 funciones de densidad de probabilidad.

Por cada usuario se realizará una identificación parcial para cada uno de los 62 canales (esto es, caracteres) en que se separan sus muestras, obteniendo una identidad probable de la base de datos para cada uno de ellos. Posteriormente, se aplicará una fusión a nivel de decisión utilizando la regla de la mayoría, de manera que el primer candidato de la lista de identidades probables que devuelva el sistema será aquél que haya resultado ganador en el mayor número de canales, la segunda identidad será el siguiente escritor con mayor número de canales ganadores, etc. Además, existen las siguientes 4 reglas, que se aplicarán en orden descendiente de importancia, para evitar

empates en el caso de que dos o más usuarios posean el mismo número de canales ganadores:

1. Media de las sub-distancias ganadoras

2. Sub-distancia ganadora mínima

3. Media de las 62 sub-distancias entre los escritores de entrenamiento y test.

4. Mínima de las 62 sub-distancias entre los escritores de entrenamiento y test.

Para ilustrar con un ejemplo simplificado las 5 reglas a utilizar en la fusión a nivel de decisión, supongamos que utilizamos el sistema biométrico con 4 usuarios en la base de datos (en lugar de los 30 del proyecto) y 4 canales distintos (en lugar de 62): a, b, c y d, siendo las distancias obtenidas para cada usuario y canal las que se muestran en la Tabla 5.2.

Canal	Usuario	Distancia
a	1	0,19
	2	0,28
	3	0,6
	4	0,75
b	1	0,23
	2	0,14
	3	0,8
	4	0,43
c	1	0,12
	2	0,17
	3	0,65
	4	0,32
d	1	0,34
	2	0,11
	3	0,64
	4	0,75

Tabla 5.2. Distancias obtenidas por canal y usuario en un caso de ejemplo.

Podemos observar que el usuario 1 será el ganador para los canales a y c, mientras que el usuario 2 será el ganador para los canales b y d. Del mismo modo, ni el usuario 3 ni el usuario 4 habrán sido elegidos para alguno de los 4 canales, por lo que habrá que aplicar los criterios antes descritos para poder obtener una lista final ordenada de los 4 usuarios. En la Tabla 5.3 se muestran los valores correspondientes a los criterios a aplicar.

Criterio	Usuario			
	1	2	3	4
Canales ganadores	2	2	0	0
Media distancias ganadoras	0,155	0,125	------	------
Mínima distancia ganadora	0,12	0,11	------	------
Media todas distancias	0,22	0,175	0,6725	0,5625
Mínima todas distancias	0,12	0,11	0,6	0,32

Tabla 5.3. Criterios de desempate para cada usuario en un caso de ejemplo.

Para el desempate entre los usuarios 1 y 2, ambos con 2 canales ganadores, atenderemos en primer lugar a la media de las distancias ganadoras. Como esta media es menor para el usuario 2, ésta será la identidad que aparecerá en primer lugar en la lista de identidades probables a la salida del sistema. Respecto a los usuarios 3 y 4, deberemos utilizar como criterio la media de las distancias obtenidas en los 4 canales, siendo menor la del usuario 4, que por lo tanto, ocupará el tercer lugar en la lista ordenada. Tras estos desempates, la lista que ofrecerá el sistema a su salida será:

1º. Usuario 2
2º. Usuario 1
3º. Usuario 4
4º. Usuario 3

Las distancias a calcular en el sistema en este escenario serán 62 x 30 x 30 = 55.800 distancias, debido a que se realiza una identificación parcial para cada uno de los 62 caracteres.

Al igual que se realizaron en el escenario 1 diversos experimentos variando el tamaño del catálogo utilizado, en este caso también se analizarán dichas variaciones de tamaño para los 62 sub-catálogos. En primer lugar, se probará a utilizar el mismo tamaño de catálogo para todos los caracteres, y a partir del porcentaje de acierto individual para cada carácter, realizaremos un último experimento en el que se utilizará el catálogo de tamaño óptimo en cada uno de los canales.

El uso de la información de carácter nos permite también utilizar en el sistema un subconjunto del total de canales alfanuméricos respecto al total de 62 canales disponibles, con el fin de acelerar la velocidad de proceso. Para analizar este efecto, y tras obtener el rendimiento individual de cada canal, se han realizado pruebas de identificación en las que se emplean desde 1 a 62 canales, siempre en orden descendiente de porcentaje de acierto individual de canal, es decir, utilizando el canal más discriminativo, los 2 canales más discriminativos…, hasta los 62 canales totales.

5.3 Resultados

En esta sección mostraremos los resultados obtenidos en los diversos experimentos realizados a lo largo del presente proyecto, tanto utilizando el sistema disponible (método del gradiente), como con el nuevo sistema que hemos desarrollado (alógrafos).

En relación al sistema de alógrafos implementado, y como se ha explicado previamente, hemos realizado experimentos focalizados en dos escenarios distintos: un catálogo global vs. un catálogo para cada uno de los caracteres. Uno de los objetivos ha sido determinar cuál de estos escenarios produce una tasa de acierto mayor en el contexto forense estudiado. Por otra parte, también hemos estudiado la influencia que tiene el tamaño de dichos catálogos en el rendimiento del sistema, así como el número de canales/caracteres a emplear. Se han comparado, posteriormente, los resultados del sistema de alógrafos con los del sistema del gradiente. Por último, se ha realizado una fusión a nivel de decisión de ambos sistemas, con el objetivo de mejorar el porcentaje de acierto.

5.3.1 Resultados del sistema disponible evaluado

Como se describió en el capítulo 4, el sistema de referencia utilizado en este proyecto utiliza el cálculo del gradiente de la imagen como método de caracterización e identificación de los usuarios de la base de datos.

Sobre el sistema de referencia disponible se han probado una serie de modificaciones con el objetivo de mejorar las tasas de identificación. Estas modificaciones se basan principalmente en dos aspectos del sistema:

1. Binarización de la imagen. Se han realizado experimentos con el objetivo de evaluar si la binarización de las imágenes de la base de datos produce una mejora en el rendimiento del sistema o si, por el contrario, se obtiene mayor acierto con las imágenes en escala de gris.

2. Normalización de histogramas. Tras dividir la imagen en celdas y calcular el gradiente en cada una de ellas, se puede optar por obtener un vector de características binario (sistema original) o por normalizar el histograma de manera que lo convirtamos en una función densidad de probabilidad.

Los resultados obtenidos con este sistema se muestran en la Figura 5.7

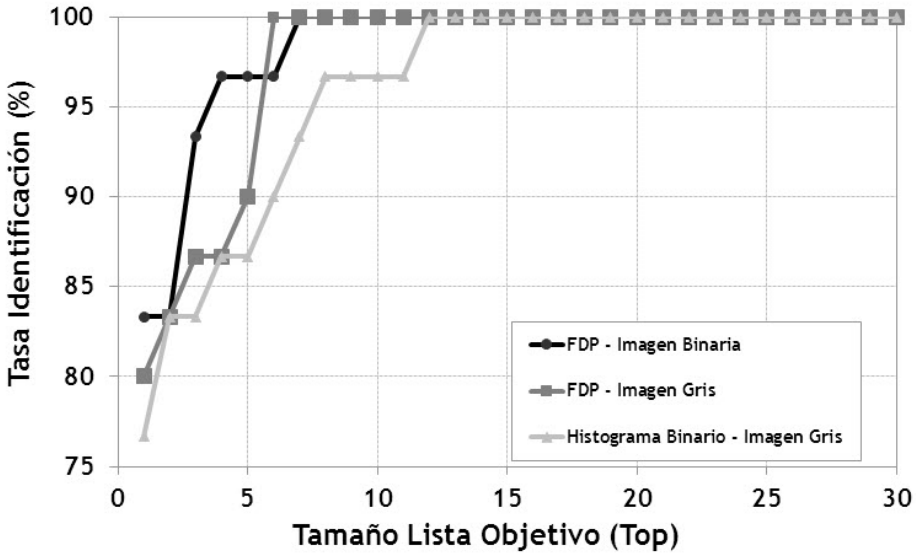


Figura 5.7. Resultados de identificación del sistema de gradiente

Como vemos en los resultados de la Figura 5.7, binarizar la imagen (gráfica negra) produce mejores resultados que utilizar las imágenes en escala de gris. Esto es debido al realce en los bordes que produce dicha binarización, haciendo que la magnitud del gradiente sea más discriminante. Por otra parte, la normalización del histograma a una función densidad de probabilidad (gráfica gris oscuro) también nos devuelve mejores resultados que utilizar histogramas binarizados. La causa de esto se debe a que la función densidad de probabilidad permite caracterizar con mayor precisión la estructura de gradiente de cada carácter.

5.3.2 Resultados del sistema desarrollado en el marco de este PFC

En el sistema de alógrafos utilizado en el proyecto el punto de partida es la obtención de un catálogo que represente los diversos alógrafos o formas características que puedan aparecer en una muestra de escritura dada. Dicho catálogo marcará los resultados que posteriormente obtengamos en los experimentos de identificación y, por lo tanto, en cada uno de los experimentos, el tamaño del catálogo de alógrafos será uno de los parámetros más importantes.

5.3.2.1 Escenario 1. Catálogo global

En el escenario 1 contamos con un único catálogo global, es decir, no existe diferenciación de caracteres en el sistema. Para este escenario, realizamos pruebas de identificación para tamaños de dicho catálogo global de entre 1 y 1000 centroides.

En la Figura 5.8 se muestran los resultados Top 1 (tamaño de lista = 1) en función del tamaño del catálogo global.

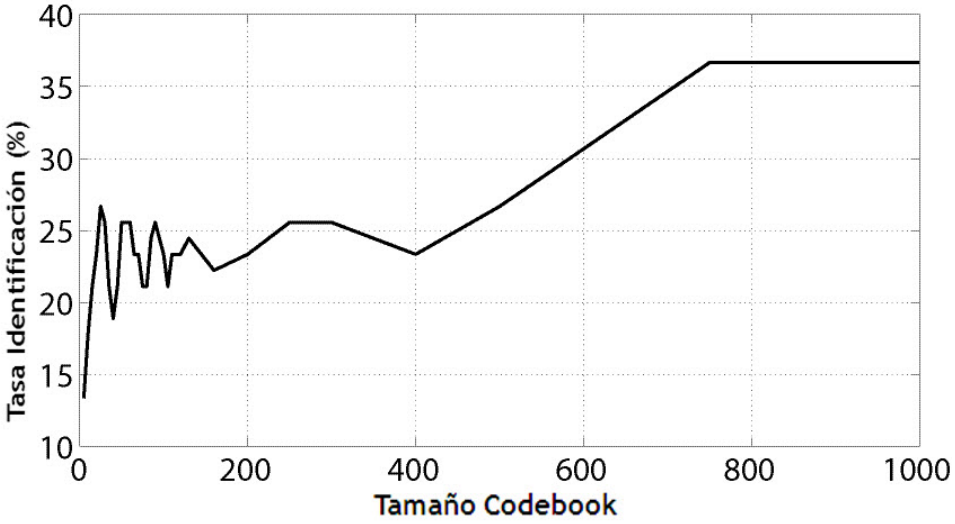


Figura 5.8. Resultados de identificación del sistema de alógrafos en el escenario 1, en función del tamaño del catálogo global.

Como podemos observar, la tasa de identificación utilizando un catálogo global oscila para tamaños de catálogo pequeños (hasta 200 centroides aproximadamente), y luego tiende a incrementarse con tamaños superiores a 400 centroides, estabilizándose a partir de los 750 centroides.

Al utilizar un catálogo único para todos los caracteres es coherente que los mejores resultados se alcancen con tamaños grandes de catálogo, ya que permiten modelar de forma completa los distintos alógrafos que pueden aparecer en una muestra escrita de la base de datos. Los catálogos con pocos centroides no consiguen, por el contrario, recoger toda la variedad de alógrafos existente, obteniendo peores resultados de identificación.

En todo caso, la mejor tasa de identificación obtenida en este escenario no alcanza el 40%, lo que muestra que el uso de un único catálogo global no permite aprovechar de manera efectiva el potencial de discriminación propio de cada carácter.

5.3.2.2 Escenario 2. Sub-catálogos por carácter

En el escenario 2 utilizamos la información de carácter proporcionada por el segmentado y etiquetado manual de la base de datos. Contamos con 62 sub-catálogos de alógrafos (uno por carácter), los cuales iremos variando de tamaño con el objetivo de encontrar el tamaño óptimo de catálogo para cada carácter.

En primer lugar, realizamos un primer experimento de identificación utilizando los 62 caracteres y variando el tamaño de los catálogos de alógrafos entre 1 y 500 centroides. En este caso, el tamaño de catálogo será igual para todos los caracteres, es decir, se realizará una identificación combinando los 62 subcatálogos con tamaño 1 centroide,

después los 62 sub-catálogos con tamaño 2 centroides, y así sucesivamente. Los resultados de este experimento se muestran en la Figura 5.9, y con más detalle en las figuras 5.10 y 5.11

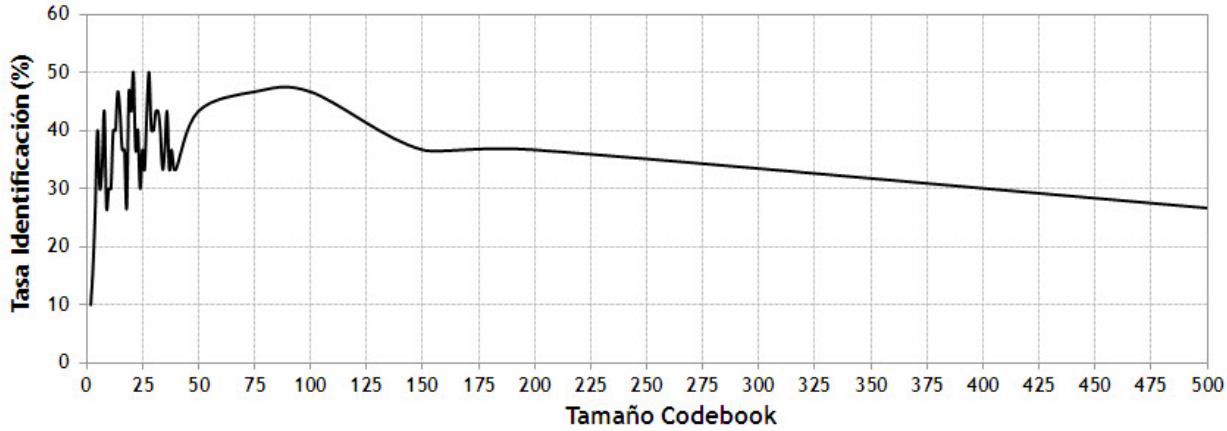


Figura 5.9. Resultados de identificación del sistema de alógrafos en el escenario 2, utilizando el mismo tamaño para todos los sub-catálogos.

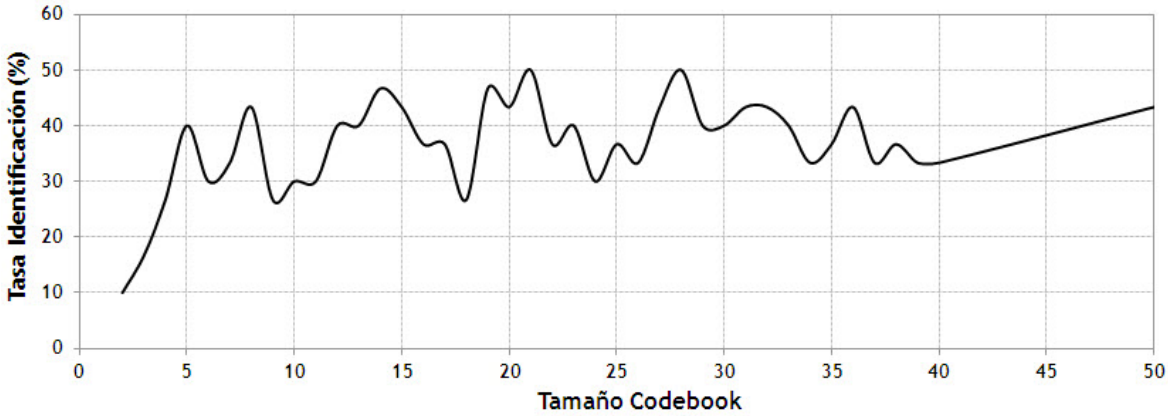


Figura 5.10. Resultados de identificación del sistema de alógrafos en el escenario 2, utilizando el mismo tamaño para todos los sub-catálogos (detalle entre 2 y 50 centroides)

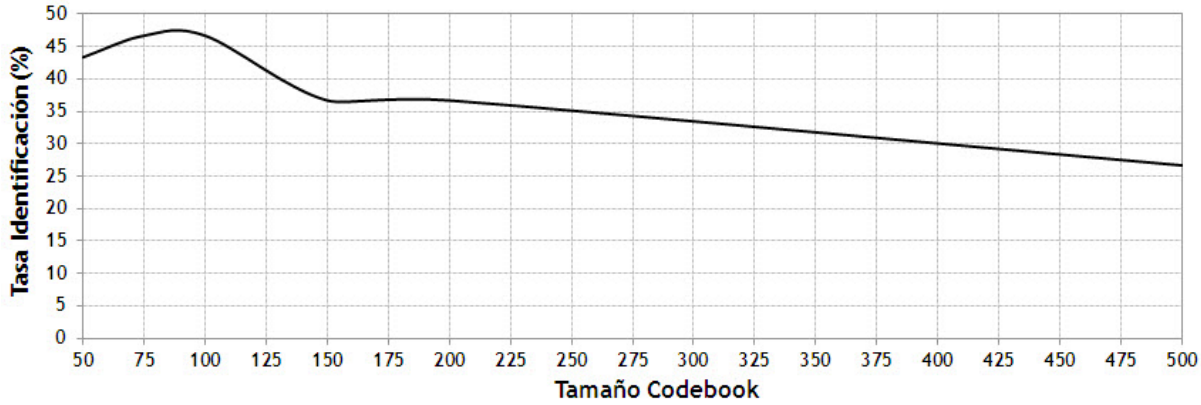


Figura 5.11. Resultados de identificación del sistema de alógrafos en el escenario 2, utilizando el mismo tamaño para todos los sub-catálogos (detalle entre 40 y 500 centroides)

En los resultados obtenidos observamos que, al igual que ocurre en el escenario 1 donde se usa un único catálogo global, de nuevo las tasas de acierto para tamaños de catálogo bajos oscilan bastante, estabilizándose a partir de tamaños de catálogo superiores a 50 centroides. La mejor tasa de acierto obtenida es del 50%.

La inestabilidad de los resultados obtenidos es debida a la utilización del mismo tamaño de catálogo para todos los caracteres, lo que conduce a la conclusión de que cada uno de los caracteres tiene un tamaño óptimo de catálogo distinto, por lo que combinar la información de los 62 caracteres no es suficiente si para cada uno de ellos no optimizamos el tamaño de su catálogo de alógrafos.

Para analizar qué tamaño de catálogo es el óptimo para cada uno de los 62 caracteres, nuestro siguiente experimento consiste en realizar las pruebas de identificación utilizando únicamente 1 carácter cada vez. De esta forma, mediremos la capacidad identificativa de cada uno de los caracteres de forma aislada, así como el tamaño de catálogo con el que obtenemos mejores resultados para cada uno de ellos. El tamaño óptimo para cada uno de los canales alfanuméricos (caracteres) será fijado como aquél para el que se obtiene una mayor tasa de identificación para un tamaño de lista (*hit list size*) de 1.

En la Figura 5.12 se muestra la tasa de identificación obtenida por cada uno de los caracteres trabajando de forma aislada, junto con el tamaño óptimo de cada sub-catálogo (aquél para el cual el carácter obtuvo la mejor tasa). Podemos ver que los caracteres con mejor tasa de acierto son “d”, “r”, “s” y “N”. Existen algunos caracteres, como “j”, “q”, “Q”, “w” y “W” con tasas de identificación nulas. Como se explicó anteriormente, para los caracteres “q”, “Q” y “j” sólo fue posible obtener catálogos de tamaño muy pequeño (hasta 2 o 3 centroides), por lo que sus FDPs no resultan para nada discriminativas. En el caso de los caracteres “w” y “W” sí se pudieron obtener catálogos de tamaño mayor, pero sin embargo en la base de datos forense utilizada en los experimentos realizados en el proyecto no existen muestras de dichos caracteres, al no ser muy frecuentemente utilizados en castellano.

Adicionalmente, también podemos observar en la Figura 5.12 que, como ya se dedujo de los resultados obtenidos en el experimento anterior (en el que usábamos el mismo tamaño de catálogo para todos los caracteres), cada uno de los canales alfanuméricos posee un tamaño óptimo de catálogo distinto. Es conveniente indicar que estos tamaños óptimos de catálogo se han obtenido para la base de datos forense con muestras escritas en castellano con la que hemos trabajado, pero cabe esperar que dependiendo del tamaño e idioma de la base de datos, el tamaño óptimo de cada sub-catálogo pueda variar.

La Tabla 5.4 recoge una clasificación ordenada de los 62 caracteres en función de su tasa individual de identificación.

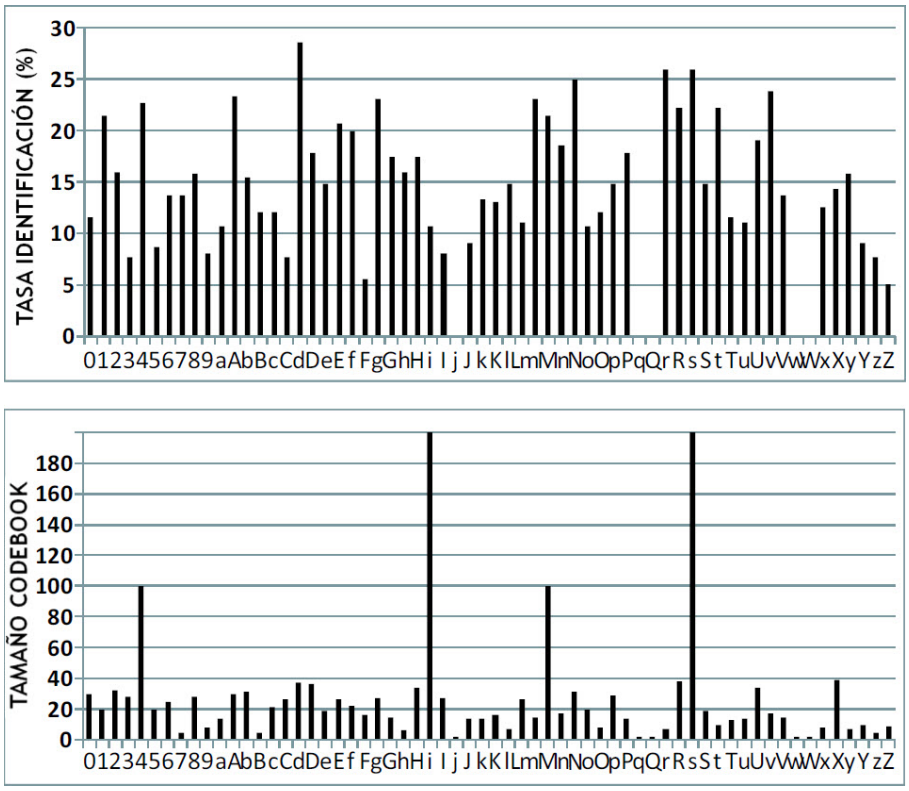


Figura 5.12. Tasas de identificación individual de cada carácter (arriba). Tamaño de catálogo con el que se obtiene la mejor tasa de identificación por carácter (abajo).

Carácter	Tasa (%)	Carácter	Tasa (%)	Carácter	Tasa (%)
d	28,6	h	16,0	u	11,1
s	25,9	2	16,0	L	11,1
r	25,9	y	15,8	o	10,7
N	25,0	8	15,8	i	10,7
v	23,8	b	15,4	a	10,7
A	23,3	S	14,8	Y	9,1
m	23,1	p	14,8	J	9,1
g	23,1	l	14,8	5	8,7
4	22,7	e	14,8	I	8,0
t	22,2	X	14,3	9	8,0
R	22,2	V	13,6	z	7,7
M	21,4	7	13,6	C	7,7
1	21,4	6	13,6	3	7,7
E	20,7	k	13,3	F	5,6
f	20,0	K	13,0	Z	5,0
U	19,0	x	12,5	Q	0,0
n	18,5	O	12,0	q	0,0
P	17,9	c	12,0	W	0,0
D	17,9	B	12,0	w	0,0
H	17,4	T	11,5	j	0,0
G	17,4	0	11,5		

Tabla 5.4. Clasificación ordenada de los caracteres en función de su tasa de identificación individual (Tasas de acierto para Top 1)

Una vez obtenido el tamaño óptimo para cada canal alfanumérico, realizamos experimentos de identificación combinando los 62 canales, esta vez con el tamaño óptimo para cada uno de ellos. Los resultados de dichos experimentos para tamaños de lista (*hit list size*) desde 1 a 30 se muestran en la Figura 5.13.

Podemos observar que en este caso las tasas de identificación del sistema mejoran notablemente respecto a los experimentos anteriores, llegando a alcanzar un 100% de efectividad para tamaños de lista superiores a 10. Esto implica que si a la salida del sistema se ofrece una lista con los 10 usuarios de la base de datos con mayor parecido respecto al escritor dubitado al que pertenece la muestra de entrada, en todos los casos se encontrará en dicha lista el escritor correcto al que pertenece dicha muestra. Para un tamaño de lista de N=1 (es decir, mostrando a la salida del sistema una sola identidad probable), la tasa de identificación supera el 75%, frente al 50% que hemos obtenido en el mejor de los casos anteriores (Figura 5.9)

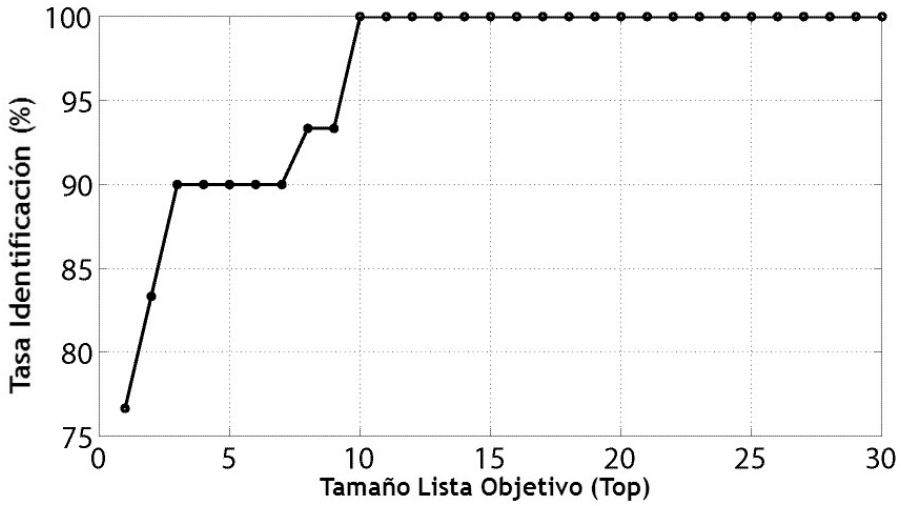


Figura 5.13. Tasas de identificación del sistema de alógrafos con tamaño de catálogo optimizado para cada carácter, para tamaños de lista de salida entre 1 y 30

Otro de los experimentos llevados a cabo consiste en combinar un subconjunto de los 62 canales, para comprobar el efecto que tiene no usar la totalidad de ellos.

En la Figura 5.14 podemos ver los resultados de los experimentos de identificación realizados en función del número de canales combinados para un tamaño de lista (*hit list size*) de N=1 (Top 1). Los canales se clasifican en orden descendente de acuerdo a su tasa individual de identificación, mostrada en la Figura 5.12, y se combinan en cada caso los canales con mejor tasa de identificación (p.ej, el canal con la mayor tasa de identificación, los dos canales con la mayor tasa de identificación, etc.).

Se observa que la tasa de identificación aumenta con el número de canales, alcanzando un máximo para alrededor de 40 canales combinados, y manteniéndose

aproximadamente constante a partir de dicho punto en torno al 76/77% observado en la Figura 5.13

También podemos ver, a partir de los resultados del experimento de identificación, que utilizando únicamente 15 de los 62 caracteres obtenemos una tasa de identificación superior al 60%. Además, la tasa de identificación entre 50 y 62 caracteres se mantiene constante, lo que indica que los 12 caracteres con peor rendimiento individual no aportan realmente información discriminante de cada usuario, y pueden ser excluidos en el sistema biométrico desarrollado.

El uso de un menor número de caracteres permite disminuir el tiempo de proceso del sistema, así como el tiempo que conlleva la operativa manual de segmentado y etiquetado de las muestras escritas.

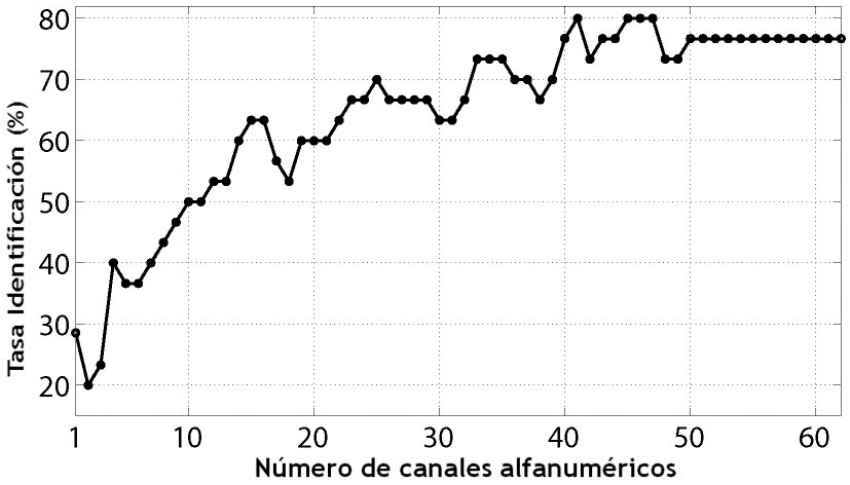


Figura 5.14. Tasas de identificación del sistema de alógrafos en función del número de caracteres utilizados para un tamaño de lista Top 1.

5.3.3 Comparativa Escenarios

En la Figura 5.15 se muestran las tasas de identificación obtenidas variando el tamaño de la lista cuando se combinan 5, 10, 20, 30, 40 y los 62 canales alfanuméricos. También se muestran los resultados para el escenario 1 (catálogo global), con un tamaño de 750 centroides (tamaño para el cual se obtenía el mejor rendimiento en dicho escenario, como se vio anteriormente).

Podemos observar en el gráfico comparativo que trabajar con sub-catálogos locales resulta en un mucho mejor rendimiento que usar un único catálogo global, lo que implica que la información de clase dada por la segmentación y etiquetado de los caracteres llevada a cabo por el experto forense proporciona una mejora considerable.

Este resultado justifica el modelo de identificación de escritor utilizado en nuestro sistema forense, en el que se invierte una considerable cantidad de tiempo cada vez que se incluye un nuevo escritor en la base de datos.

Para el sistema que trabaja con sub-catálogos locales, observamos en la Figura 5.15 que sólo existen ligeras diferencias en el rendimiento entre combinar 40 o todos los canales alfanuméricos, como ya vimos previamente en la Figura 5.14. Podemos observar, de igual modo, que si permitimos una lista de tamaño 8-10 (Top 8-10), la combinación de sólo los 10 mejores canales alfanuméricos funcionaría tan bien como otras combinaciones con mayor número de canales. Por el contrario, si queremos que la identidad correcta se encuentre en las primeras posiciones de la lista (Top 1-2), un mayor número de canales alfanuméricos son necesarios.

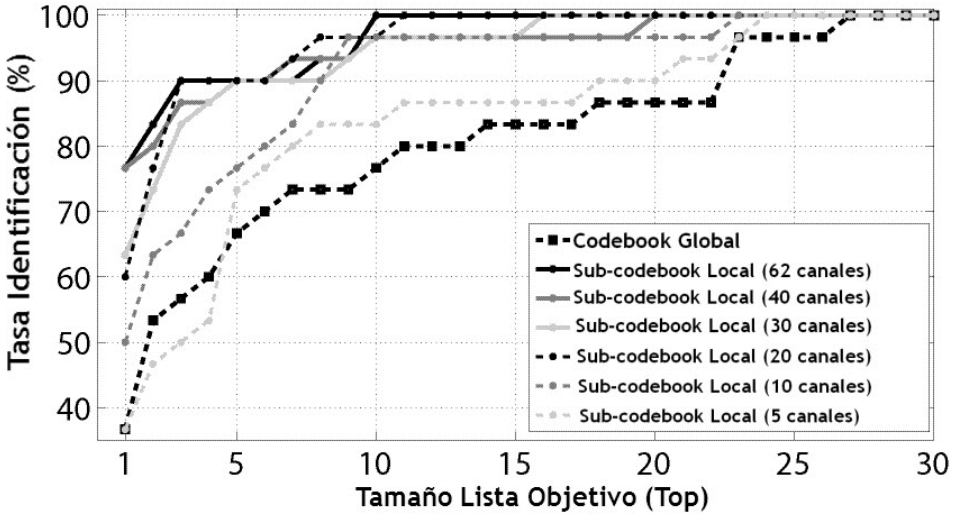


Figura 5.15. Comparativa del rendimiento del sistema de alógrafos en los diversos escenarios considerados.

5.3.4 Fusión sistemas

En la Figura 5.16 podemos observar las tasas de identificación obtenidas si fusionamos el sistema basado en gradiente con el nuevo sistema desarrollado basado en características alográficas.

La fusión se ha desarrollado a nivel de decisión, y en base al número de canales ganadores. En concreto, como se explicó en el capítulo 4 al describir el funcionamiento de los sistemas de identificación, dado un usuario dubitado a la entrada del sistema, sus caracteres se comparan con los de todos los usuarios indubitados que existen en la base de datos, y para cada uno de los 62 caracteres/canales del usuario dubitado existirá un usuario de la base de datos "ganador" de dicho canal, aplicando posteriormente una

regla de mayoría, que ordena los usuarios en la base de datos en función del número de canales ganadores, utilizando una regla de mayoría. Al realizar la fusión de los sistemas de alógrafos y gradiente, para cada usuario de la base de datos se obtiene la media de canales ganadores entre los dos sistemas, siendo éste el dato final con el que se obtiene la lista ordenada de usuarios probables a la salida del sistema

Como observamos en la Figura 5.16, las tasas de identificación que se obtienen con la fusión de sistemas son elevadas, alcanzando el 90 % para un tamaño de lista N=1 (Top 1), y llegando al 100 % con un tamaño de lista de 3 usuarios.

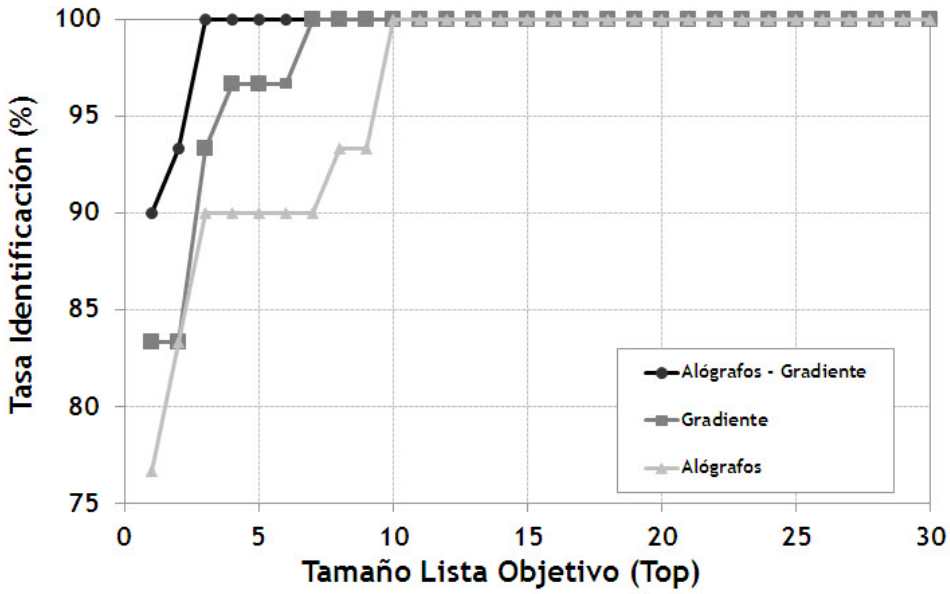


Figura 5.16. Comparativa del rendimiento entre el sistema de gradiente, el sistema de alógrafos y la fusión de ambos sistemas

Además, en la propia Figura 5.16 se compara el sistema basado en la fusión con los dos sistemas individuales en su configuración óptima,

Como ya se ha visto, el sistema basado en la fusión nos proporciona las mejores tasas de identificación (90 % de acierto para Top 1, y alcanzando un 100 % en Top 3), seguido por el sistema optimizado basado en gradiente (83 % de acierto para Top 1, y alcanzando un 100 % en Top 7) y por último el sistema basado en alógrafos (76 % de acierto para Top 1, y alcanzando un 100 % en Top 10).

A pesar de funcionar algo peor que el sistema del gradiente optimizado, el sistema de alógrafos desarrollado en este proyecto nos permite trabajar con un subconjunto de los 62 caracteres sin que se resientan en gran medida las tasas de identificación, como pudimos ver en la Figura 5.15, lo que puede ser útil cuando necesitamos disminuir el tiempo de proceso del sistema.

6

Conclusiones y trabajo futuro

6.1 Conclusiones

En el presente proyecto se ha desarrollado un sistema de reconocimiento biométrico de escritor que hace uso de características de emisión alográfica. Dicho sistema ha sido evaluado con una base de datos forense compuesta por muestras reales de 30 escritores distintos.

Haciendo uso de otra base de datos, se han generado diversos catálogos de alógrafos que se han utilizado como base para calcular las funciones de densidad de probabilidad del uso de alógrafos para cada usuario de la base de datos forense. Estas funciones de densidad de probabilidad son la característica utilizada para comparar usuarios en el sistema desarrollado, en el que se considera al escritor como un generador estocástico de alógrafos. Los experimentos han sido realizados en modo identificación (uno a muchos), por ser ésta la situación típica en casos forenses y criminales.

Se han planteado dos escenarios distintos para evaluar el rendimiento del sistema, en los que se ha comparado el uso de un único catálogo de alógrafos global (que no hace uso de la información de clase de carácter) respecto al uso de un conjunto de sub-catálogos locales (uno por carácter alfanumérico, explotando la información de clase dada por un previo etiquetado manual). Asimismo, para el segundo escenario, se ha determinado el tamaño óptimo del catálogo de alógrafos para cada carácter, logrando una mayor tasa de acierto del sistema. Los experimentos llevados a cabo han demostrado que se obtiene mucho mejor rendimiento con el uso de sub-catálogos locales, lo que justifica la considerable cantidad de tiempo utilizada por el experto forense en los procesos de segmentado y etiquetado manual. En concreto, en el caso de usar un catálogo global, la tasa de acierto para Top 1 ha sido cercana al 40 %, mientras que utilizando el conjunto de sub-catálogos locales ésta se aproxima al 80 %.

Para el segundo escenario también se ha analizado la influencia del número de canales (caracteres) utilizados, realizando experimentos con una cantidad de canales inferior a 62, llegando a la conclusión de que para alcanzar el máximo rendimiento para Top 1 basta con usar 40 canales diferentes. Otro resultado remarcable es que si utilizamos sólo 10 canales, podemos alcanzar casi un 100 % de acierto para Top 10, lo que muestra que el poder de identificación del sistema recae especialmente en un

subconjunto pequeño de los 62 caracteres utilizados. En conclusión, si buscamos mucha precisión a la salida para listas pequeñas (Top N bajos), un mayor número de canales es necesario, pero si simplemente buscamos realizar una criba de usuarios (lista de tamaño mayor a la salida) para un posterior cotejo manual del experto forense, sería suficiente con usar unos pocos del total de canales alfanuméricos.

Por último, se ha evaluado el sistema resultante de la fusión del sistema de alógrafos desarrollado en el proyecto y un sistema basado en gradiente disponible en el grupo ATVS. En este caso, la tasa de acierto para un tamaño de lista Top 1 ha sido del 90 %, y se alcanza un 100 % de acierto para Top 3, lo que mejora los resultados alcanzados por ambos sistemas de manera individual.

El análisis de estos resultados con una base de datos de tamaño limitado sugiere que la aproximación propuesta puede ser utilizada de una forma efectiva para la identificación forense de escritor, demostrando de igual manera que utilizar la información de clase de cada carácter permite mejorar ostensiblemente la tasa de acierto del sistema.

6.2 Trabajo futuro

A partir del presente proyecto y de los resultados obtenidos en el mismo, se abren diversas líneas de posible trabajo futuro, entre las que destacamos las siguientes:

- Realizar pruebas de identificación de escritor a partir de una base de datos forense de mayor tamaño, comprobando si se consolidan los resultados alcanzados en nuestros experimentos.

- Desarrollar un sistema de identificación de escritor basado en características alográficas a partir de textos escritos por escritores en otro idioma distinto al castellano. De esta manera, se podrían analizar las diferencias a la hora de generar los catálogos de alógrafos, así como comparar los resultados con el sistema desarrollado en este proyecto.

- Aplicar métodos de selección de características avanzados [43] para la combinación de canales alfanuméricos, así como técnicas basadas en la selección dependiente de usuario [44]

- Programar una aplicación práctica que utilice los algoritmos desarrollados en este proyecto, y que pueda ser utilizada por profesionales que requieran de un sistema de identificación de escritor en un entorno forense.

Bibliografía

[1] **K. Jain, A. Ross and S. Pankanti**. Biometrics: A Tool for Information Security. *IEEE Transactions on Information Forensics and Security Vol. 1, No.2, pp. 125- 143*. 2006.

[2] **Srihari, S.N., Cha, S.H., Arora, H., Lee, S**. Individuality of handwriting. *Journal of Forensic Sciences 47(4), 856- 872*. 2002

[3] **B. Bidyut Chaudhuri.** Digital Document Processing: Major Directions and Recent Advances. *Springer*. 2006.

[4] **Srihari, S., Huang, C., Srinivasan, H., Shah, V**. Biometric and Forensic Aspects of Digital *Document* Processing. *Digital Document Processing. Springer*. 2007

[5] **Plamondon, R., Srihari, S.** On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. on PAMI 22(1), 63-84*. 2000

[6] **Srihari, S., Leedham, G**. A survey of computer methods in forensic document examination. *Proc. IGS Conference, 278-281*. 2003

[7] **Schomaker, L**. Writer identification and verification. *Sensors, Systems and Algorithms, Advances in Biometrics. Springer Verlag*. 2008

[8] **Schomaker, L**. Advances in writer identification and verification. *Proc. ICDAR 2, 1268-1273*. 2007

[9] **Bensefia, A., Paquet, T., Heutte, L**. Information retrieval-based writer identification. *Proc. ICDAR, 946-950*. 2003

[10] **Schomaker, L., Bulacu, M.** Automatic writer identification using connected-component contours and edge-based features of upper-case western script. *IEEE Trans. on PAMI 26(6), 787-798*. 2004

[11] **Schomaker, L., Bulacu, M., Franke, K**. Automatic writer identification using fragmented connected-component contours. *Proc. IWFHR, 185-190*. 2004

[12] **Bulacu, M., Schomaker, L.** Text-independent writer identification and verification using textural and allographic features. *IEEE Trans. PAMI 29(4), 701-717*. 2007

[13] **Tapiador, M., Sigüenza, J**. Writer identification method based on forensic knowledge. Proc. ICBA, Springer LNCS-3072, 555-560. 2004.

[14] **Jain, A., Flynn, P., Ross, A.** Eds.: Handbook of Biometrics. *Springer*. 2008

[15] **Signature Competition ICFHR**. 2010 http://www.isical.ac.in/~icfhr2010/CallforParticipation4NSigComp2010.html. .

[16] **R. Fernandez-de-Sevilla, F. Alonso-Fernandez, J. Fiérrez-Aguilar, and J. Ortega-Garcia**. Forensic Writer Identification Using Allographic Features. *Proc. ICFHR, pp.308-313*. 2010

[17] **R. Fernandez-de-Sevilla, F. Alonso-Fernandez, J. Fierrez and J.Ortega-Garcia**. Identificación Forense de Escritor Usando Características de Emisión Alográfica. *V Jornadas de Reconocimiento Biométrico de Personas JRBP10*. 2010

[18] **A. K. Jain, A. Ross and S. Prabhakar**. An Introduction to Biometric Recognition. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Image-and-Video-Based Biometrics, Vol. 14, No. 1, pp. 4-20*. 2004.

[19] **Dessimoz, D. and Champod, C**. Linkages Between Biometrics and Forensic Science. *[ed.] A.K. Jain, P.J. Flynn and A. Ross. Handbook of biometrics. Springer, New York*. 2007.

[20] **Champod, C**. Forensic Applications, Overview. *Encyclopedia of Biometrics.S.Z. Li, editor. Springer*. 2009

[21] **Champod, C**. Keynote Speech. *International Conference of the European Academy of Forensic Science (EAFS)*. 2006.

[22] **A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki**. The DET Curve in Assessment of Detection Task Performance. *Proc. Eurospeech '97*, pages 1895-1898, Rhodes, Greece. 1997.

[23] **D. Maltoni, D. Maio, A.K. Jain, and S. Prabhakar**. Handbook of Fingerprint Recognition. *Springer, New York*. 2003

[24] **M. Bulacu**. Statistical Pattern Recognition for Automatic Writer Identification and Verification. *PhD Thesis, University of Groningen, The Netherlands*. 2006.

[25] **S. Impedovo, editor.** Fundamentals in Handwriting Recognition. s.l. *Springer Verlag*. 1994.

[26] **A. Schlapbach and H. Bunke**. A Writer Identification and Verification System Using HMM Based Recognizers. *Pattern Analysis and Applications, Vol 10, No. 1, pp. 33-43*. 2007.

[27] **R. Plamodon, G. Lorette.** Automatic Signature Verification and Writer Identification - The State of the Art. *s.l. : Pattern Recognition, Vol 22, No. 2, pp. 107-131.* 1989.

[28] **I.Yoshimura, M. Yoshimura.** Writer identification based on the arc pattern transform. *s.l. : 9th International Conference on Pattern Recognition, Editorial Computer Society Press*. 1988.

[29] **I. Yoshimura, M. Yoshimura.** Off-line writer verification using ordinary characters as the object. *s.l. : Pattern Recognition, vol. 24, no. 9, pp.909-915*. 1991.

[30] **Arazi, B.** Handwriting identification by means of run-length measurements. *s.l. : IEEE Trans. Syst., Manand Cybernetics, no. 7, vol. 12, pp.878-881*. 1997.

[31] **E. Zois, V. Anastossopoulus.** Morphological Waveform Coding for Writer Identification. *s.l. : Pattern Recognition, Vol. 33, No. 3, pp. 385-398*. 2000.

[32] **U.V. Marti, R. Messerli, H. Bunke.** Writer identification using text line based features. *s.l. : ICDAR*. 2001.

[33] **S.H. Lee, S.N. Cha, Srihari.** Combining macro and micro features for writer identification. *s.l. : SPIE, Document Recognition and retrieval IX, San Jose, CA, pp. 129-142.* 2002.

[34] **X. Wang, X. Ding, H. Liu.** Writer identification using directional element features and linear transform. *s.l. : ICDAR*. 2003.

[35] **H E S Said, K D Baker, T N Tan**. Personal identification based on handwriting. *Proceedings Fourteenth International Conference on Pattern Recognition Volume: 2, Publisher: Spie, Pages: 1761-1764 vol.2*. 1998

[36] **Cohen, F.S., Huang, Z., Yang, Z**. Invariant matching and identification of curves using B-splines curve representation. *IEEE Transactions on Image Processing, 1-10*. 1995

[37] **González R.C., Woods R.E.** Digital Image Processing. *Editorial Addison-Wesley*. 1992.

[38] **Otsu N**. A threshold selection method from gray-scale histogram. *IEEE Transactions System, Man and Cybernetics, vol. 9, pp. 62-66*. 1979

[39] **Tapiador, Marino.** Análisis de las Características de Identificación Biométrica de la Escritura Manuscrita y Mecanográfica. *s.l. : PhD. Thesis*. 2006.

[40] **Hull, J**. A database for handwritten text recognition research. *IEEE Trans. On PAMI 16(5), 550-554*. 1994

[41] **Duda, R., Hart, P., Stork, D**. Pattern Classification - *2nd Edition*. 2004

[42] **Bulacu, M., Schomaker, L**. A comparison of clustering methods for writer identification and verification. *Proc. ICDAR*. 2005

[43] **Galbally, J., Fierrez, J., Freire, M.R., Ortega-Garcia, J**. Feature selection based on genetic algorithms for on-line signature verification. *Proc. AutoID 198-203*. 2007

[44] **Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J**. Adapted user-dependent multimodal biometric authentication exploiting general information. *Pattern Recognition Letters 26, 2628-2639*. 2005

[45] **Francisco Vargas, Miguel A. Ferrer, Carlos M. Travieso, Jesús B.Alonso.** Off-line Handwritten Signature GPDS-960 Corpus. *s.l. : IAPR 9th International Conference on Document Analysis and Recognition, ISBN: 978-0-7695-2822.9, pp.764-768*. 2007.

[46] **A. Gilperez, F. Alonso-Fernandez, S. Pecharroman, J. Fierrez, J. Ortega-Garcia**. Off-line signature verification using contour features. *s.l. : ICFHR*. 2008.

[47] **A. Gilperez**. Reconocimiento off-line de escritura basado en fusión de características locales y globales. *MSc. Thesis*. 2010.

[48] **J. Fierrez, J. Galbally, J. Ortega-Garcia, M. R. Freire, F. Alonso-Fernandez, D. Ramos, D. T. Toledano, J. Gonzalez-Rodriguez, J. A. Siguenza, J. Garrido-Salas, E. Anguiano, G. Gonzalez-de-Rivera, R. Ribalda, M. Faundez-Zanuy, J. A. Ortega et. al**. BiosecurID: A Multimodal Biometric Database. *s.l. : Pattern Analysis and Applications, Vol.13, n.2, pp. 235-246*. 2010.

[49] **Brummer, N**. "Focal toolkit," *Available in http://www.dsp.sun.ac.za/nbrummer/focal.*

[50] **F. Alonso-Fernandez, J. Fierrez, D. Ramos, J. Gonzalez-Rodriguez**. Quality-Based Conditional Processing in Multi-Biometrics: application to Sensor Interoperability. *s.l. : IEEE Tansactions on Systems, Man and Cybernetics Part A, (article in press)*. 2010.

A

Presupuesto

1) Ejecución Material

• Compra de ordenador personal (Software incluido)	2.000 €
• Alquiler de impresora láser durante 6 meses	220 €
• Material de oficina	150 €
• Total de ejecución material	2.370 €

2) Gastos generales

• sobre Ejecución Material	379 €

3) Beneficio Industrial

• sobre Ejecución Material	142 €

4) Honorarios Proyecto

• 1000 horas a 15 € / hora	15.000 €

5) Material fungible

• Gastos de impresión	150 €
• Encuadernación	200 €

6) Subtotal del presupuesto

• Subtotal Presupuesto	18.241 €

7) I.V.A. aplicable

• 18% Subtotal Presupuesto	3.283,38 €

8) Total presupuesto

• Total Presupuesto	21.524,38 €

Madrid, Abril de 2012

El Ingeniero Jefe de Proyecto

Fdo.: Rubén Fernández de Sevilla García
Ingeniero Superior de Telecomunicación

B

Pliego de condiciones

Este documento contiene las condiciones legales que guiarán la realización, en este proyecto, de un *sistema biométrico de reconocimiento de escritor*. En lo que sigue, se supondrá que el proyecto ha sido encargado por una empresa cliente a una empresa consultora con la finalidad de realizar dicho sistema. Dicha empresa ha debido desarrollar una línea de investigación con objeto de elaborar el proyecto. Esta línea de investigación, junto con el posterior desarrollo de los programas está amparada por las condiciones particulares del siguiente pliego.

Supuesto que la utilización industrial de los métodos recogidos en el presente proyecto ha sido decidida por parte de la empresa cliente o de otras, la obra a realizar se regulará por las siguientes:

Condiciones generales

1. La modalidad de contratación será el concurso. La adjudicación se hará, por tanto, a la proposición más favorable sin atender exclusivamente al valor económico, dependiendo de las mayores garantías ofrecidas. La empresa que somete el proyecto a concurso se reserva el derecho a declararlo desierto.

2. El montaje y mecanización completa de los equipos que intervengan será realizado totalmente por la empresa licitadora.

3. En la oferta, se hará constar el precio total por el que se compromete a realizar la obra y el tanto por ciento de baja que supone este precio en relación con un importe límite si este se hubiera fijado.

4. La obra se realizará bajo la dirección técnica de un Ingeniero Superior de Telecomunicación, auxiliado por el número de Ingenieros Técnicos y Programadores que se estime preciso para el desarrollo de la misma.

5. Aparte del Ingeniero Director, el contratista tendrá derecho a contratar al resto del personal, pudiendo ceder esta prerrogativa a favor del Ingeniero Director, quien no estará obligado a aceptarla.

6. El contratista tiene derecho a sacar copias a su costa de los planos, pliego de condiciones y presupuestos. El Ingeniero autor del proyecto autorizará con su firma las copias solicitadas por el contratista después de confrontarlas.

7. Se abonará al contratista la obra que realmente ejecute con sujeción al proyecto que sirvió de base para la contratación, a las modificaciones autorizadas por la superioridad o a las órdenes que con arreglo a sus facultades le hayan comunicado por escrito al Ingeniero Director de obras siempre que dicha obra se haya ajustado a los preceptos de los pliegos de condiciones, con arreglo a los cuales, se harán las modificaciones y la

valoración de las diversas unidades sin que el importe total pueda exceder de los presupuestos aprobados. Por consiguiente, el número de unidades que se consignan en el proyecto o en el presupuesto, no podrá servirle de fundamento para entablar reclamaciones de ninguna clase, salvo en los casos de rescisión.

8. Tanto en las certificaciones de obras como en la liquidación final, se abonarán los trabajos realizados por el contratista a los precios de ejecución material que figuran en el presupuesto para cada unidad de la obra.

9. Si excepcionalmente se hubiera ejecutado algún trabajo que no se ajustase a las condiciones de la contrata pero que sin embargo es admisible a juicio del Ingeniero Director de obras, se dará conocimiento a la Dirección, proponiendo a la vez la rebaja de precios que el Ingeniero estime justa y si la Dirección resolviera aceptar la obra, quedará el contratista obligado a conformarse con la rebaja acordada.

10. Cuando se juzgue necesario emplear materiales o ejecutar obras que no figuren en el presupuesto de la contrata, se evaluará su importe a los precios asignados a otras obras o materiales análogos si los hubiere y cuando no, se discutirán entre el Ingeniero Director y el contratista, sometiéndolos a la aprobación de la Dirección. Los nuevos precios convenidos por uno u otro procedimiento, se sujetarán siempre al establecido en el punto anterior.

11. Cuando el contratista, con autorización del Ingeniero Director de obras, emplee materiales de calidad más elevada o de mayores dimensiones de lo estipulado en el proyecto, o sustituya una clase de fabricación por otra que tenga asignado mayor precio o ejecute con mayores dimensiones cualquier otra parte de las obras, o en general, introduzca en ellas cualquier modificación que sea beneficiosa a juicio del Ingeniero Director de obras, no tendrá derecho sin embargo, sino a lo que le correspondería si hubiera realizado la obra con estricta sujeción a lo proyectado y contratado.

12. Las cantidades calculadas para obras accesorias, aunque figuren por partida alzada en el presupuesto final (general), no serán abonadas sino a los precios de la contrata, según las condiciones de la misma y los proyectos particulares que para ellas se formen, o en su defecto, por lo que resulte de su medición final.

13. El contratista queda obligado a abonar al Ingeniero autor del proyecto y director de obras así como a los Ingenieros Técnicos, el importe de sus respectivos honorarios facultativos por formación del proyecto, dirección técnica y administración en su caso, con arreglo a las tarifas y honorarios vigentes.

14. Concluida la ejecución de la obra, será reconocida por el Ingeniero Director que a tal efecto designe la empresa.

15. La garantía definitiva será del 4% del presupuesto y la provisional del 2%.

16. La forma de pago será por certificaciones mensuales de la obra ejecutada, de acuerdo con los precios del presupuesto, deducida la baja si la hubiera.

17. La fecha de comienzo de las obras será a partir de los 15 días naturales del replanteo oficial de las mismas y la definitiva, al año de haber ejecutado la provisional, procediéndose si no existe reclamación alguna, a la reclamación de la fianza.

18. Si el contratista al efectuar el replanteo, observase algún error en el proyecto, deberá comunicarlo en el plazo de quince días al Ingeniero Director de obras, pues transcurrido ese plazo será responsable de la exactitud del proyecto.

19. El contratista está obligado a designar una persona responsable que se entenderá con el Ingeniero Director de obras, o con el delegado que éste designe, para todo relacionado con ella. Al ser el Ingeniero Director de obras el que interpreta el proyecto, el contratista deberá consultarle cualquier duda que surja en su realización.

20. Durante la realización de la obra, se girarán visitas de inspección por personal facultativo de la empresa cliente, para hacer las comprobaciones que se crean oportunas. Es obligación del contratista, la conservación de la obra ya ejecutada hasta la recepción de la misma, por lo que el deterioro parcial o total de ella, aunque sea por agentes atmosféricos u otras causas, deberá ser reparado o reconstruido por su cuenta.

21. El contratista, deberá realizar la obra en el plazo mencionado a partir de la fecha del contrato, incurriendo en multa, por retraso de la ejecución siempre que éste no sea debido a causas de fuerza mayor. A la terminación de la obra, se hará una recepción provisional previo reconocimiento y examen por la dirección técnica, el depositario de efectos, el interventor y el jefe de servicio o un representante, estampando su conformidad el contratista.

22. Hecha la recepción provisional, se certificará al contratista el resto de la obra, reservándose la administración el importe de los gastos de conservación de la misma hasta su recepción definitiva y la fianza durante el tiempo señalado como plazo de garantía. La recepción definitiva se hará en las mismas condiciones que la provisional, extendiéndose el acta correspondiente. El Director Técnico propondrá a la Junta Económica la devolución de la fianza al contratista de acuerdo con las condiciones económicas legales establecidas.

23. Las tarifas para la determinación de honorarios, reguladas por orden de la Presidencia del Gobierno el 19 de Octubre de 1961, se aplicarán sobre el denominado en la actualidad “Presupuesto de Ejecución de Contrata” y anteriormente llamado ”Presupuesto de Ejecución Material” que hoy designa otro concepto.

Condiciones particulares

La empresa consultora, que ha desarrollado el presente proyecto, lo entregará a la empresa cliente bajo las condiciones generales ya formuladas, debiendo añadirse las siguientes condiciones particulares:

1. La propiedad intelectual de los procesos descritos y analizados en el presente trabajo, pertenece por entero a la empresa consultora representada por el Ingeniero Director del Proyecto.

2. La empresa consultora se reserva el derecho a la utilización total o parcial de los resultados de la investigación realizada para desarrollar el siguiente proyecto, bien para su publicación o bien para su uso en trabajos o proyectos posteriores, para la misma empresa cliente o para otra.

3. Cualquier tipo de reproducción aparte de las reseñadas en las condiciones generales, bien sea para uso particular de la empresa cliente, o para cualquier otra aplicación, contará con autorización expresa y por escrito del Ingeniero Director del Proyecto, que actuará en representación de la empresa consultora.

4. En la autorización se ha de hacer constar la aplicación a que se destinan sus reproducciones así como su cantidad.

5. En todas las reproducciones se indicará su procedencia, explicitando el nombre del proyecto, nombre del Ingeniero Director y de la empresa consultora.

6. Si el proyecto pasa la etapa de desarrollo, cualquier modificación que se realice sobre él, deberá ser notificada al Ingeniero Director del Proyecto y a criterio de éste, la empresa consultora decidirá aceptar o no la modificación propuesta.

7. Si la modificación se acepta, la empresa consultora se hará responsable al mismo nivel que el proyecto inicial del que resulta el añadirla.

8. Si la modificación no es aceptada, por el contrario, la empresa consultora declinará toda responsabilidad que se derive de la aplicación o influencia de la misma.

9. Si la empresa cliente decide desarrollar industrialmente uno o varios productos en los que resulte parcial o totalmente aplicable el estudio de este proyecto, deberá comunicarlo a la empresa consultora.

10. La empresa consultora no se responsabiliza de los efectos laterales que se puedan producir en el momento en que se utilice la herramienta objeto del presente proyecto para la realización de otras aplicaciones.

11. La empresa consultora tendrá prioridad respecto a otras en la elaboración de los proyectos auxiliares que fuese necesario desarrollar para dicha aplicación industrial, siempre que no haga explícita renuncia a este hecho. En este caso, deberá autorizar expresamente los proyectos presentados por otros.

12. El Ingeniero Director del presente proyecto, será el responsable de la dirección de la aplicación industrial siempre que la empresa consultora lo estime oportuno. En caso contrario, la persona designada deberá contar con la autorización del mismo, quien delegará en él las responsabilidades que ostente.

C

Competición 4NSigComp2010

El grupo ATVS participó en 2010 en la competición de verificación de firma forense *off-line 4NSigComp2010 - Forensic Signature Verification Competition* [15], que se celebró en el marco del congreso internacional ICFHR 2010 – *International Conference in Frontiers of Handwriting Recognition.*

Para dicha competición el grupo envió un sistema basado en la fusión de varios sistemas disponibles. Entre dichos sistemas, se encuentra el sistema de alógrafos desarrollado en el presente proyecto, que se adaptó para poder operar en modo verificación y con firmas en lugar de escritura.

Al tratarse de un trabajo en equipo, la descripción de esta tarea no se presenta como material principal del proyecto fin de carrera, sino como un anexo al mismo.

C.1. Base de datos

En esta competición se utilizan datos procedentes de la base de datos GPDS [45], capturada por el Grupo de Procesado Digital de Señales de la Universidad de Las Palmas de Gran Canaria, que participó como co-organizador de la competición. Para la fase de entrenamiento, se distribuyó un conjunto de firmas pertenecientes a 300 individuos, con 24 firmas genuinas por individuo (capturadas en una única sesión) y 30 imitaciones de su firma, resultando un total de 16200 imágenes de firmas. Las imitaciones fueron realizadas mostrando al imitador la imagen de la firma a imitar y permitiendo práctica previa. Cada uno de los imitadores produjo 3 imitaciones de 5 individuos diferentes de la base de datos, por lo que la fima de cada individuo ha sido imitada por 10 imitadores distintos. Los datos de test de la competición (no distribuidos a los participantes) consisten en un conjunto de firmas de 400 individuos. En la Figura C.1 podemos ver ejemplos de firmas genuinas y firmas imitadas de la base de datos de entrenamiento.

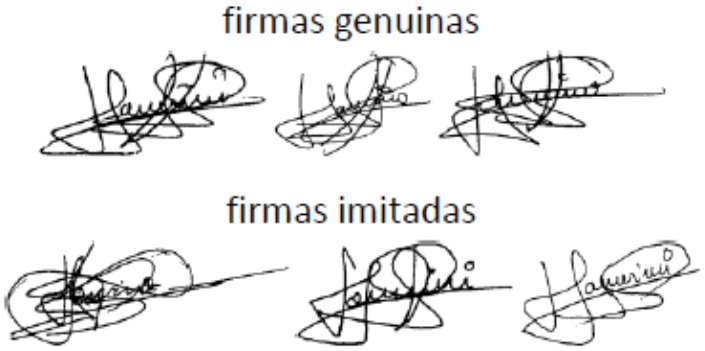


Figura C.1. Ejemplos de firmas genuinas e imitadas de la base de datos de entrenamiento

Para la generación del catálogo a utilizar en el sistema de alógrafos se ha utilizado la base de datos BIOSECUR-ID [48], capturada en el grupo ATVS. Esta base de datos contiene firmas de 133 individuos capturadas en 4 sesiones diferentes. Por cada usuario existen 4 firmas genuinas y 3 imitaciones entrenadas por sesión hechas por otros 3 imitadores. A cada persona se le solicitó que firmara en una hoja de papel sobre una tableta digitalizadora (utilizando el mismo bolígrafo todos los usuarios) y las firmas fueron escaneadas a 600 dpi.

C.2. Protocolo experimental

El protocolo experimental seguido viene marcado por el de la propia competición, en la que los sistemas se van a evaluar en modo verificación. Para modelar la identidad de cada individuo, se utilizan 4 firmas genuinas. El resto de sus firmas genuinas se utilizarán para generar intentos de acceso de usuario genuino. Por otra parte, las imitaciones se utilizan para generar intentos de acceso de impostor entrenado (*skilled forgeries*). Esto resulta en 24 – 4 = 20 *scores* de acceso genuino y 30 *scores* de acceso de impostor entrenado para cada individuo de la base de datos. Del mismo modo, para un individuo concreto, se calculan *scores* de impostor casual (*random forgeries)* utilizando una firma genuina del resto de individuos de la base de datos. Como existe un conjunto de 300 usuarios de entrenamiento, existirá un total de 299 *scores* de acceso de impostor casual por cada individuo de la base de datos. Estos impostores casuales simulan el hecho de intentar acceder al sistema con una firma aleatoria, sin tener conocimiento previo de la firma del usuario cuya identidad se desea suplantar.

En un sistema en modo verificación, en el módulo comparador se realiza una comparación 1:1, a diferencia del modo identificación en el que la comparación es 1:N. Esta comparación produce una puntuación o *score*, que comparado con un umbral que establezca el sistema emitirá una decisión de “aceptado” o “rechazado”.

Como se describió en el Capítulo 2 del proyecto, en este modo de operación existen dos tipos de errores: Falsa Aceptación (FA) y Falso Rechazo (FR). La FA mide la probabilidad de que un usuario incorrecto sea aceptado en el sistema, y el FR la probabilidad de que un usuario genuino del sistema sea rechazado. El rendimiento de los sistemas que se presentan a la competición vendrá dado en términos de un error global (OE: *Overall Error*), que se obtiene a partir de los errores de Falsa Aceptación y Falso Rechazo, de acuerdo a la siguiente formula:

$$OE = 0.5 \times \frac{nGFR}{nG} + 0.25 \times \left(\frac{nSFA}{nSF} + \frac{nRFA}{nRF} \right)$$

donde nG indica el número total de accesos de usuario genuino, nSF el número total de accesos de impostor entrenado, nRF el número total de accesos de impostor casual, nGFR el número de accesos de usuario genuino incorrectamente rechazados, nSFA el número de accesos de impostor entrenado incorrectamente aceptados y nRFA el número de accesos de impostor casual incorrectamente aceptados.

C.3. Resultados

En esta sección se muestran en primer lugar los resultados de los sistemas individuales del grupo sobre los datos de entrenamiento distribuidos por los organizadores. En segundo lugar se mostrarán los resultados de los experimentos de fusión que se llevaron a cabo para optimizar el sistema finalmente enviado. Por último, se muestran los resultados de la competición publicados por los organizadores sobre el conjunto de datos de test.

Los sistemas disponibles en el grupo ATVS que se han evaluado de cara a participar en esta competición son los siguientes:

- Sistema basado en alógrafos (desarrollado en este PFC).
- Sistema basado en características de gradiente (descrito en este PFC).
- Sistema basado en características de contorno (descrito en [46])
- Sistema basado en característica GSC estructural (descrito en [47])

Para la generación del catálogo en el sistema de alógrafos se ha hecho uso de una firma genuina de cada usuario de la base de datos BIOSECUR-ID, lo que hace un total de 133 firmas, considerándose en este caso la unidad alográfica de trabajo los bloques por los que está formada la firma, es decir, en lugar de tener un catálogo de alógrafos a partir de caracteres individuales, en este caso tendremos un catálogo de bloques de firma (trazos característicos que podemos encontrar en una firma).

Dada una imagen de firma a la que ya se ha aplicado la caja limítrofe, aplicamos una ventana deslizante que recorre la firma tanto horizontal como verticalmente, permitiendo cierto solape. En la Figura C.2 podemos ver un ejemplo de la aplicación de la ventana deslizante a una firma de la base de datos. Siguiendo este procedimiento se extraen los bloques por los que está formada la firma, descartándose aquellos que no contienen ningún trazo. Seguido este procedimiento con el total de 133 firmas utilizadas para generar el catálogo, se aplica el algoritmo *k-means* tal y como se describió en el capítulo 4, obteniéndose el catálogo de bloques de firma.

Se llevaron a cabo diversas pruebas de ajuste, modificando el tamaño de la ventana deslizante, así como el solapamiento de la misma, llegándose a la conclusión de que el mejor rendimiento se obtiene con una ventana deslizante de 16x16 con un solape del

50%. Tras aplicar esta configuración a las 133 firmas de usuario se obtienen 93735 bloques de firma disponibles para generar el catálogo mediante *k-means*. Un ejemplo de catálogo generado de bloques de firma podemos observarlo en la Figura C.3

Figura C.2. Ejemplo de aplicación de ventana deslizante a una imagen de firma. (Ventana de tamaño 32x32, con solape del 50%)

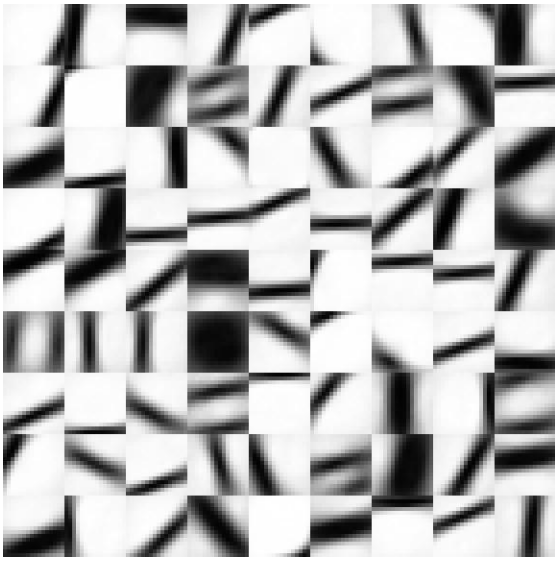


Figura C.3. Ejemplo de catálogo de bloques de firma (81 centroides)

Al igual que se realizó con el sistema de alógrafos para identificación de escritor desarrollado en este proyecto, en este caso también se realizaron pruebas obteniendo catálogos de diversos tamaños (centroides), con el objetivo de encontrar el tamaño de catálogo que obtuviese mejor rendimiento. El mejor rendimiento vendrá dado, al tratarse de un sistema en modo verificación en este caso, por aquel tamaño de catálogo que minimice la EER. (Tasa de Igual Error).

Para la base de datos de firma de entrenamiento de la competición, el tamaño óptimo de catálogo de bloques de firma fue el formado por 49 centroides.

C.3.1. Resultados de los sistemas individuales

En la Figura C.4 podemos ver el rendimiento individual en forma de curvas DET de cada uno de los sistemas sobre la base de datos de entrenamiento proporcionada por los organizadores de la competición, tanto para impostores entrenados (*skilled forgeries*) como para impostores casuales (*random forgeries*). La Tabla C.1 muestra el rendimiento en términos de EER.

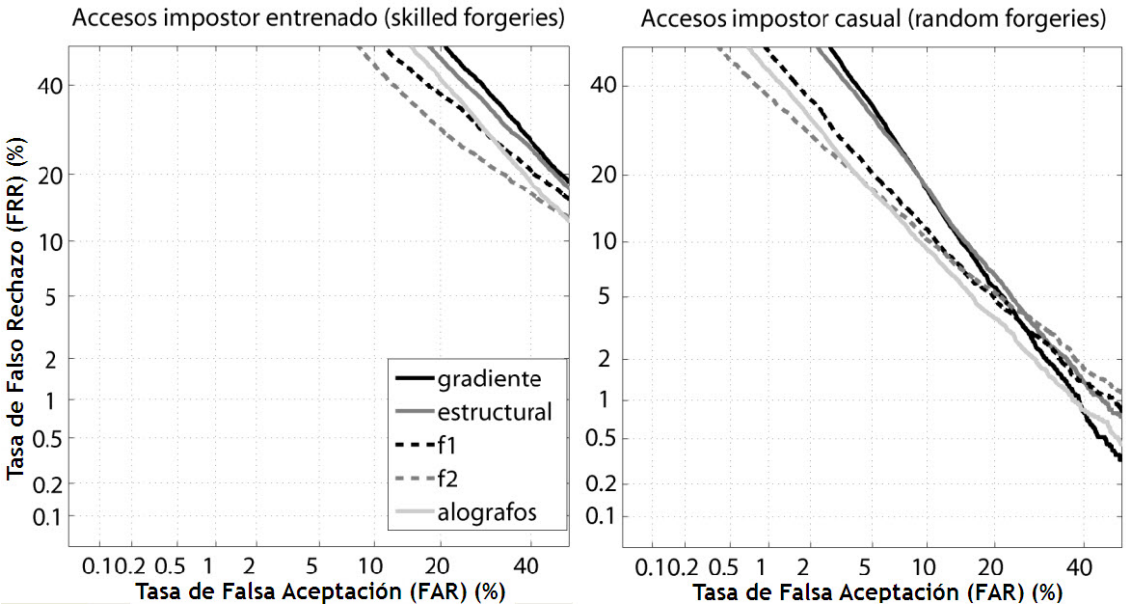


Figura C.4. Rendimiento (curvas DET) de los sistemas individuales sobre la base de datos de entrenamiento de la competición.

	EER impostor entrenado (skilled forgeries)	EER impostor casual (random forgeries)
Gradiente	33,58%	12,54%
Estructural	31,93%	12,68%
Contorno *f1*	28,75%	10,63%
Contorno *f2*	24,92%	10,14%
Alógrafos	29,01%	9,60%

Tabla C.1. Rendimiento (EER) de los sistemas individuales sobre la base de datos de entrenamiento de la competición.

Atendiendo a los resultados, es visible que el rendimiento con impostores casuales (*random forgeries*) es notablemente superior al rendimiento con impostores entrenados (*skilled forgeries)*. Este dato era previsible, pues el conocimiento y entrenamiento previo de la firma a imitar permite que la imitación sea más parecida a la original.

Por otra parte, vemos que el sistema de alógrafos es el que mejor rendimiento obtiene en el caso de impostores casuales (9,60% EER), siendo éste algo peor con impostores entrenados. (29,01% EER). En este caso, el mejor sistema es el de característica de contorno *f2*.

C.3.2. Resultados de la fusión de sistemas

Para obtener el mejor sistema posible, se han realizado pruebas de fusión de los sistemas individuales, con el objetivo de obtener la combinación óptima en términos del error global (OE), que es la medida de rendimiento utilizada por los organizadores.

Al acceder al sistema con una imagen de firma, se devolverán *N scores*¸ que corresponderán con los *N* sistemas individuales disponibles. Dados estos *N* sistemas individuales y una firma de entrada *j,* se ha efectuado una fusión lineal de los *N scores* devueltos:

$$f_j = a_0 + a_1 \cdot s_{1j} + a_2 \cdot s_{2j} + \ldots + a_N \cdot s_{Nj}$$

En base a dicha fórmula de la fusión lineal, el siguiente paso a seguir consistirá en obtener el conjunto de pesos (a_0, a_1, …, a_N) que logre optimizar la fusión de ambos sistemas.

El conjunto de pesos se entrena por regresión logística lineal de tal manera que el *score* fusionado tienda a una log-relación de verosimilitud (LLR por sus siglas en inglés

"Log-Likelihood Ratio") entre las dos posibles hipótesis de decisión, es decir, el logaritmo de la probabilidad de que los dos modelos bajo comparación pertenezcan al mismo usuario frente a la probabilidad de que los modelos bajo comparación no pertenezcan al mismo usuario. De acuerdo con la definición de LLR, idealmente los accesos de usuario genuino deberían producir un *score* fusionado positivo, indicando que hay mayor probabilidad de que los modelos bajo comparación pertenezcan al mismo usuario, y los accesos de usuario impostor deberían producir un *score* fusionado negativo, indicando que hay mayor probabilidad de que los modelos bajo comparación no pertenezcan al mismo usuario. Si el LLR tuviese valor cero, indicaría que no hay apoyo a ninguna de ambas hipótesis.

Para el entrenamiento de los pesos se han utilizado todos los *scores* obtenidos a partir de los datos de entrenamiento de la competición, a excepción de los procedentes de accesos de impostor casual, pues en este caso se obtendrán valores de LLR negativo más alejados del cero que con los impostores entrenados (como se vio en los resultados del apartado anterior), por lo que es suficiente con entrenar sólo con estos últimos. Se ha utilizado el *toolbox* Focal para el entrenamiento. Este *toolbox* es de libre disposición [49]. El procedimiento seguido para el entrenamiento de la fusión viene descrito en [50].

Se han realizado pruebas con la fusión de dos, tres, cuatro y todos los sistemas individuales, calculando en cada caso las tasas EER y OE. En la Tabla C.2 podemos ver los resultados para cada una de estas combinaciones. La mejor de ellas en términos de OE aparece recuadrada, y es ésta la enviada a la competición de verificación de firma 4NSigComp2010.

Como podemos observar, la mejor combinación ha resultado ser la formada por los sistemas de características estructurales, características de contorno *f1* y *f2* y el sistema de alógrafos adaptado del presente proyecto. En la Tabla C.2 podemos ver también que el umbral óptimo de decisión en todos los casos está próximo a cero, lo que en términos de LLR significa que está situado en torno al valor donde no existe apoyo para ninguna de las hipótesis de apoyo o rechazo al comparar los modelos.

A la vista de los resultados se puede concluir que la fusión de sistemas produce mejora sobre el mejor de los sistemas individuales. En concreto, la mejor combinación para impostores entrenados (contorno *f1,* contorno *f2*, y alógrafos) tiene un EER del 23,21%, respecto al 24,92% del mejor sistema individual (contorno *f2),* lo que supone una mejora de casi un 7%. Respecto al escenario de impostores casuales, la mejor combinación (contorno *f2* y alógrafos) produce un EER del 7,77%, y el mejor sistema individual (alógrafos) produce un EER del 9,60%, siendo la mejora en este caso de casi un 20%. Asimismo, observamos que el sistema de alógrafos desarrollado en este proyecto se encuentra en la mejor combinación en términos de EER en el escenario de impostores casuales, en la mejor combinación en términos de EER en el escenario de impostores entrenados, y en la mejor combinación en términos de OE.

DOS SISTEMAS

	EER skilled	EER random	OE	umbral optimo
grad-estr	32,50	**12,27**	23,51	0,00
grad-f1	28,78	10,63	21,71	0,10
grad-f2	**24,89**	10,17	19,18	0,10
grad-alogr	**27,71**	**8,07**	19,36	-0,10
estr-f1	29,19	**10,58**	21,71	0,10
estr-f2	**24,86**	10,14	19,16	0,10
estr-alogr	**27,42**	**8,16**	19,23	-0,10
f1-f2	25,01	10,90	19,04	0,00
f1-alogr	**26,51**	**7,84**	18,93	-0,20
f2-alogr	**24,11**	**7,77**	**17,82**	-0,20

TRES SISTEMAS

	EER skilled	EER random	OE	umbral optimo
grad-estr-f1	**28,71**	**10,55**	21,64	0,10
grad-estr-f2	**24,88**	10,15	19,16	0,10
grad-estr-alogr	**27,64**	**8,03**	19,21	-0,30
grad-f1-f2	25,57	11,08	19,28	0,00
grad-f1-alogr	**26,41**	**7,88**	18,93	-0,20
grad-f2-alogr	**23,69**	**8,03**	17,70	-0,10
estr-f1-f2	25,01	10,73	18,91	0,10
estr-f1-alogr	**26,48**	**7,82**	18,93	-0,20
estr-f2-alogr	**23,89**	**7,86**	17,84	-0,10
f1-f2-alogr	**23,21**	**8,43**	**17,43**	-0,10

CUATRO SISTEMAS

	EER skilled	EER random	OE	umbral optimo
grad-estr-f1-f2	25,50	11,15	19,31	0,00
grad-estr-f1-alogr	**26,29**	**7,82**	18,85	-0,20
grad-estr-f2-alogr	**23,79**	**8,04**	17,75	0,10
grad-f1-f2-alogr	**23,64**	**8,35**	17,38	-0,10
estr-f1-f2-alogr	**23,28**	**8,37**	**17,29**	-0,10

	EER skilled	EER random	OE	umbral optimo
TODOS	**23,63**	**8,35**	**17,31**	-0,10

Tabla C.2. Rendimiento de las diversas combinaciones de fusión de los sistemas individuales sobre la base de datos de entrenamiento de la competición

C.3.3. Resultados de la competición

Como se vio en el apartado anterior, el grupo ATVS envió a la competición 4NSigComp2010 el sistema formado por la fusión de los sistemas individuales de características estructurales, características de contorno *f1,* características de contorno *f2* y características alográficas.

Los resultados oficiales obtenidos en dicha competición, proporcionados por los organizadores de la misma son los que se muestran en la Tabla C.3. En esta tabla se muestran las tasas de Falso Rechazo (FRR, 2ª columna), Falsa Aceptación para impostores casuales (FARR, 3ª columna), Falsa Aceptación para impostores entrenados (FARS, 4ª columna), y Error General (OE, 5ª columna), que es el que determina la clasificación final de la competición (*Rank*, 6ª columna). El grupo ATVS viene representado por el Id 8. Podemos observar que los resultados obtenidos por el sistema enviado por el grupo son excelentes, quedando clasificados en 2º lugar. Pese a obtener una tasa de Falsa Aceptación de impostores casuales bastante elevada, en términos de Falso Rechazo, Falsa Aceptación de impostores entrenados y Error General, el sistema del grupo se encuentra entre los mejores. (4ª en FRR, 3ª en FARS y 2º en OE). Estos resultados son una excelente validación de los desarrollos y experimentos llevados a cabo en esta competición.

Id	*FRR (%)*	*FARR (%)*	*FARS (%)*	*OE (%)*	*Rank*
6	13,96	0,01	7,81	8,94	1st
8	18	4,31	25,38	16,42	2nd
9	21,49	1,22	22,84	16,76	3rd
1	10,43	9,18	39,18	17,31	4th
10	11,19	0,96	47,8	17,79	5th
2	22,06	0,07	46,27	22,62	6th
3	37,86	0,15	28,4	26,07	7th
4	40,83	0,09	28,36	27,53	8th
5	41,84	0,09	29,35	28,28	9th
7	46,08	0,24	29,73	30,53	10th

Tabla C.3. Resultados oficiales de la competición 4NSigComp2010. El grupo ATVS viene representado por el Id 8.

D

Publicaciones

A continuación se anexan las dos publicaciones generadas a partir de las investigaciones llevadas a cabo en este proyecto. Ambas fueron aceptadas y publicadas en sus respectivos congresos:

Título: Identificación Forense de Escritor Usando Características de Emisión Alográfica
Autores: Rubén Fernández de Sevilla García, Fernando Alonso Fernández, Julián Fiérrez Aguilar y Javier Ortega García.
Conferencia: V Jornadas de Reconocimiento Biométrico de Personas (JRBP). September 2010. Zaragoza (España)
Referencia bibliográfica: R. Fernandez-de-Sevilla, F. Alonso-Fernandez, J. Fierrez and J. Ortega-Garcia, "Identificación Forense de Escritor Usando Características de Emisión Alográfica", in V Jornadas de Reconocimiento Biométrico de Personas, JRBP10, September 2010

Título: Forensic Writer Identification Using Allographic Features
Autores: Rubén Fernández de Sevilla García, Fernando Alonso Fernández, Julián Fiérrez Aguilar y Javier Ortega García.
Conferencia: International Conference on Frontiers in Handwriting Recognition (ICFHR). November 2010. Calcuta (India).
Referencia bibliográfica: R. Fernandez-de-Sevilla, F. Alonso-Fernandez, J. Fiérrez-Aguilar, and J. Ortega-Garcia, "Forensic Writer Identification Using Allographic Features", in Proc. ICFHR, 2010, pp.308-313.

Identificación Forense de Escritor Usando Características de Emisión Alográfica

Ruben Fernandez-de-Sevilla, 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
Campus de Cantoblanco, 28049 Madrid, Spain
ruben.fernandezdesevilla, fernando.alonso, julian.fierrez, javier.ortega@uam.es

Abstract. El examen de documentos cuestionados se usa ampliamente en identificación criminal. Se presenta aquí un sistema de identificación de escritor basado en características alográficas que opera al nivel de caracteres aislados, considerando que cada persona usa un número reducido de formas para cada uno. Dichos caracteres se segmentan manualmente por un experto y se asignan a una de entre 62 clases alfanuméricas (10 números y 52 letras, incluyendo minúsculas y mayúsculas), siendo ésta la configuración particular usada por el laboratorio forense que participa en este trabajo. El sistema usa un catálogo de alógrafos generado mediante técnicas de agrupamiento (clustering) y la función de distribución de probabilidad del uso de alógrafos es la característica discriminante utilizada para el reconocimiento. Los resultados obtenidos usando una base de 30 escritores de documentos forenses reales muestran que la información a nivel de carácter proporciona una valiosa fuente de mejora, justificando la aproximación propuesta. También hemos evaluado la selección de diferentes canales alfanuméricos, mostrando una dependencia entre el tamaño de la lista objetivo ("hit list") y el número de canales necesarios para el funcionamiento óptimo.

1 Introducción

El análisis de documentos escritos con el objetivo de determinar la identidad del escritor es una importante área de aplicación en el campo forense, con numerosos casos en juicios a lo largo de los años en los que se ha utilizado la evidencia provista por estos documentos [1]. La escritura es considerada algo individual, como muestra el alto grado de aceptación social y legal de las firmas como un medio de validación de la identidad, lo que también está apoyado por estudios experimentales [2]. El objetivo del reconocimiento de escritor es determinar si dos documentos escritos, referidos como documento dubitado y documento indubitado, fueron escritos por la misma persona o no. Con este propósito, se han aplicado técnicas basadas en la visión artificial y el reconocimiento de patrones a este problema para dar soporte a los expertos forenses [3, 4].

El escenario forense presenta algunas dificultades debido a sus particulares características de [5]: reducido número de muestras escritas, variabilidad del

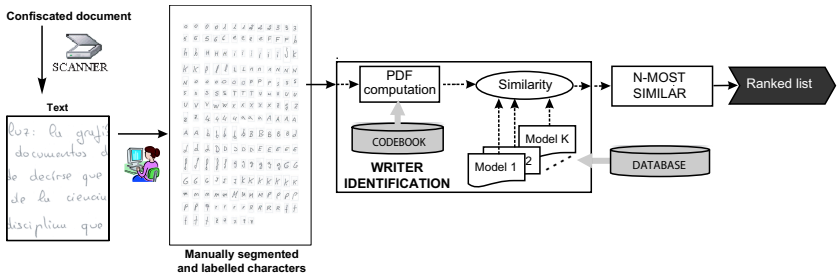


Fig. 1. Modelo del sistema de identificación forense de escritor basado en características alográficas.

estilo de escritura, lápiz o tipo de papel, presencia de patrones de ruido, etc. o no disponibilidad de información on-line (dinámica). Como consecuencia de ello, este dominio de aplicación aún se basa fuertemente en la interacción del experto humano. El uso de sistemas de reconocimiento semi-automáticos es muy útil para, dada una muestra de escritura dubitada, obtener una lista reducida de posibles candidatos que se encuentran en una base de datos de identidades conocidas, haciendo más fácil el posterior cotejo del experto forense [5, 4].

En los últimos años, se han descrito varios algoritmos de reconocimiento de escritor basados en diferentes grupos de características [6]. El presente trabajo presenta un sistema que hace uso de características del nivel alográfico, basado en discriminar escritores codificando sus alógrafos más utilizados en base a su probabilidad de ocurrencia. Trabajos previos en este sentido usan imágenes de componentes conectadas [7] o contornos [8, 9] usando segmentación automática. La segmentación automática perfecta de caracteres individuales aún es un problema sin resolver [5], pero los componentes conectados compuestos por varios caracteres o sílabas pueden segmentarse fácilmente, y los elementos generados también capturan detalles de la forma de los alógrafos utilizados por el escritor [10]. El sistema propuesto, sin embargo, usa caracteres individuales segmentados manualmente por un experto forense, a la vez que asigna cada carácter a una de las 62 clases alfanuméricas: dígitos ("0"-"9"), letras minúsculas ("a"-"z") y mayúsculas ("A"-"Z"). Ésta es la configuración usada por el grupo forense que participa en este trabajo. Para cada individuo, se escanea el documento autentificado y después se aplica una herramienta de software para la segmentación de caracteres. La segmentación se hace manualmente por un experto forense, que realiza la selección del carácter mediante el ratón del ordenador y etiqueta la muestra correspondiente de acuerdo a las 62 clases mencionadas. En este trabajo, adaptamos el algoritmo de reconocimiento basado en características alográficas de [10] para trabajar con esta configuración. Adicionalmente, el sistema se evalúa utilizando una base de datos creada a partir de documentos forenses reales (confiscados a criminales reales o autentificados en presencia de un agente de la policía), lo que es una diferencia importante en comparación con los experimen-

tos de otros trabajos, en los que las muestras de escritura eran obtenidas con la colaboración de voluntarios y bajo condiciones controladas [11].

El sistema se evalúa en modo identificación, donde cada individuo se identifica por una búsqueda entre todos los integrantes de la base de datos (búsqueda uno a muchos). Como resultado, se devuelve una clasificación ordenada de candidatos. Idealmente, la primera posición (Top 1) debería corresponder con la identidad correcta del individuo, pero se puede considerar un tamaño de lista más grande (p.ej. Top 10) para incrementar las posibilidades de encontrar la identidad correcta. La identificación es un componente crítico en aplicaciones forenses y criminales, donde el objetivo es comprobar si la persona es quien él/ella (implícita o explícitamente) niega ser [12].

El resto de este documento está organizado en varias partes. En la Sección 2 se describen las principales etapas de nuestro sistema de reconocimiento. La base de datos y el protocolo experimental utilizado se describen en la Sección 3. Los resultados experimentales se presentan en la Sección 4. Finalmente, las conclusiones se presentan en la Sección 5.

2 Descripción del sistema

El sistema de reconocimiento de escritor utilizado en este trabajo es una implementación del sistema presentado en [10], adaptado a la configuración utilizada. Se considera al escritor como un generador estocástico de formas escritas (alógrafos). La función de distribución de probabilidad (FDP) de estas formas en una muestra de escritura dada es lo que se utiliza para caracterizar al escritor. Para calcularla, se usa un catálogo común de alógrafos obtenido por medio de técnicas de agrupamiento (clustering). De esta manera, el catálogo proporciona un espacio común de alógrafos y la FDP de cada escritor captura su preferencia en el uso de estos alógrafos. Este sistema de identificación de escritor incluye tres fases principales: *i*) preprocesado, *ii*) generación del catálogo de alógrafos, y *iii*) cálculo de la FDP específica de cada escritor. En la Figura 1 se muestra el modelo de sistema de identificación utilizado en este trabajo.

Preprocesado

El método de identificación de escritor utilizado por el grupo forense participante en este trabajo se basa en la revisión manual del material escrito, como se mencionó en la Sección 1. Después de la segmentación manual y etiquetado de los caracteres alfanuméricos de un documento dado, se binarizan utilizando el algoritmo de Otsu [13], aplicando posteriormente un recorte de los márgenes útiles (caja limítrofe) y una normalización de tamaño a 32×32 píxeles, manteniendo la relación de aspecto.

Generación del catálogo de alógrafos

El objetivo de esta etapa es generar un catálogo común de formas que podemos observar en una muestra de escritura, para lo cual se utiliza una base de datos externa con caracteres alfanuméricos segmentados (obtenida a partir de

Fig. 2. Catálogos globales de diferentes tamaños.

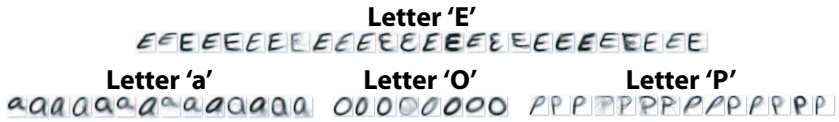


Fig. 3. Ejemplo de subcatálogos óptimos para algunos caracteres.

un conjunto independiente de escritores que no están incluidos en el material forense). Para este propósito, hacemos uso de la base de datos CEDAR [14]. Esta base de datos (disponible bajo pago en http://www.cedar.buffalo.edu/Databases) contiene imágenes digitalizadas de palabras escritas y códigos postales (300 ppp, 1 bit). Los datos fueron escaneados de sobres en una oficina postal de Búfalo, en Estados Unidos, por lo que no existen restricciones en cuanto a estilo, lápiz usado, etc. En este trabajo se hace uso de un conjunto de dígitos y caracteres alfanuméricos aislados. En concreto, se utilizaron 27.837 caracteres alfanuméricos segmentados de bloques de direcciones postales y 21.837 dígitos segmentados de códigos postales. Como la base de datos fue extraída de texto escrito en cartas postales reales, la distribución de muestras no es uniforme, exitiendo para algunos caracteres, como "1", más de 1000 muestras, y menos de 10 muestras de otros caracteres, como "j". Para los experimentos de este trabajo, reducimos el margen de las imágenes binarias calculando la caja limítrofe de cada una de ellas. Posteriormente, se procede a una normalización de tamaño a 32×32 píxeles, preservando la relación de aspecto de la muestra escrita. En este trabajo se evalúan dos escenarios para la generación del catálogo de alógrafos:

- Un catálogo global que no utiliza información de carácter. Simplemente se utilizan como entradas todas las imágenes de caracteres alfanuméricos de la base de datos CEDAR y se genera un catálogo global único.

- Un catálogo local basado en caracteres, compuesto por 62 "sub-catálogos", uno por carácter (10 números y 52 letras, incluyendo minúsculas y mayúsculas). Este caso trata de aprovechar la información de clase dada por la segmentación y etiquetado llevada a cabo por el experto forense.

Tras ello, se aplica un algoritmo de agrupamiento (clustering) a la base de datos CEDAR con el objetivo de obtener los catálogos de alógrafos correspondientes a los escenarios descritos. La técnica de agrupamiento utilizada es "k-means" [15], debido a su simplicidad y eficiencia computacional [16]. Se generan catálogos de diferentes tamaños para poder obtener el tamaño óptimo para cada escenario (es decir, aquel tamaño que consiga un mejor rendimiento). El tamaño

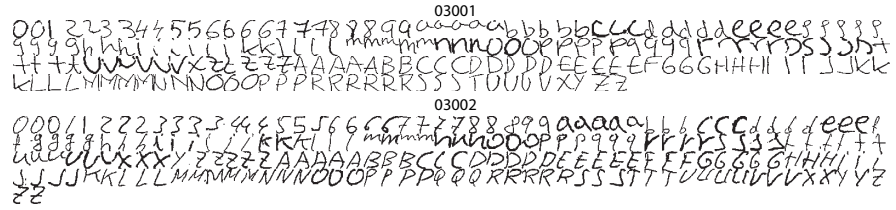


Fig. 4. Muestras de entrenamiento de dos escritores distintos de la base de datos forense.

máximo de cada subcatálogo en el escenario 2 depende del número de muestras del carácter correspondiente en la base de datos CEDAR. Por ejemplo, caracteres como "q" o "j'" permiten solamente catálogos de tamaño 2 o 3, mientras que "0" o "A" permiten tamaños de catálogo de hasta 500 centroides (clusters). La Figura 2 muestra algunos catálogos globales de diferentes tamaños de acuerdo a este protocolo, mientras que en la Figura 3 se muestran algunos de los 62 "sub-catálogos" óptimos obtenidos en los experimentos de la Sección 4.

Cálculo de la FDP y comparación.

En esta etapa, se pretende obtener la FDP discriminante de cada escritor que describa su preferencia en el uso de alógrafos. Para calcularla, se construye un histograma en el que cada caja representa a una muestra del catálogo. Para cada muestra alfanumérica de un escritor, se busca la muestra del catálogo más cercana utilizando la distancia Euclídea. Así, para cada escritor obtenemos 1 histograma (en el caso del catálogo global de alógrafos) o 62 histogramas (uno por carácter, en el caso de sub-catálogos locales). Para finalizar, cada histograma se normaliza a una FDP, que será la característica discriminante usada para reconocimiento. Para calcular la similaridad entre dos FDPs $\mathbf{o}$ y $\boldsymbol{\mu}$ de dos escritores distintos, se utiliza la distancia χ^2, la cual se calcula como: $\chi^2_{\mathbf{o}\boldsymbol{\mu}} = \sum_{i=1}^{N} \left[(o_i - \mu_i)^2 / (o_i + \mu_i) \right]$, donde N es la dimensión de los vectores $\mathbf{o}$ y $\boldsymbol{\mu}$. En el caso del catálogo global, sólo se obtiene una distancia. Cuando se utilizan los 62 sub-catálogos basados en la información de carácter, se obtienen 62 sub-distancias entre dos escritores dados, una por cada canal alfanumérico.

3 Base de datos y Protocolo.

Para evaluar el sistema se utiliza una base de datos forense real formada por documentos originales confiscados o autentificados proporcionada por el laboratorio forense de la Dirección General de la Guardia Civil (DGGC). Como se describió en la Sección 2, los caracteres alfanuméricos de las muestras escritas se segmentan y etiquetan por un experto forense de la DGGC. La base de datos contiene 9.297 muestras de caracteres de casos forenses reales provenientes de 30 escritores diferentes, con una media de unas 300 muestras por escritor, distribuidas entre

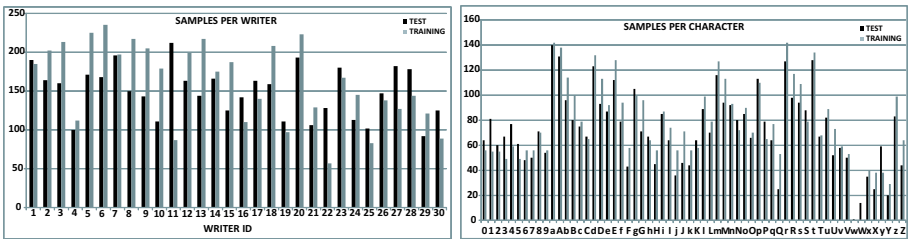


Fig. 5. Distribución de muestras por escritor (izquierda) y por carácter (derecha) de la base de datos forense utilizada en este trabajo.

un conjunto de entrenamiento y un conjunto de test. En la Figura 4 se observan las muestras de entrenamiento de dos escritores de la base de datos. Para cada escritor, los datos de entrenamiento y test se extraen de documentos confiscados diferentes, lo cual significa que se "capturaron" en distintos momentos. Al igual que la base de datos CEDAR, y dada su naturaleza, no contiene un número uniforme de muestras por caracter. La Figura 5 muestra la distribución de muestras por escritor y por carácter de la base de datos utilizada.

Dado un escritor del conjunto de test, los experimentos de *identificación* se hacen devolviendo las N identidades más cercanas del conjunto de entrenamiento. Un intento de identificación se considera exitoso si la identidad correcta se encuentra entre las N devueltas. Cuando se usa un catálogo global, solamente se calcula una distancia entre dos escritores, la cual se usa para identificación. Esto resulta en $30\times30=900$ distancias. Cuando se utilizan 62 subcatálogos, calculamos la identidad más cercana a cada carácter alfanumérico basándonos en la sub-distancia de cada canal. Se toma una decisión utilizando la regla de mayoría: la identidad de salida ganadora será aquélla que tenga el mayor número de canales alfanuméricos ganadores, la segunda identidad ganadora será el siguiente escritor con mayor número de canales ganadores, etc. Esto resulta en $62\times30\times30=55.800$ distancias calculadas. En el caso de que dos o más escritores posean el mismo número de canales ganadores, se ordenan utilizando los siguientes 4 criterios, en orden descendiente de importancia: 1) media de las sub-distancias ganadoras, 2) sub-distancia ganadora mínima, 3) media de las 62 sub-distancias entre los escritores de entrenamiento y test y 4) mínima de las 62 sub-distancias entre los escritores de entrenamiento y test.

4 Resultados

El primer paso es obtener el tamaño óptimo de los catálogos de alógrafos. En la Figura 6 se muestran los resultados de identificación en función del tamaño del catálogo global para un tamaño de lista (hit list size) de $N=1$ (Top 1). Se observa que la tasa de identificación oscila para tamaños de catálogo pequeños y tiende a incrementarse con tamaños superiores a 400 centroides, alcanzando un máximo alrededor de un tamaño de 750.

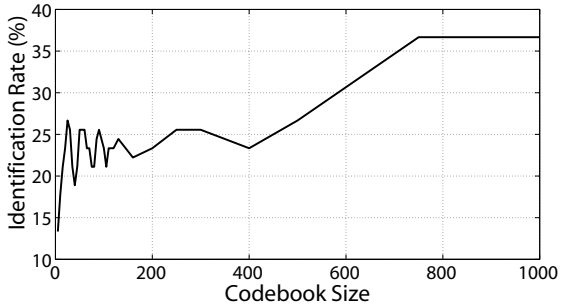


Fig. 6. Tasas de identificación de escritor en función del tamaño del catálogo (catálogo global, tamaño de lista=1).

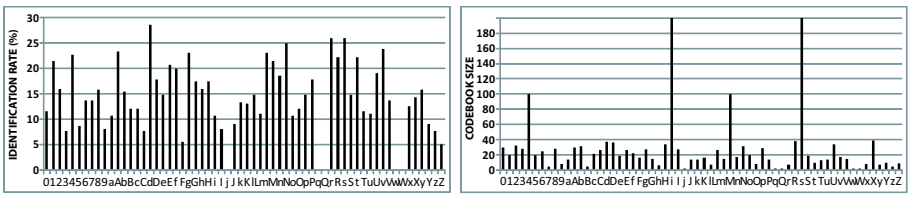


Fig. 7. Mejores tasas de identificación (izquierda) y tamaño óptimo del sub-catálogo (derecha) para cada canal alfanumérico (tamaño de lista=1).

De forma similar, variamos el tamaño de cada uno de los 62 sub-catálogos por separado en el escenario correspondiente de la Sección 2, obteniendo tasas de identificación para cada canal alfanumérico. El tamaño óptimo de cada sub-catálogo se fija como aquél para el que se obtiene la mayor tasa de identificación para un tamaño de lista (hit list size) de 1. En la Figura 7 se muestra la mejor tasa de identificación obtenida para cada canal, junto con el tamaño óptimo de cada subcatálogo. Se observa que los caracteres con las mejores tasas de acierto son “d”, “r”, “s” y “N”. Para algunos caracteres, como ”j”, ”q”, ”Q”, ”w” y ”W”, las tasas de identificación son nulas. Como se explicó en la Sección 2, para los caracteres ”q”, ”Q” y ”j” sólo se pudieron generar catálogos muy pequeños (de hasta 2 o 3 centroides) por lo que sus FDPs no son muy discriminantes. Para los caracteres “w” y “W” sí se generaron catálogos de tamaño suficiente, pero en la base de datos forense no hay muestras de dichos caracteres, al no ser frecuentemente utilizados en castellano (ver Figura 5). Podemos observar también, en la Figura 7, que para cada carácter alcanzamos la mejor tasa de identificación con un tamaño de catálogo distinto. Estos tamaños óptimos se han obtenido para nuestra base de datos real basada en muestras escritas en castellano, pero es esperable que dependiendo del tamaño y del idioma de la base de datos, el tamaño óptimo de los sub-catálogos pueda variar.

Una vez obtenido el tamaño óptimo de catálogo para cada canal, se evalúa la combinación de los 62 canales alfanuméricos. En la Figura 8 se muestran

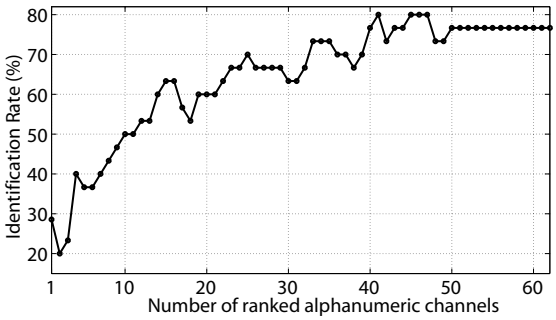


Fig. 8. Tasas de identificación de escritor en función del número de canales alfanuméricos combinados (sub-catálogos locales, tamaño de lista=1)

los resultados de los experimentos de identificación en función del número de canales combinados para un tamaño de lista (hit list size) de N=1 (Top 1). Los canales individuales son clasificados en orden descendente y seleccionados de acuerdo a su tasa de identificación, mostrada en la Figura 7 (p.ej, el canal con la mayor tasa de identificación, los dos canales con mayor tasa de identificación, etc.) Se observa que la tasa de identificación aumenta con el número de canales, alcanzando el máximo para alrededor de 40 canales combinados, manteniéndose aproximadamente constante a partir de ese punto.

También se muestran en la Figura 9 las tasas de identificación variando el tamaño de la lista cuando se combinan 5, 10, 20, 30, 40 y los 62 canales alfanuméricos. Los resultados se muestran para el catálogo global con un tamaño de 750 centroides (de acuerdo a la Figura 6). Se observa que trabajar con sub-catálogos locales resulta en un mucho mejor rendimiento que usar un único catálogo, lo que implica que la información de clase dada por la segmentación y etiquetado de caracteres llevada a cabo por el experto forense proporciona una mejora considerable. Este resultado justifica el modelo de identificación de escritor utilizado en nuestro sistema forense, en el que se invierte una considerable cantidad de tiempo cada vez que se incluye un nuevo escritor en la base de datos.

Para el sistema que trabaja con sub-catálogos locales, observamos en la Figura 9 que sólo existen ligeras diferencias en el rendimiento entre combinar 40 o todos los 62 canales alfanuméricos, como se vio previamente en la Figura 8. Podemos observar, de igual modo, que si permitimos una lista de tamaño 8-10 (Top 8-10), la combinación de sólo los 10 mejores canales alfanuméricos funciona tan bien como otras combinaciones con mayor número de canales. Por el contrario, si queremos que la identidad correcta se encuentre en las primeras posiciones de la lista (Top 1-2), se necesitan más canales alfanuméricos.

5 Conclusiones y trabajo futuro

En este trabajo, presentamos un sistema de reconocimiento de escritor que usa características de emisión alográfica. Se basa en la revisión manual de los

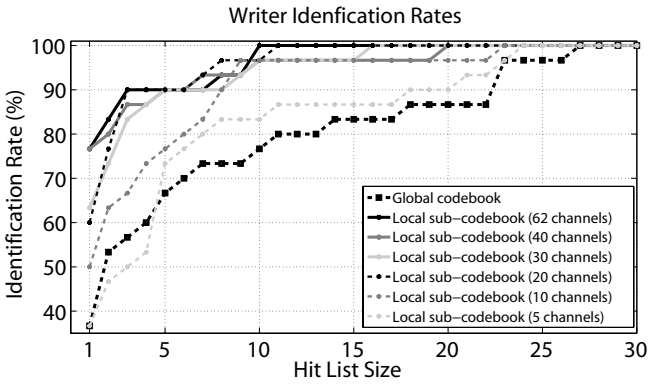


Fig. 9. Tasas de identificación de escritor en función del tamaño de la lista.

documentos escritos, realizándose, mediante una aplicación software, una segmentación y etiquetado de los caracteres de acuerdo a 62 clases alfanuméricas (10 números y 52 letras, incluyendo minúsculas y mayúsculas). Esta configuración es la usada por el grupo forense participante en este trabajo, que además ha proporcionado una base de datos de documentos forenses reales de 30 escritores distintos, lo que supone una importante diferencia respecto a otros trabajos previos en los que los datos eran obtenidos en condiciones controladas y con escritores colaborativos. Los experimentos se han realizado en modo identificación (uno a muchos), que es la situación típica en casos forenses y criminales.

El sistema presentado considera al escritor como un generador estocástico de alógrafos. Usando un catálogo común de formas escritas (alógrafos), se obtiene el conjunto personalizado de alógrafos que cada persona usa al escribir calculando su probabilidad de ocurrencia. Se han llevado a cabo experimentos usando un catálogo *global* (que no hace uso de la información de clase de carácter) y un conjunto de sub-catálogos *locales* (uno por carácter alfanumérico, explotando la información de clase dada por el etiquetado manual). Los resultados muestran que se obtiene mucho mejor rendimiento con sub-catálogos locales, justificando la considerable cantidad de tiempo utilizada por el experto forense en el proceso de segmentación y etiquetado. Para el caso local, también se ha evaluado el uso de un número diferente de canales alfanuméricos basados en su tasa de identificación individual. Observamos que la mejor tasa de identificación se obtiene cuando se usan 40 canales, sin obtener una mejora adicional al incorporar más canales. Se observa también que en el caso de listas grandes, el mejor rendimiento se obtiene ya con el uso de sólo 10 canales alfanuméricos. Sin embargo, para listas más pequeñas, se necesita un mayor número de canales alfanuméricos.

El análisis de estos resultados con una base de datos limitada sugiere que la aproximación propuesta puede ser utilizada de forma efectiva para identificación forense de escritor. Entre el trabajo futuro se incluye evaluar nuestro sistema con una base de datos forense de mayor tamaño y aplicar métodos de selección

de características avanzados [17] para la combinación de canales alfanuméricos, incluyendo aproximaciones basadas en la selección dependiente de usuario [18].

6 Agradecimientos

Este trabajo ha sido parcialmente financiado por los proyectos Bio-Challenge (TEC2009-11186), BBfor2 (FP7 ITN-2009-238803) y "Cátedra UAM-Telefónica". El trabajo postdoctoral del autor F. A.-F. ha sido financiado por un contrato del programa Juan de la Cierva del MICINN. Los autores agradecen al Laboratorio de Grafística de la Dirección General de la Guardia Civil por su inestimable apoyo.

References

1. Srihari, S., Huang, C., Srinivasan, H., Shah, V.: 17. Biometric and Forensic Aspects of Digital Document Processing. In: Digital Document Processing. Springer (2007)
2. Srihari, S.N., Cha, S.H., Arora, H., Lee, S.: Individuality of handwriting. Journal of Forensic Sciences **47**(4) (2002) 856–872
3. Plamondon, R., Srihari, S.: On-line and off-line handwriting recognition: A comprehensive survey. IEEE Trans. on PAMI **22**(1) (2000) 63–84
4. Srihari, S., Leedham, G.: A survey of computer methods in forensic document examination. Proc. IGS Conference (2003) 278–281
5. Schomaker, L.: Writer identification and verification. In: Sensors, Systems and Algorithms, Advances in Biometrics. Springer Verlag (2008)
6. Schomaker, L.: Advances in writer identification and verification. Proc. ICDAR **2** (2007) 1268–1273
7. Bensefia, A., Paquet, T., Heutte, L.: Information retrieval-based writer identification. Proc. ICDAR (2003) 946–950
8. Schomaker, L., Bulacu, M.: Automatic writer identification using connected-component contours and edge-based features of upper-case western script. IEEE Trans. on PAMI **26**(6) (2004) 787–798
9. Schomaker, L., Bulacu, M., Franke, K.: Automatic writer identification using fragmented connected-component contours. Proc. IWFHR (2004) 185–190
10. Bulacu, M., Schomaker, L.: Text-independent writer identification and verification using textural and allographic features. IEEE Trans. PAMI **29**(4) (2007) 701–717
11. Tapiador, M., Sigenza, J.: Writer identification method based on forensic knowledge. Proc. ICBA, Springer LNCS-3072 (2004) 555–560
12. Jain, A., Flynn, P., Ross, A., eds.: Handbook of Biometrics. Springer (2008)
13. Otsu, N.: A threshold selection method for gray-level histograms. IEEE Trans. on SMC **9** (1979) 62–66
14. Hull, J.: A database for handwritten text recognition research. IEEE Trans. on PAMI **16**(5) (1994) 550–554
15. Duda, R., Hart, P., Stork, D.: Pattern Classification - 2nd Edition. (2004)
16. Bulacu, M., Schomaker, L.: A comparison of clustering methods for writer identification and verification. Proc. ICDAR (2005)
17. Galbally, J., Fierrez, J., Freire, M.R., Ortega-Garcia, J.: Feature selection based on genetic algorithms for on-line signature verification. Proc. AutoID (2007) 198–203
18. Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J.: Adapted user-dependent multimodal biometric authentication exploiting general information. Pattern Recognition Letters **26** (2005) 2628–2639

Forensic Writer Identification Using Allographic Features

Ruben Fernandez-de-Sevilla, 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
Campus de Cantoblanco, 28049 Madrid, Spain
ruben.fernandezdesevilla, fernando.alonso, julian.fierrez, javier.ortega@uam.es

Abstract

Questioned document examination is extensively used by forensic specialists for criminal identification. This paper presents a writer recognition system based on allographic features operating in identification mode (one-to-many). It works at the level of isolated characters, considering that each writer uses a reduced number of shapes for each one. Individual characters of a writer are manually segmented and labeled by an expert as pertaining to one of 62 alphanumeric classes (10 numbers and 52 letters, including lowercase and uppercase letters), being the particular setup used by the forensic laboratory participating in this work. A codebook of shapes is then generated by clustering and the probability distribution function of allograph usage is the discriminative feature used for recognition. Results obtained on a database of 30 writers from real forensic documents show that the character class information given by the manual analysis provides a valuable source of improvement, justifying the proposed approach. We also evaluate the selection of different alphanumeric channels, showing a dependence between the size of the hit list and the number of channels needed for optimal performance.

1. Introduction

Analysis of handwritten documents with the aim of determining the writer is an important application area in forensic casework, with numerous cases in courts over the years that have dealt with evidence provided by these documents [14]. Handwriting is considered individual, as shown by the wide social and legal acceptance of signatures as a mean of identity validation, which is also supported by experimental studies [16]. The goal of writer recognition is to determine whether two handwritten documents, referred as to the known and the questioned document, were written by the same person or not. For this purpose, computer vision and pattern recognition techniques have been applied to this problem to support forensic experts [9, 15].

The forensic scenario present some difficulties due to their particular characteristics in terms of [11]: frequently reduced number of handwriting samples, variability of writing style, pencil or type of paper, the presence of noise patterns, etc. or the unavailability of online information. As a result, this application domain still heavily relies on human-expert interaction. The use of semi-automatic recognition systems is very useful to, given a questioned handwriting sample, narrow down a list of possible candidates which are into a database of known identities, therefore making easier the subsequent confrontation for the forensic expert [11, 15].

In the last years, several writer recognition algorithms have been described in literature based on different group of features [10]. This paper presents a writer recognition system making use of features at the allographic level, which focuses on discriminating writers encoding their preferred or more used allographic elements by capturing their occurrence probability. Previous works following this direction used connected-component images [1] or contours [12, 13] using automatic segmentation. Perfect automatic segmentation of individual characters still remains an unsolved problem [11], but connected components encompassing several characters or syllables can be easily segmented, and the elements generated also capture shape details of the allographs used by the writer [3]. The system in this paper, however, makes use of individual characters segmented manually by a forensic expert which are also assigned to one of the 62 alphanumeric classes among digits “0”∼“9”, lowercase letters “a”∼“z”, and uppercase letters “A”∼“Z”. This is the setup used by the Spanish forensic group participating in this work. For a particular individual, the authenticated document is scanned and next, a dedicated software tool for charac-

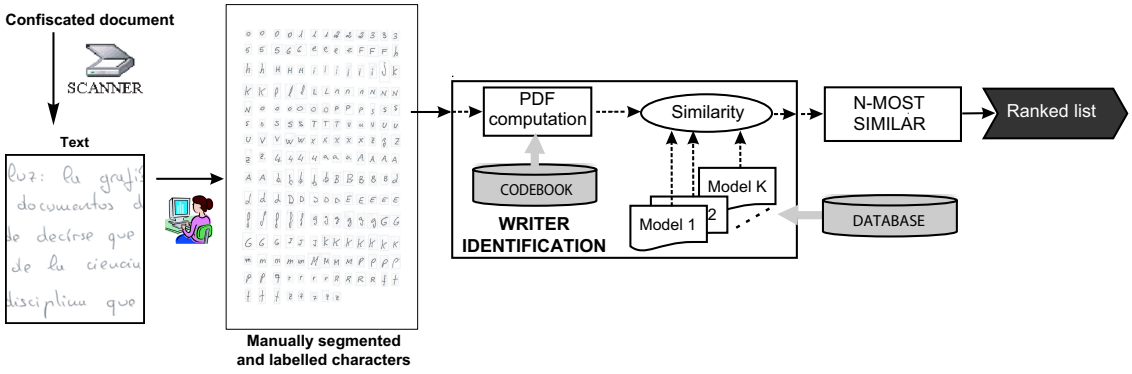


Figure 1. System model for forensic writer identification based on allographic features.

ter segmentation is used. Segmentation is done manually by a trained operator (a forensic expert), who draw a character selection with the computer mouse and label the corresponding sample according to the 62 classes mentioned. In this work, we adapt the recognition method based on allographic features from [3] to work with this setup. Additionally, the system is evaluated using a database created from real forensic documents (i.e. confiscated to real criminals or authenticated in the presence of a police officer), which is an important point compared with experiments of other works where the writing samples are obtained with the collaboration of volunteers under controlled conditions [18].

The system is evaluated in identification mode, in which an individual is recognized by searching the reference models of all the subjects in the database for a match (one-to-many). As a result, the system returns a ranked list of candidates. Ideally, the first ranked candidate (Top 1) should correspond with the correct identity of the individual, but one can choose to consider a longer list (e.g. Top 10) to increase the chances of finding the correct identity. Identification is a critical component in negative recognition applications (or watchlists) where the aim is checking if the person is who he/she (implicitly or explicitly) denies to be, which the typical situation in forensic/criminal cases [7].

The rest of the paper is structured as follows. In Section 2 we describe the main stages of our recognition system. Section 3 describes the database and the experimental protocol used. Experimental results are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. System Description

The writer recognition system used in this paper is an implementation of the system presented in [3], which is adapted to the particular setup of this paper. It considers the writer as a stochastic pattern generator of handwritten shapes. The probability distribution function (PDF) of these shapes in a given handwritten sample is used to characterize the writer, which is computed using a common codebook of shapes obtained by means of clustering techniques. This way, the codebook provides a common shape space and the PDF captures the individual shape usage preference of the writer. This writer identification system includes three main stages: i) handwriting preprocessing, ii) shape codebook generation, and iii) computation of the writer-specific PDF. In Figure 1, the overall model of the identification system used in this work is depicted.

Handwriting Preprocessing

The writer identification method used by the forensic group participating in this work is based on manually reviewing the handwritten material, as mentioned in Section 1. After manual segmentation and labeling of alphanumeric characters from a given document, they are binarized using the Otsu algorithm [8], followed by a margin drop and a size normalization to 32×32 pixels, preserving the aspect ratio.

Codebook Generation

The objective of this stage is to generate a common codebook of shapes that we can observe on a handwriting sample, for which an external database of segmented alphanumeric characters is used (obtained from an independent set of writers not "participating" in the forensic material). For this purpose, we make use of the CEDAR database [6]. This database[1] contains digitized images of handwritten words and ZIP codes (300 dpi, 8-bit) and binary handwritten isolated digits and alphanumeric characters (300 dpi, 1-bit). Data were scanned from envelopes in a working post office at Buffalo in the US, therefore no constraints were imposed in the

[1] Available for a fee at http://www.cedar.buffalo.edu/Databases

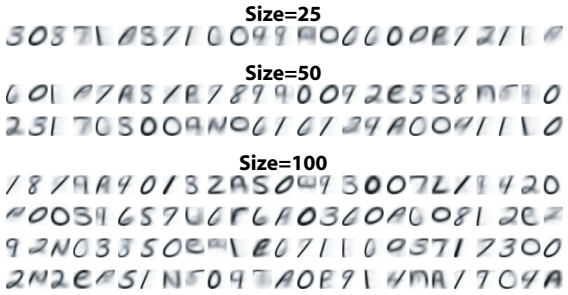


Figure 2. Global codebooks of different sizes.

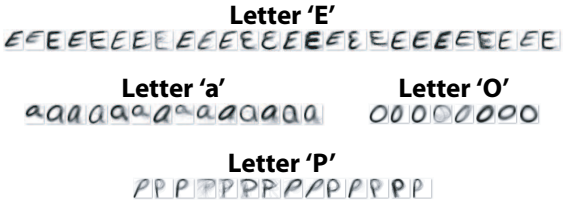


Figure 3. Example of optimal "sub-codebooks" for different characters.

writer, style, pencil, etc. In this paper, we use the set of isolated digits and alphanumeric characters, which contains 27,837 mixed alphas and numerics segmented from address blocks and 21,179 digits segmented from ZIP Codes. Since the database was extracted from handwritten text on real postal letters, the distribution of samples is not uniform, having over 1000 samples of some characters, like '1', and less than 10 samples of another ones, like 'j'. For the experiments of this paper, we drop the margins of the binary images by calculating their bounding boxes, followed by a size normalization to 32×32 pixels, preserving the aspect ratio of the handwritten sample.

In this paper, we evaluate the following two scenarios for codebook generation:

1. A *global* codebook that does not use the character class information. We just use as input all the alphanumeric character images of the CEDAR database and generate a unique global codebook.

2. A *local* character-based codebook, composed of 62 "sub-codebooks", one per each character (10 numbers and 52 letters, including lowercase and uppercase letters). In this case, we exploit the class information given by the character segmentation and labeling carried out by the forensic expert.

Clustering is then applied to the CEDAR database in order to obtain the codebooks according to these scenarios. The clustering technique used is k-means [4] because of its simplicity and computational efficiency [2]. We generate codebooks with different sizes in order to obtain the optimal size for each scenario (i.e. yielding the best performance). The maximum size for each sub-codebook in the scenario 2 depends on the number of samples of the corresponding character in the CEDAR database. For example, characters like "q" or "j" allow only codebooks of size 2 or 3, while "0" or "A" allow codebooks of up to 500 clusters. Figure 2 depicts several global codebooks of different sizes obtained according to this protocol, whereas Figure 3 depicts some of the 62 optimal "sub-codebooks" obtained in the experiments of Section 4.

PDF Computation and Matching

In this stage, the main objective is to obtain the discriminative PDF of each writer describing his/her individual shape usage preference. It is computed by building an histogram in which one bin is allocated to every codebook sample. For each alphanumeric sample of a writer, we find the nearest codebook sample using Euclidean distance. Therefore, for each writer we obtain 1 histogram (in the case of global codebook) or 62 histograms (one per character, in the case of local sub-codebooks). Each histogram is finally normalized to a PDF, which will be the discriminative feature used for recognition. To compute the similarity between two PDFs $\mathbf{o}$ and $\boldsymbol{\mu}$ from two different writers, the χ^2 distance is used:

$$\chi^2_{\mathbf{o}\boldsymbol{\mu}} = \sum_{i=1}^{N} \frac{(o_i - \mu_i)^2}{o_i + \mu_i} \qquad (1)$$

where N is the dimensionality of the vectors $\mathbf{o}$ and $\boldsymbol{\mu}$. When using a global codebook, only one distance is obtained. In the case of using 62 character-based sub-codebooks, 62 sub-distances between any two given writers are obtained, one per alphanumeric channel.

3. Database and Protocol

To evaluate the system, we use a real forensic database from original confiscated/authenticated documents provided by the Spanish forensic laboratory of the Dirección General de la Guardia Civil (DGGC). As described in Section 2, alphanumeric characters of the handwritten samples are segmented and labeled by a forensic expert of the DGGC. The whole database contains 9,297 character samples of real forensic cases from 30 different writers, with around 300 samples on average per writer distributed between a training and a

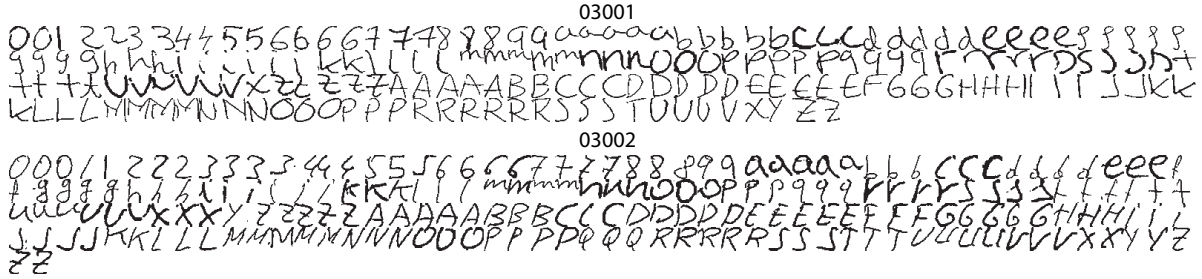


Figure 4. Training samples of two different writers of the forensic database used in this paper.

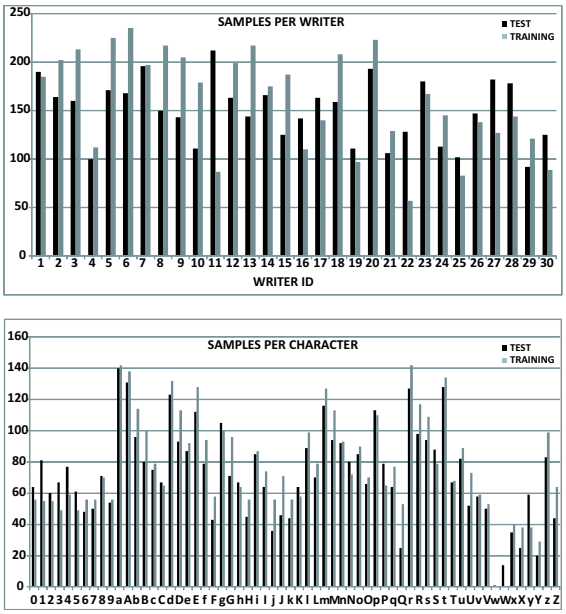


Figure 5. Distribution of samples per writer (top) and per character (bottom) of the forensic database used in this paper.

testing data set. In Figure 4 we plot the training samples of two different writers of the database. For each writer, training and testing data are extracted from different confiscated documents, meaning that they were "acquired" at different moments. As in the CEDAR database, and given the nature of the database, it does not contain uniformly distributed samples of every character. Figure 5 shows the distribution of samples per writer and per character of our database.

Given a writer of the test set, *identification experiments* are done by outputting the N closest identities of the training set. An identification is considered successful if the correct identity is among the N outputted ones. When using a global codebook, only one distance between two writers is obtained, which is used for identification. This results in 30×30=900 computed distances. When using 62 sub-codebooks, we compute the closest identity to each alphanumeric character based on the sub-distance of each channel. A decision is then made based on the majority rule: the winning output identity will be the writer having the maximum number of winning alphanumeric channels, the second winning identity will be the next writer, and so on. This results in $62\times30\times30$=55,800 computed distances. In case of writers having the same number of winning channels, they are subsequently ranked using the next 4 criteria, listed in descending order of weight: 1) average of winning sub-distances, 2) minimum winning sub-distance, 3) average of the 62 sub-distances between the test and the training writers and 4) minimum of the 62 sub-distances between the test and the training writers.

4. Results

The first step is to obtain the optimal size of the codebooks. We plot in Figure 6 results of the identification experiments depending on the size of the global codebook for a hit list size of N=1 (Top 1). We observe that the identification rate oscillates for small codebook sizes and it tends to increase with codebooks of size above 400 clusters, reaching a plateau around a size of 750.

Similarly, we vary the size of each of the 62 sub-codebooks separately in the corresponding scenario of Section 2, obtaining the identification rates of each alphanumeric channel. The optimal size of each sub-codebook is set at that resulting in the highest identification rate for a hit list size of 1. In Figure 7 we plot the best identification rate obtained for each channel, together with the optimal size of each sub-codebook. We observe that the characters with best rates are "d", "r", "s" and "N". For some characters, like "j", "q", "Q", "w" and "W", the identification rates are null. As explained in Section 2, for characters "q", "Q" and "j" we

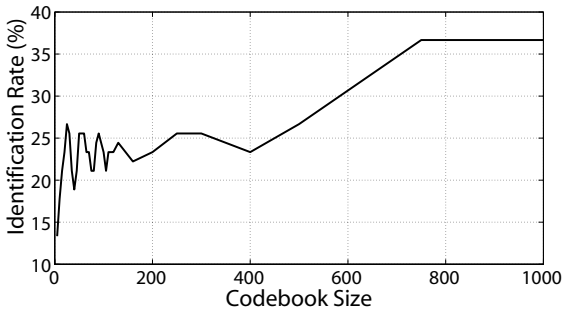


Figure 6. Writer identification rates depending on the size of the codebook (global codebook, hit list size=1).

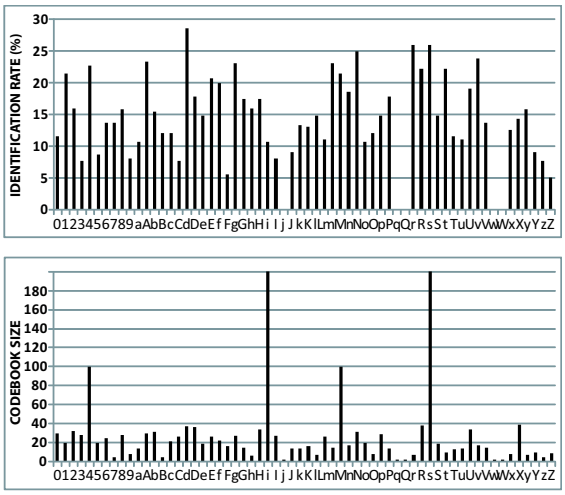


Figure 7. Best identification rates (top) and optimal size of the sub-codebook (bottom) for each individual alphanumeric channel (hit list size=1).

could only generate very small codebooks (up to 2 or 3 clusters) so their PDFs are not very discriminative. For the characters 'w' and 'W', codebooks of enough size can be generated, but in the forensic database there are not samples of them for most users, as long as this characters are not often used in the Spanish language (see Figure 5). We also observe in Figure 7 that for each character, we achieve the best identification rate with a codebook of different size. These optimal sizes are obtained for our forensic database in Spanish language, but it is expected that depending on the size and the language of the database, the size of the optimal sub-codebooks could vary.

Once we have obtained the optimal size of the codebook for each single channel, we evaluate the combination of the 62 alphanumeric channels. We plot in Figure 8 results of the identification experiments depending on the number of channels combined for a hit list size of N=1 (Top 1). Individual channels are ranked in descending order and selected according to its identification rate depicted top in Figure 7 (e.g. the channel with the highest identification rate, the two channels with the highest identification rates, etc.) We observe that the identification rate is increased with the number of channels, reaching its maximum at around 40 channels combined, and then it remains more or less constant.

We also plot in Figure 9 the identification rates varying the size of the hit list when combining 5, 10, 20, 30, 40 and all the 62 alphanumeric channels. Results are also shown for the global codebook with a size of 750 clusters (according to Figure 6). We observe that working with local sub-codebooks results in much better performance that using a unique single codebook, meaning that the class information given by the character segmentation and labeling carried out by the forensic expert provides a considerable improvement. This justifies the writer identification approach used in our forensic system, in which a considerable amount of time is spent every time a new writer is included in the database.

Concerning the system working with local sub-codebooks, we observe in Figure 9 that there are only slightly differences in performance between combining 40 and all the 62 alphanumeric channels, as mirrored previously in Figure 8. Interestingly enough, if we allow a hit list of size 8-10 (Top 8-10), the combination of only the best 10 alphanumeric channels works as well as other combinations involving more channels. On the other hand, if we want the target identity to be in the first positions of the list (Top 1-2), more alphanumeric channels are needed.

5. Conclusions and Future Work

A writer recognition system based on allographic features has been presented. It is based on manual review of the handwritten material, in which segmentation and labeling of characters is made using a dedicated software tool according to 62 alphanumeric classes (10 numbers and 52 letters, including lowercase and uppercase letters). This particular setup is used by the Spanish forensic group participating in this work, which has also provided us with a database of real forensic documents from 30 different writers, an important point in comparison with other works where data is obtained from collaborative writers under controlled conditions. Experiments are done in identification mode (one-to-

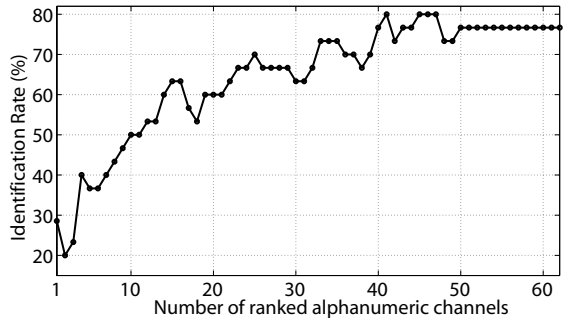


Figure 8. Writer identification rates depending on the number of alphanumeric channels combined (local sub-codebooks, hit list size=1).

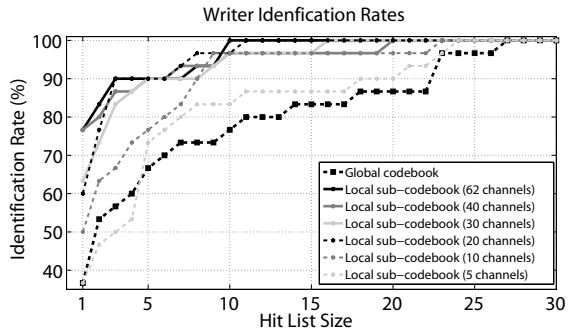


Figure 9. Writer identification rates depending on the size of the hit list size.

many), which the typical situation in forensic/criminal cases.

The system of this paper considers the writer as a stochastic pattern generator. Using a common codebook of handwritten shapes (also called allographs), the personalized set of shapes that each person uses in writing is obtained by computing their occurrence probability. Experiments are carried out using a *global* codebook (i.e. that does not use the character class information) and a set of *local* character-based sub-codebooks (i.e one per alphanumeric character, exploiting the class information given by the manual labeling). Results show that much better performance is obtained with local sub-codebooks, justifying the considerable amount of time spent by the forensic expert in the segmentation and labeling process. For the local case, we also evaluate the use of a different number of alphanumeric channels based on its individual identification rate. We observe that the best identification rate is obtained when using 40 channels, with no additional improvement given by the incorporation of additional ones. It is also worthy to note that in the case of big hit lists, the best performance is already obtained with the use of only 10 alphanumeric channels. However, for small hit lists, more alphanumeric channels are needed.

The analysis of these results with a limited database suggest that the proposed approach can be effectively used for forensic writer identification. Future work includes evaluating of our system with a bigger forensic database and applying advanced feature selection methods [5] to the combination of alphanumeric channels, including used-dependent selection approaches [19].

6. Acknowledgements

This work has been partially supported by projects Bio-Challenge (TEC2009-11186), BBfor2 (FP7 ITN-2009-238803) and "Cátedra UAM-Telefónica". Author F. A.-F. is supported by a Juan de la Cierva Fellowship from the Spanish MICINN. Author J. F. is supported by a Marie Curie Fellowship from the European Commission. The authors would like to thank to the forensic "Laboratorio de Grafística" of the 'Dirección General de la Guardia Civil' for its valuable support.

References

[1] A. Bensefia, T. Paquet, L. Heutte. Information retrieval-based writer identification. *Proc. ICDAR*, 2003.
[2] M. Bulacu, L. Schomaker. A comparison of clustering methods for writer identification and verification. *Proc. ICDAR*, 2005.
[3] M. Bulacu, L. Schomaker. Text-independent writer identification and verification using textural and allographic features. *IEEE PAMI*, 29(4):701–717, April 2007.
[4] R. Duda, P. Hart, D. Stork. *Pattern Classification*, 2004.
[5] J. Galbally, J. Fierrez, M. R. Freire, J. Ortega-Garcia. Feature selection based on genetic algorithms for on-line signature verification. *Proc. AutoID*, 2007.
[6] J. Hull. A database for handwritten text recognition research. *IEEE PAMI*, 16(5):550–554, May 1994.
[7] A. Jain, P. Flynn, and A. Ross, editors. *Handbook of Biometrics*. Springer, 2008.
[8] N. Otsu. A threshold selection method for gray-level histograms. *IEEE SMC*, 9:62–66, December 1979.
[9] R. Plamondon, S. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE PAMI*, 22(1):63–84, 2000.
[10] L. Schomaker. Advances in writer identification and verification. *Proc. ICDAR*, 2007.
[11] L. Schomaker. *Sensors, Systems and Algorithms, Advances in Biometrics*, chapter Writer identification and verification. Springer Verlag, 2008.
[12] L. Schomaker, M. Bulacu. Automatic writer identification using connected-component contours and edge-based features of upper-case western script. *IEEE PAMI*, 26(6):787-798, 2004.
[13] L. Schomaker, M. Bulacu, and K. Franke. Automatic writer identification using fragmented connected-component contours. *Proc. IWFHR*, 2004.
[14] S. Srihari, C. Huang, H. Srinivasan, V. Shah. *Digital Document Processing*, ch. 17. Biometric and Forensic Aspects of Digital Document Processing. Springer, 2007.
[15] S. Srihari and G. Leedham. A survey of computer methods in forensic document examination. *Proc. IGS*, 2003.
[16] S. N. Srihari, S.-H. Cha, H. Arora, S. Lee. Individuality of handwriting. *J. Forensic Sc.*, 47(4):856–872, 2002.
[17] M. Tapiador. *Análisis de las Características de Identificación Biométrica de la Escritura Manuscrita y Mecanográfica*. PhD thesis, Escuela Politécnica Superior, Universidad Autónoma de Madrid, 2006.
[18] M. Tapiador, J. Sigüenza. Writer identification method based on forensic knowledge. *Proc. ICBA*, 2004.
[19] J. Fierrez-Aguilar, D. Garcia-Romero, J. Ortega-Garcia, J. Gonzalez-Rodriguez. Adapted user-dependent multimodal biometric authentication exploiting general information. *Patt. Recogn. Letters*, 26:2628–2639, 2005.