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

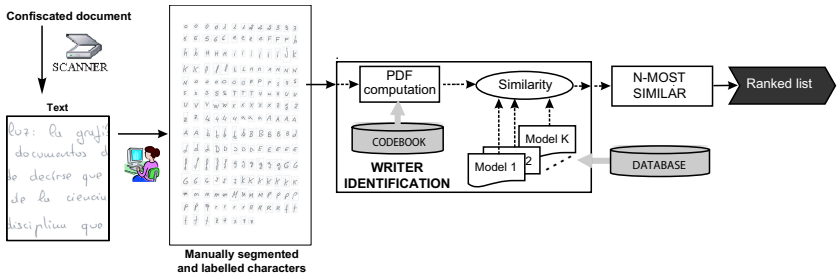

Fig. 1. Modelo del sistema de identificación forense de escritor basado en características alográficas.

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.

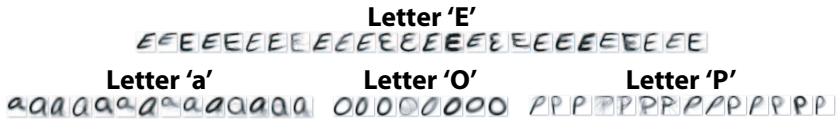

Fig. 3. Ejemplo de subcatálogos óptimos para algunos caracteres.

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

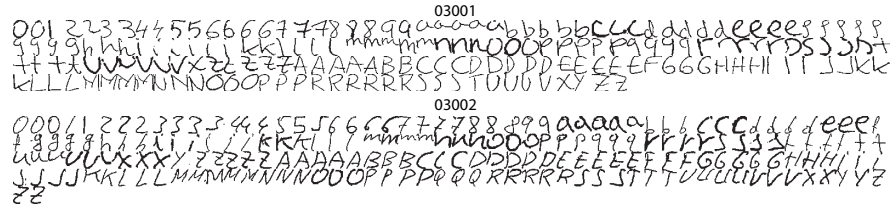

Fig. 4. Muestras de entrenamiento de dos escritores distintos de la base de datos forense.

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

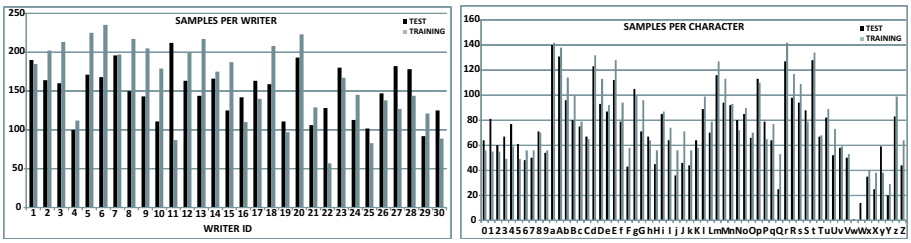

Fig. 5. Distribución de muestras por escritor (izquierda) y por carácter (derecha) de la base de datos forense utilizada en este trabajo.

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.

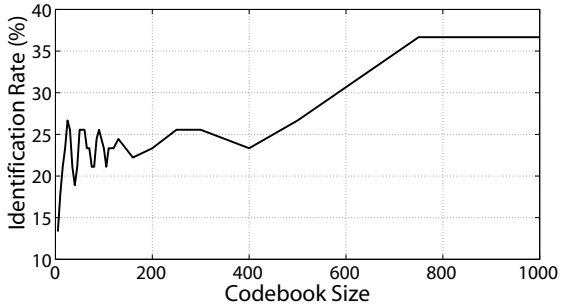

Fig. 6. Tasas de identificación de escritor en función del tamaño del catálogo (catálogo global, tamaño de lista=1).

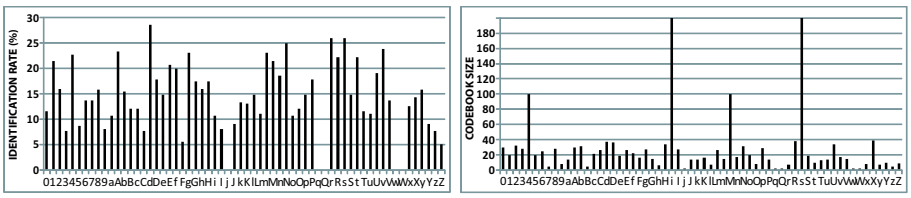

Fig. 7. Mejores tasas de identificación (izquierda) y tamaño óptimo del sub-catálogo (derecha) para cada canal alfanumérico (tamaño de lista=1).

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

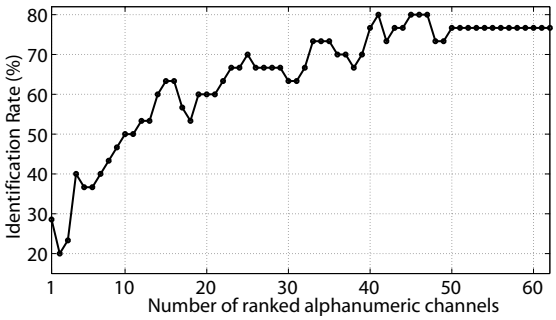

Fig. 8. Tasas de identificación de escritor en función del número de canales alfanuméricos combinados (sub-catálogos locales, tamaño de lista=1)

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

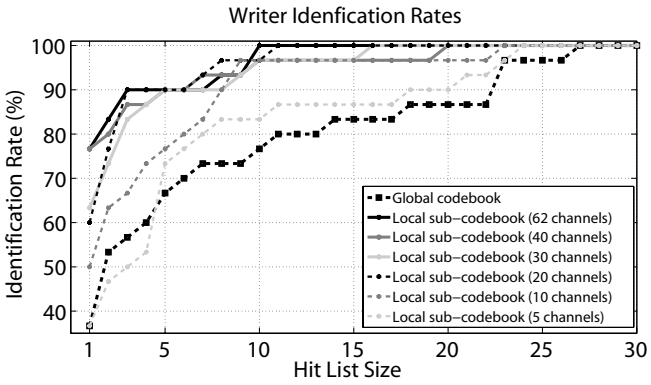

Fig. 9. Tasas de identificación de escritor en función del tamaño de la lista.

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