

Un nuevo descriptor para la identificación de personas mediante caracteres simples

Verónica Aubin(1), Marco Mora(2), and Matilde Santos(3)

(1)Universidad Nacional de La Matanza, Argentina

(2)Universidad Católica del Maule, Chile

(3)Universidad Complutense de Madrid, España

vaubin@unlam.edu.ar, mora@spock.ucm.cl, msantos@ucm.es

Abstract. En este artículo se presenta un nuevo descriptor para la verificación de la identidad de personas en base al análisis de imágenes en escala de grises de caracteres manuscritos individuales y simples. El descriptor corresponde a los coeficientes B-Spline de la curva de posición relativa de los puntos de mínimo valor de gris dentro del carácter. Estos puntos corresponden a los píxeles de menor valor de gris sobre la línea recta perpendicular a los puntos del esqueleto morfológico del trazo. La posición relativa se computa como la distancia euclidia entre el punto de mínimo gris y su correspondiente en el esqueleto morfológico del trazo. Se utilizó un clasificador multiclase, basado en Máquinas de Vectores Soporte de salida binaria, para evaluar la capacidad de discriminación del descriptor propuesto. Se utilizó una base de datos con 50 muestras de 6 símbolos simples realizadas por 50 personas. La experimentación de la base de datos muestra resultados muy satisfactorios, con un promedio de aciertos del 97%, y permiten pensar que es factible desarrollar un método de identificación de personas en base al descriptor presentado.

Keywords: Reconocimiento de personas, Análisis off-line, Trazos manuscritos, Puntos de mayor presión.

1. Introduction

La escritura manuscrita como patrón biométrico de comportamiento ha cobrado un renovado interés por parte de los investigadores en los últimos años [1]. A pesar de avanzar hacia el sueño de un futuro puramente digital, el estudio de la escritura usando imágenes digitales ha conservado su lugar debido a sus aplicaciones en la vida real: en el análisis forense de documentos, la clasificación de los archivos históricos [2], la verificación de firmas [3], estudios sobre la correlación entre la escritura y diferentes trastornos neurológicos [4], así también como en diferentes campos como la seguridad y control de acceso [5] [6].

La escritura a mano es una de las técnicas biométricas primitivamente utilizadas para autenticar un individuo. Los trazos manuscritos de una persona tienen una gran variabilidad; no obstante, existen características invariantes que permiten el reconocimiento manual o automático del autor. En el modo verificación, el

sistema valida la identidad de una persona comparando el rasgo biométrico capturado en la entrada con su propia plantilla biométrica previamente almacenada en la base de datos. En general, el usuario indicará su identidad mediante un número de identificación personal, un nombre de usuario o algún tipo de código. Posteriormente el sistema realizará una comparación uno a uno para determinar si el individuo es quien dice ser.

Muchos de los trabajos relacionados con esta investigación provienen del problema de la verificación de firmas. La verificación off line de firmas es un tema bien documentado y se ha abordado con muchos enfoques diferentes. Una serie de documentos cubren los avances en este campo [7] [3]. Recientemente en [8] se presenta un estudio de la literatura, hasta la fecha, sobre la identificación y verificación off line del autor de escrituras a mano considerando las distintas características y enfoques de clasificación, en diferentes idiomas y alfabetos. En [9] se presenta un análisis de imágenes off-line de escritura a mano basado en textura para la identificación del escritor; cercana a nuestra propuesta. La técnica que usan divide una letra en fragmentos pequeños y considera cada fragmento como una textura. En [10] se consideran caracteres bengalí aislados para la verificación del escritor. Se presenta el rendimiento de diferentes características texturales. Una colección de 500 documentos en bengalí de 100 escritores se utilizan en este sentido. La combinación de características obtiene un mejor rendimiento. La evaluación de los resultados muestra que el método es efectivo y se puede aplicar en la base de datos de gran tamaño. En [11] se plantea la identificación de personas a través de características dinámicas extraídas de la imagen en escala de grises de un carácter. El descriptor presentado considera el patrón de la distribución de presión a lo largo del trazo que es discriminante entre individuos, es decir considera el nivel de gris de los píxeles del trazo y la ubicación en el mismo. Se encontró que, la distancia relativa, medida en una perpendicular al esqueleto del trazo, entre la línea que une los puntos más oscuros de la imagen respecto del esqueleto, es característica del autor, y varía de un individuo a otro.

Al momento de implementar un sistema biométrico deben considerarse aspectos prácticos, tales como el rendimiento computacional y el espacio de almacenamiento. El uso de técnicas de selección de características son necesarias para reducir el espacio de almacenamiento y mejorar el tiempo de compilación.

En este artículo se presenta un nuevo descriptor para la identificación de personas, basado en la extracción de características pseudo-dinámicas de imágenes en escala de grises de caracteres manuscritos individuales y simples. El descriptor corresponde a los coeficientes B-Spline de la curva de posición relativa de los puntos de mínimo valor de gris dentro del carácter. Estos puntos corresponden a los píxeles de menor valor de gris sobre la línea recta perpendicular a los puntos del esqueleto morfológico del trazo. La posición relativa se computa como la distancia euclidia entre el punto de mínimo gris y su correspondiente en el esqueleto morfológico del trazo. Es de destacar la importancia de que el descriptor propuesto extrae características de un trazo simple, en lugar de un trazo complejo como puede ser la firma de un individuo, lo que permite tomar muchas muestras de un mismo documento. Además de lo anterior, el descriptor presenta

la particularidad de ser invariante a la rotación, lo que aporta una ventaja importante dentro del procesamiento de imágenes. La clasificación se realizó con una SVM multiclase, contemplando el problema de clases desbalanceadas, para evaluar la capacidad de discriminación del descriptor propuesto. Se analizan los resultados obtenidos de la clasificación para las muestras realizadas por 50 personas, correspondientes a la base de datos creada especialmente con 6 símbolos simples.

La estructura de este trabajo es la siguiente. La sección 2 presenta conceptos de aproximación B-spline aplicados en el trabajo. La sección 3 presenta el descriptor propuesto. La sección 4 explica en detalle la metodología empleada para la validación de los resultados y la base de datos creada especialmente. La sección 5 presenta los resultados de la evaluación objetiva del descriptor. Finalmente, en la sección 6, se presentan las conclusiones del trabajo.

2. Conceptos de aproximación B-spline

La teoría de las B-spline es muy amplia y se expone en detalle en [12]. En ésta sección sólo se desarrollan los elementos necesarios de dicha teoría para la comprensión del uso de las B-spline en esta investigación.

La B-spline es una combinación lineal de $n+1$ puntos de control $c_1, \dots, c_n, n \geq k$, con funciones bases $N_{i,k}$. Estas funciones son polinomios formados a partir del parámetro u tal como se muestra en la siguiente expresión:

$$f(u) = \sum_{i=0}^L c_i N_{i,k}(u) \quad (1)$$

siendo k el grado de la curva B-spline y L se corresponde con el número de segmentos en que se divide el espacio de aproximación.

Las bases B-spline $N_{i,k}$ pueden definirse según la siguiente expresión recursiva:

$$N_{i,j}(t) = \frac{t - t_i}{t_{i+j} - t_i} N_{i,j-1}(t) + \frac{t_{i+j+1} - t}{t_{i+j+1} - t_{i+1}} N_{i+1,j-1}(t) \quad (2)$$

donde el corte de la recursividad corresponde a la expresión

$$N_{i,0}(t) = \begin{cases} 1 & t_i \leq t \leq t_{i+1} \\ 0 & \text{otros casos} \end{cases} \quad (3)$$

Dados m puntos (x_i, y_i) de una curva, se puede expresar el problema de aproximación de la curva mediante el siguiente sistema de ecuaciones lineales:

$$y_i = \sum_{j=0}^{n-1} c_j N_{j,k}(x_i) \quad (4)$$

donde los coeficiente c_j son desconocidos.

Matricialmente el sistema de ecuaciones se expresa de la siguiente manera:

$$Y = N \times C \quad (5)$$

donde:

$$Y = [y_0 y_2 y_3 \dots y_{m-1}]^T \quad (6)$$

$$N = \begin{pmatrix} N_{0,3}(x_0) & N_{2,3}(x_0) & \dots & N_{n-1,3}(x_0) \\ N_{0,3}(x_1) & N_{2,3}(x_1) & \dots & N_{n-1,3}(x_1) \\ \dots & \dots & \dots & \dots \\ N_{0,3}(x_{m-1}) & N_{2,3}(x_{m-1}) & \dots & N_{n-1,3}(x_{m-1}) \end{pmatrix} \quad (7)$$

$$C = [c_0 c_2 c_3 \dots c_{n-1}]^T \quad (8)$$

El sistema planteado está sobredeterminado, es decir, tiene más ecuaciones que incógnitas. Este sistema sobredeterminado se resuelve a través de la matriz pseudoinversa por izquierda (L) según la siguiente expresión:

$$L = M^{-1} N^T \quad (9)$$

siendo $L_{(n \times m)} \cdot N_{(m \times n)} = I_{(n \times n)}$. Para $m > n$ puede calcularse $L_{(n \times m)}$ construyendo la matriz $M_{(n \times n)}$ como:

$$M = N^T N \quad (10)$$

Si el determinante de (M) $\neq 0$, existe M^{-1} y puede escribirse:

$$M^{-1} N^T N = M^{-1} M \quad (11)$$

La solución del sistema sera:

$$C = LY \quad (12)$$

3. Presentación del Descriptor

El descriptor de este trabajo corresponde a los coeficientes de la aproximación B-spline de la curva que representa la posición relativa de los puntos de mínimo gris dentro del grafema. Los puntos de mínimo gris corresponden al pixel de menor valor de gris sobre la línea perpendicular al esqueleto del grafema. La posición relativa corresponde a la distancia euclídea entre el punto de mínimo gris y el borde del trazo, como se puede observar en la figura 1(a). Considerando las distancias de todos los puntos de mínimo gris se conforma una señal que es característica para cada individuo. A esta señal o curva se la denomina, de aquí en adelante, como señal de posición relativa de los puntos de mínimo gris.

En general las señales de distancia de posición relativa tienen distinta cantidad de puntos dependiendo de la imagen. Lo anterior implica que es necesaria una etapa de normalización de la longitud de la curva, representada en la figura 1(b). Esta normalización se realizó redimensionando la señal a la longitud

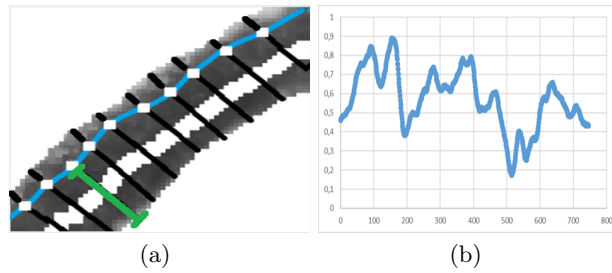


Figura 1. Descriptor: (a) señal de posición relativa de los puntos de mínimo gris (b) Normalización de la señal caracter C

promedio de todas las señales de la base de datos. Luego de la normalización anterior se observa que las señales tienen una longitud demasiado grande para implementar una etapa de clasificación supervisada (sobre 700 puntos aproximadamente). Para reducir la dimensionalidad del descriptor se recurre al cómputo de los coeficientes de la interpolación B-spline de la señal de distancia relativa normalizada.

Como es ampliamente sabido, si se utilizan pocas bases la aproximación de interpolación tiene un error elevado respecto de la señal original. A medida que se van aumentando las bases el error disminuye y la aproximación se va ajustando cada vez más respecto de la original. No obstante lo anterior llega un punto que el error deja de disminuir de manera significativa, reduciéndose el error de manera muy marginal. Para determinar la cantidad adecuada de coeficientes

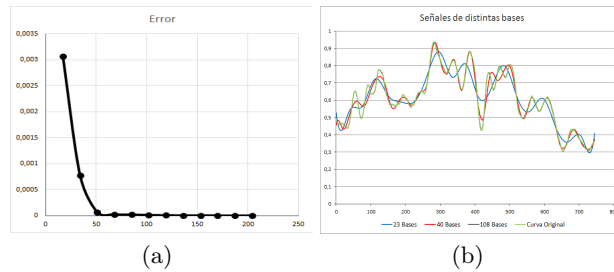


Figura 2. Error: (a) Disminución del Error (b) Señales de distintas bases y curva original

se comienza a aumentar de forma paulatina la cantidad de bases hasta que la disminución del error es despreciable. Para mostrar este fenómeno se presenta la figura 2. Se escogió aleatoriamente una de las señales de distancia relativa de todo el conjunto disponible y con esta señal se realizan las aproximaciones aumentando las bases. En la figura 2(a) se observa que basta con 74 coeficientes para reconstruir nuevamente la señal original con un muy bajo error. En la figura

2(b) se presentan 3 casos de aproximaciones para 23 (azul), 40(rojo) y 108(gris) bases y la curva original(verde). Se observa que la aproximación con la mayor cantidad de bases se superpone casi perfectamente a la curva original.

4. Materiales y Métodos

Para evaluar la capacidad de discriminación de personas del descriptor propuesto se implementó un clasificador multiclase, basado en SVM de salida binaria, entrenado con Validación Cruzada y considerando el problema de desbalanceo de clases usado en [13]. La base de datos de símbolos simples de donde obtener las muestras para la experimentación se detalla a continuación .

4.1. Base de Datos

Un total de 15.000 imágenes de grafemas forman la base de datos construida especialmente. Las imágenes son a color de 24 bits y de mediana resolución, aproximadamente de 800 por 800 píxeles. Se consideraron las 6 grafemas que se muestran en la figura 3, por ser rasgos simples con trazos ascendentes, descendentes y curvos en distintas direcciones, que dan lugar a los gestos gráficos que revelan las características personales de su autor. 50 personas realizaron 50 muestras de cada una de los grafemas escogidos, de acuerdo con los modelos presentados. Las muestras se recogieron bajo condiciones controladas, los cola-

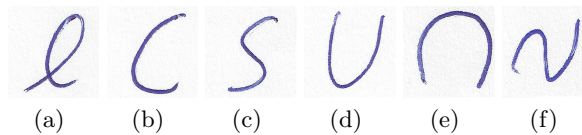


Figura 3. Símbolos de la base de datos.

boradores escribieron sobre una hoja de papel de $75(g/m^2)$, utilizando 5 hojas del mismo tipo como base de apoyo y el instrumento de escritura fue un bolígrafo bic trazo grueso de color azul, cuya bolita está fabricada en tungsteno de 0,7 mm. La captura de las imágenes se realiza mediante un escáner convencional, lo que permite realizar una captura masiva de imágenes utilizando un dispositivo de uso cotidiano.

4.2. Modelo de clasificación

A continuación se presenta en detalle la metodología de entrenamiento y clasificación. El reconocimiento de personas por un carácter simple, planteado en este artículo, es un problema multiclase. Las muestras correspondientes a un

grafema, realizadas por una misma persona, forman una clase. Por consiguiente en la base de datos presentada se definen 50 clases diferentes para un símbolo. Para resolver el problema se construyó un sistema de múltiples clasificadores binarios (SVM). Cada clasificador distingue entre una de las clases y el resto como segunda clase, técnica conocida como uno-contra-todos (one-versus-all) [14]. Para formar un grupo balanceado, se decidió optar por eliminar casos de la clase mayoritaria, consiguiendo una relación de 1 a 5. El conjunto queda de 295 muestra, 50 de la persona elegida como clase positiva y 5 muestras de cada una de las restantes personas, clase negativa. SVM necesita dos etapas: el entrenamiento y la fase de pruebas. El conjunto de entrenamiento se utiliza para ajustar los parámetros del modelo, y el de prueba, para evaluar los parámetros de los mismos. Para garantizar que los resultados sean independientes de la partición entre datos de entrenamiento y prueba se utilizó en el modelo validación cruzada aleatoria con 10 iteraciones, dividiendo las muestras de datos balanceadas en un 70 % para el entrenamiento y 30 % para el test. Los datos en el modelo planteado, no son linealmente separables a través de un hiperplano óptimo en el espacio de entrada, por lo que se construye un sistema con función Kernel de base radial (RBF). Siguiendo el desarrollo de [15], la expresión del hiperplano de separación L_D es la siguiente:

$$L_D = \sum_j \alpha_j - \frac{1}{2} \sum_j \sum_k \alpha_j \alpha_k y_j y_k x'_i x_k \quad (13)$$

el problema de SVM consiste en maximizar el funcional L_D sujeto a las siguientes restricciones:

$$\begin{aligned} 0 &\leq \alpha_j \leq C \\ \sum_j \alpha_j y_j &= 0 \end{aligned} \quad (14)$$

donde α_j son los multiplicadores de Lagrange, y C es un parametro que permite mantener acotados dichos multiplicadores. Los α_j no nulos corresponden a los vectores de soporte y determinan frontera de decisión. En particular, la expresión de la función de Base Radial es la siguiente:

$$K(x, y) = \exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right) \quad (15)$$

donde σ es el parámetro de la función kernel.

Se realiza el cómputo de los parámetros del SVM con optimización aleatoria. Se itera 100 veces el entrenamiento del clasificador SVM con validación cruzada. Por cada iteración se obtiene la pérdida, error cuadrático medio, de los 10 folds. Se calculan los promedios por iteración y se obtienen los parámetros óptimos del modelo, correspondientes al promedio mínimo de las pérdidas por iteración.

Se entrena el clasificador SVM de los parámetros computados con validación cruzada. Para lo cual se hace una nueva partición y se calcula el promedio de las pérdidas y se elige como modelo el que tiene pérdidas más cercanas al promedio. Se evalúa este modelo SVM. Se calcula la matriz de confusión para el grupo de muestras balanceado considerado. Para obtener mayor independencia de los

datos de entrada se forman 20 grupos de muestras balanceados por persona, que se usan para entrenar y evaluar el modelo, obteniéndose como resultado una matriz de confusión por grupo. Luego se calcula la matriz de confusión promedio para cada persona. Se obtienen los siguientes indicadores a partir de la matriz de confusión:

- $TP = \text{AciertosPositivos}$
- $TN = \text{AciertosNegativos}$
- $FP = \text{FalsosPositivos}$
- $FN = \text{FalsosNegativos}$
- $ACIERTOS = TP + TN$
- Sensibilidad o la verdadera tasa positiva

$$TPR = \frac{TP}{TP+FN}$$
- Especificidad o la verdadera tasa negativa

$$TNR = \frac{TN}{FP+TN}$$
- $RHO = \frac{TP}{TP+FP+FN}$

5. Resultados

En esta sección se presentan resultados numéricos que permiten evaluar y comparar el funcionamiento en la verificación de personas de la señal de distancia relativa y de los coeficientes de la aproximación B-spline.

En la tabla 1 se observan los resultados obtenidos para la letra C de los 50 individuos de la base de datos, tanto para la señal de distancia relativa como para los coeficientes B-spline. Cada fila corresponde al promedio de la evaluación del conjunto de test de 20 grupos balanceados. La última fila corresponde al promedio de las 50 personas. Como se observa de la última fila, cada SVM en promedio reconoce correctamente la identidad del 97% de los individuos para ambos descriptores.

La tabla 2 presenta los resultados para todos los símbolos. Se muestran los valores en promedio obtenido para cada uno de los indicadores considerados a partir de la matriz de confusión. Se observa que el porcentaje de reconocimiento utilizando todos los símbolos es superior a 97%.

Persona	% Aciertos Reducción	% Aciertos Descriptor	Persona	% Aciertos Reducción	% Aciertos Descriptor
1	98	98	26	96	97
2	96	97	27	96	97
3	98	97	28	96	97
4	96	95	29	95	95
5	98	98	30	97	98
6	97	96	31	97	96
7	98	97	32	98	99
8	95	95	33	99	98
9	97	98	34	100	100
10	98	99	35	96	97
11	97	97	36	96	97
12	97	98	37	95	93
13	98	98	38	97	97
14	98	97	39	98	97
15	98	97	40	98	98
16	96	94	41	98	98
17	97	97	42	94	93
18	99	98	43	98	98
19	98	97	44	98	97
20	96	97	45	99	96
21	95	95	46	97	98
22	99	97	47	98	97
23	96	96	48	98	97
24	98	98	49	96	96
25	98	99	50	94	92
			Promedio	97	97

Tabla 1. Símbolo C

Simbolo	TP	TN	FP	FN	ACIERTOS %	TPR	TNR	RHO
C	43	244	1	7	97	85	100	0.83
~	44	244	1	6	97	87	100	0.85
e	42	244	1	8	97	85	100	0.83
∩	42	244	1	8	97	84	100	0.82
s	44	244	1	6	98	87	100	0.86
u	42	244	1	8	97	84	100	0.82

Tabla 2. Promedio de los resultados por símbolo

6. Conclusión

Este artículo ha presentado un nuevo descriptor para identificación de personas usando trazos manuscritos. El descriptor propuesto utiliza los coeficientes de la aproximación B-spline de la curva que representa la posición relativa de los puntos de mínimo gris dentro del grafema. Además de lo anterior, se considera una imagen de un trazo manuscrito simple, en lugar de un trazo complejo como la firma de un individuo. Para evaluar el funcionamiento del descriptor se ha realizado un estudio con un clasificador multiclase, basado en SVM de

salida binaria, entrenado con Validación Cruzada y considerando el problema de desbalanceo de clases. Se obtuvo que, en promedio, el descriptor reconoce correctamente la identidad del 97% de los individuos.

Referencias

1. Schomaker, L.: Advances in writer identification and verification. In: Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). Volume 2., IEEE (2007) 1268–1273
2. Arabadjis, D., Giannopoulos, F., Papaodysseus, C., Zannos, S., Rousopoulos, P., Panagopoulos, M., Blackwell, C.: New mathematical and algorithmic schemes for pattern classification with application to the identification of writers of important ancient documents. *Pattern Recognition* **46** (2013) 2278–2296
3. Hafemann, L.G., Sabourin, R., Oliveira, L.S.: Offline handwritten signature verification-literature review. arXiv preprint arXiv:1507.07909 (2015)
4. Smekal, Z., Mekyska, J., Rektorova, I., Faundez-Zanuy, M.: Analysis of neurological disorders based on digital processing of speech and handwritten text. In: *Signals, Circuits and Systems (ISSCS), 2013 International Symposium on*, IEEE (2013) 1–6
5. Horster, P.: *Communications and Multimedia Security II*. Springer (2016)
6. Vielhauer, C.: *Biometric user authentication for IT security: from fundamentals to handwriting*. Volume 18. Springer Science & Business Media (2005)
7. Impedovo, D., Pirlo, G., Russo, M.: Recent advances in offline signature identification. In: *2014 14th International Conference on Frontiers in Handwriting Recognition*. (2014)
8. Halder, C., Obaidullah, S.M., Roy, K.: Offline writer identification and verification-a state-of-the-art. In: *Information Systems Design and Intelligent Applications*. Springer (2016) 153–163
9. Hannad, Y., Siddiqi, I., El Kettani, M.E.Y.: Writer identification using texture descriptors of handwritten fragments. *Expert Systems with Applications* **47** (2016) 14–22
10. Halder, C., Obaidullah, S.M., Paul, J., Roy, K.: Writer verification on bangla handwritten characters. In: *Advanced Computing and Systems for Security*. Springer (2016) 53–68
11. Aubin, V., Doorn, J., Kaplan, G.: Nuevos descriptores para la Identificación de Personas basados en la Simetría del Trazo. In: XVIII Congreso Argentino de Ciencias de la Computación. (2013) Available in <http://sedici.unlp.edu.ar/handle/10915/31655>.
12. Farin, G.: *Curves and surfaces for computer-aided geometric design: a practical guide*. Elsevier (2014)
13. Aubin, V., Mora, M., Santos, M.: A new descriptor for people recognition by handwritten strokes analysis. *International Conference on Pattern Recognition Systems ICPRS-16* (2016) 14–20 doi:10.1049/ic.2016.0038.
14. Gomez-Moreno, H., Gil-Jimenez, P., Lafuente-Arroyo, S., Vicen-Bueno, R., Sanchez-Montero, R.: Color images segmentation using the support vector machines. *Recent Advances in Intelligent Systems and Signal Processing* (2003) 151–155
15. Matlab: *Statistics and Machine Learning Toolbox*. The MathWorks, Inc. (2015) Available in <http://www.mathworks.com/help/stats/support-vector-machines-for-binary-classification.html>.