Autenticación Web de Estudiantes Mediante Reconocimiento Biométrico

Biometric Student Authentication for e-Learning Platforms

Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, Javier Ortega-Garcia
{aythami.morales, julian.fierrez, ruben.vera, javier.ortega}@uam.es

Departamento de Tecnologia Electronica y de las Comunicaciones
EPS, Universidad Autonoma de Madrid
Madrid, Spain

Resumen- En este trabajo se hace un análisis de las tecnologías de reconocimiento biométrico basado en dinámica de tecleo para la autenticación de estudiantes en entornos web. A diferencia de la autenticación basada en algo que tenemos (tarjetas) o algo que sabemos (*passwords*), el reconocimiento biométrico hace uso de características propias de los individuos (algo que somos) para verificar sus identidades. En este trabajo se estudian las características de estos sistemas así como su idoneidad para su aplicación en entornos docentes. Se incluye también un experimento práctico en el que se analiza el patrón biométrico de 64 alumnos. El experimento permite analizar el rendimiento de dos sistemas de reconocimiento a través de la dinámica de tecleo de los alumnos a lo largo de 3 exámenes elaborados durante un semestre. Los resultados muestran una tasa de reconocimiento superior al 90% lo cual anima a seguir investigando esta línea para su implantación en entornos reales.

Palabras clave: reconocimiento biométrico, autenticación, dinámica de tecleo, MOOC, POOC

Abstract- this work analyzes biometric technologies for user authentication in online educational platforms. Our study provides new insights on the deployment of these technologies in educational environments with special attention to keystroke dynamics systems. Concisely, this work studies the advantages/disadvantages of keystroke dynamics recognition for online student authentication services including an end-to-end approach and experiments performed over real student data. This work includes a case study on keystroke dynamic authentication over the typing patterns of 64 students answering questions in 3 online exams over a semester. We analyze the performance of four keystroke biometric systems and the results obtained show a promising performance.

Keywords: keystroke dynamics, user authentication, biometrics, MOOC, POOC

1. INTRODUCCIÓN

La educación en línea ha dado un salto en los últimos años como una nueva forma de educación de carácter abierto, gratuito y participativo. Los MOOCs/POOCs (siglas de los términos en inglés *Massive Open Online Course* y *Participatory Open Online Course*) han supuesto una verdadera revolución en los modelos educativos. Los cursos en línea rompen con las barreras asociadas a las tradicionales lecciones presenciales y ofrecen una educación altamente accesible a través de internet. Un alumno de cualquier parte del mundo puede participar en un curso de Ingeniería Aplicada ofertado por el MIT (Boston) y complementarlo con otro de Programación Avanzada ofertado por Oxford (Reino Unido), y todo sin moverse de su casa. Este nuevo escenario educativo ha suscitado un amplio debate entre todos los actores de la comunidad educativa. Entre los puntos que ha generado mayor controversia se encuentra la certificación de los títulos no presenciales ofertados a través de estas plataformas (Stanford 2013).

¿Cómo podemos asegurar que el alumno que obtiene el título/certificado es la persona que ha realizado el curso? ¿Cómo podemos detectar a los usuarios que usan de forma malintencionada las plataformas? La naturaleza accesible y universal de este tipo de cursos incrementa la vulnerabilidad de los mismos y la autenticación de los alumnos es una tarea difícil de solventar. Los investigadores y docentes implicados en cursos en línea son conscientes de la importancia de la autenticación fiable de los alumnos para el futuro de este tipo de cursos y se han hecho grandes esfuerzos por analizar opciones y tecnologías que cumplan con las necesidades de este tipo de autenticación (Miguel, Caballe y Prieto, 2013). Las tecnologías de reconocimiento biométrico surgen como una vía para conseguir esta autenticación fiable de estudiantes. Dichas tecnologías se basan en "algo que somos" en lugar de las tecnologías tradicionales basadas en "algo que sabemos o tenemos" como PIN o passwords.

De entre las tecnologías biométricas, el reconocimiento por cadencia de tecleo (*keystroke dynamics*) ha atraído el interés de investigadores e industria debido a su fácil implantación y conveniencia en aplicaciones relacionadas con la interacción hombre-máquina (Peacock, Ke y Wilkerson. 2004; Gunetti y Picardi, 2005; Morales, Fierrez y Ortega-Garcia, 2014). La tecnología de reconocimiento por cadencia de tecleo es atractiva principalmente por dos razones: i) transparencia (permite el reconocimiento sin requerir una participación explícita del usuario en la autenticación), y ii) continua (la autenticación se realiza durante toda la actividad del usuario, no solo durante el acceso a la plataforma). La transparencia está relacionada con la experiencia de uso de la plataforma. La autenticación de usuarios debe ser simple y no afectar al normal uso de las plataformas. Mientras que la autenticación continua es crítica en un escenario como el propuesto en los MOOCs

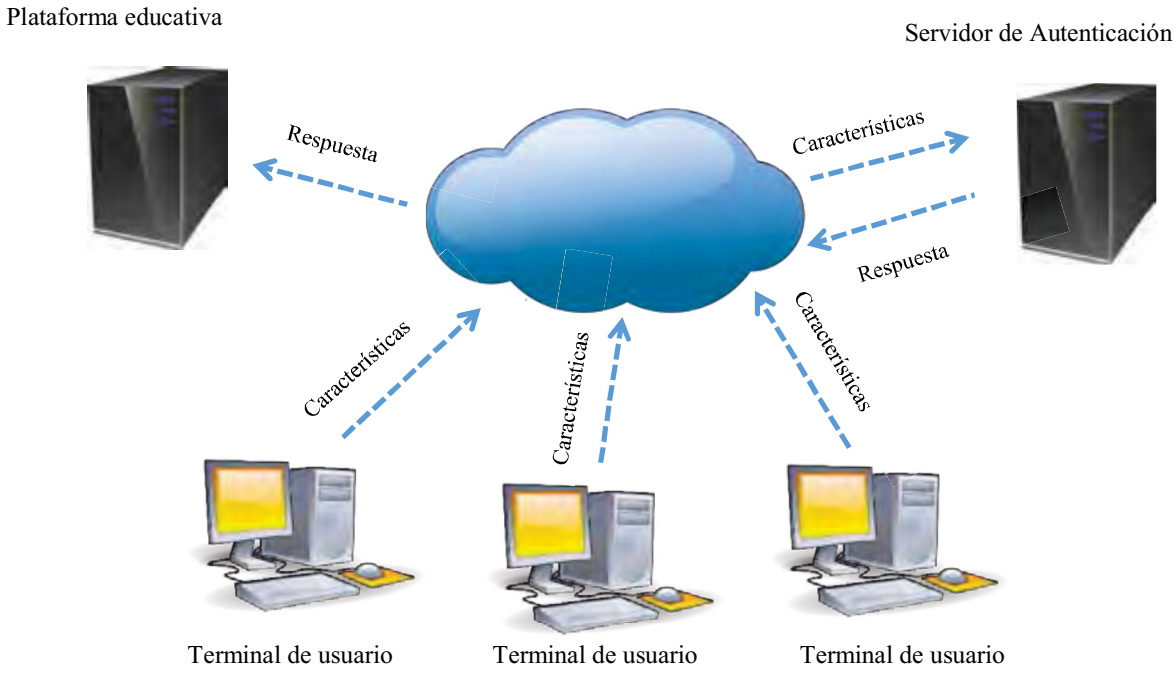

Figura 1: Proceso de autenticación de estudiante mediante arquitectura basado en servidor de autenticación en la nube

debido a que la autenticación basada en PIN/passwords asume que el usuario no cederá su clave a un tercero. Esta asunción no es válida en un escenario en el que un estudiante puede ceder su código a otro con el objetivo de que este realice un examen por él.

El objetivo principal de este trabajo es analizar el uso de tecnologías de reconocimiento biométrico en entornos docentes web. Se hace un estudio de las características de estas tecnologías con especial interés en el reconocimiento por dinámica de tecleo. Se incluye un caso de estudio en el que se hace uso de un sistema da autenticación biométrica basada en dinámica de tecleo para reconocer a los alumnos de un curso introductorio de *computer science*. Los resultados avalan el uso de estas tecnologías aunque aún hay espacio para mejoras.

El resto del trabajo se organiza de la siguiente forma: la sección 2 incluye el contexto de este trabajo así como un análisis de las principales ventajas y desventajas de las tecnologías de reconocimiento biométrico para autenticación de estudiantes. La sección 3 incluye la metodología seguida así como los algoritmos evaluados. La sección 4 presenta los resultados y por último en la sección 5 se extraen las principales conclusiones del trabajo.

2. Contexto

En una sociedad como la actual cada vez más digitalizada, la autenticación de usuarios en entornos web es una necesidad. La enseñanza en línea no es ajena a esta necesidad y el estudio y desarrollo de sistemas de autenticación de estudiantes es una importante área para investigadores, desarrolladores y usuarios finales.

A. Autenticación de estudiantes para e-learning

La naturaleza no presencial de los cursos a distancia ha promovido la necesidad de una evaluación más continua que la llevada a cabo por norma general en los tradicionales cursos presenciales. Esta evaluación continua lleva aparejada la necesidad de desarrollar sistemas de control sofisticados que aseguren la integridad del sistema. La autenticación del alumno es una de los retos a resolver que implica mayores dificultades técnicas, legales y de conveniencia de uso. A la hora de escoger un sistema de autenticación de alumnos, deberíamos analizar algunos aspectos claves:

- **Seguridad:** la robustez del sistema ante posibles ataques malintencionados, robos de identidad o de datos. Éste es un aspecto crucial en entornos web debido fundamentalmente a las múltiples vulnerabilidades que ofrecen estos sistemas. El nivel de precisión en la respuesta y las tasas de error del sistema están muy relacionados con la seguridad del mismo.
- **Usabilidad:** no hay que perder de vista que la enseñanza es un servicio que se presta a usuarios (estudiantes). La calidad del servicio no se debe ver afectada por el sistema de autenticación y el usuario debe sentirse cómodo utilizando el servicio.
- **Legalidad:** aunque muchas veces se infravaloran los aspectos legales relacionados con estos servicios, la cambiante legislación y la falta de acuerdos internacionales respecto a internet la convierten en un aspecto crítico que muchas veces dificulta la implantación de nuevas tecnologías.

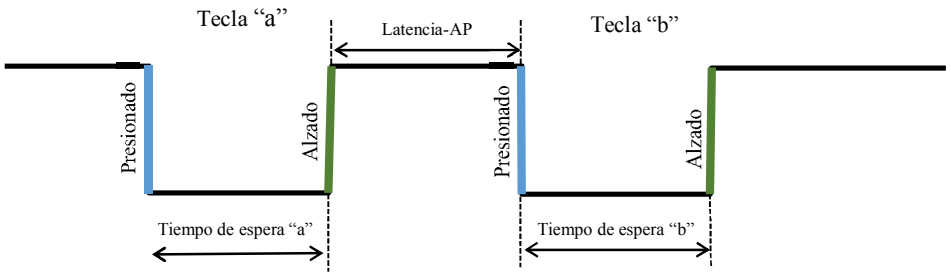

Figura 2: Características más comunes utilizadas en la autenticación por dinámica de tecleo

El sistema de reconocimiento ideal debería tener en cuenta estos tres aspectos y proponer una solución que sea capaz de ofrecer la máxima seguridad posible, sin que ello conlleve una mala percepción por parte de los usuarios ni los estamentos legislativos. En este trabajo se analiza la autenticación biométrica basada en dinámica de tecleo para su utilización en entornos de enseñanza en línea. Se propone una arquitectura tradicional basada en un servidor de autenticación en la nube, ver Figura 1. Veamos ahora como se ajusta esta tecnología a los aspectos críticos antes señalados:

- **Seguridad de un sistema de reconocimiento basado en dinámica de tecleo:** frente a la facilidad de robar una clave o un password, los sistemas biométricos ofrecen un nivel más de seguridad al ser preciso no solo conocer la clave, sino la forma en la que fue introducida (dinámica de tecleo). Esto supone un segundo nivel de autenticación capaz de detectar intrusos que se han hecho con la clave de otro usuario. Aun así, hay que tener en cuenta que los sistemas biométricos basados en dinámica de tecleo ofrecen tasas de reconocimiento moderadas (errores por encima del 10%) alejadas de sistemas más populares como la huella dactilar.
- **Usabilidad de un sistema biométrico basado en dinámica de tecleo:** a diferencia de otros rasgos como la huella dactilar, la cara o el ADN, la forma en la que tecleamos se puede considerar un rasgo poco invasivo. Es además un rasgo que puede ser reconocido de forma transparente mientras el usuario realiza tareas habituales relacionadas con el tecleo (introducción de datos, realización de pruebas, etc…).
- **Legalidad:** la ley de Protección de Datos establece requerimientos relacionados con el almacenado de información sensible entre la que se puede encuadrar los patrones biométricos. Importante destacar la existencia de estándares internacionales como ISO/IEC 19794-2 y ANSI/INCITS 378 que otorgan credibilidad y garantizan mínimos.

El resto del trabajo se centrará en el sistema de autenticación de usuarios a partir de su dinámica de tecleo.

3. DESCRIPCIÓN

En este trabajo se analiza la tecnología de reconocimiento biométrico basada en dinámica de tecleo (*kesytroke dynamics*) en el ámbito del reconocimiento de estudiantes para plataformas online. Para llevar a cabo el estudio utilizará una base de datos pública así como cuatro sistemas de reconocimiento del estado del arte.

A. Extracción de características discriminantes basadas en la dinámica de tecleo

Las dinámica de tecleo de un usuario viene definida principalmente por dos tipos de eventos: presionado de tecla (en inglés *key-press*) o alzamiento de tecla (en inglés *key-release*). Las características más comunes utilizadas para reconocer a personas a través de su dinámica de tecleo son relaciones entre las marcas de tiempo de ambos eventos. Supongamos que se quiere modelar la dinámica de pulsación de un usuario a través de N pulsaciones de teclado. El vector de características $\mathbf{t}$ que modela la dinámica de pulsaciones del usuario contendrá las marcas de tiempo de N eventos correspondientes a $\mathbf{t}^p$ marcas para los presionados de tecla (momento en el que la tecla es presionada) y $\mathbf{t}^a$ marcas para los alzamientos de tecla (momento en el que tecla se deja de presionar). Las relaciones entre estas marcas de tiempo nos darán la dinámica de tecleo de los usuarios, pero es necesario para ello realizar algún tipo de normalización respecto una referencia. Esta normalización en el tiempo se consigue considerando diferencias de tiempo entre teclas consecutivas en lugar de marcas de tiempo absolutas, ver Figura 2. Algunas de las características más populares son:

- **Tiempo de Espera (o *Hold Time*):** es la diferencia entre el tiempo de presionado y el de liberación de la tecla i-ésima:

$$h_i = t_i^a - t_i^p \qquad i = 1, \dots, N$$

- **Latencia Alzado-Presionado (o *Release-Press latency*):** es la diferencia entre el tiempo de presionado de la tecla $(i+1)$-ésima y el tiempo de alzado de la tecla i-ésima:

$$l_i^{ap} = t_{i+1}^p - t_i^a \qquad i = 1, \dots, N-1$$

Estas características se utilizarán para modelar/caracterizar la dinámica de pulsación de cada uno de los estudiantes que vendrá definida por la concatenación de los vectores de tiempo de espera y Latencia $\{\mathbf{h}, \mathbf{l}^{ap}\}$.

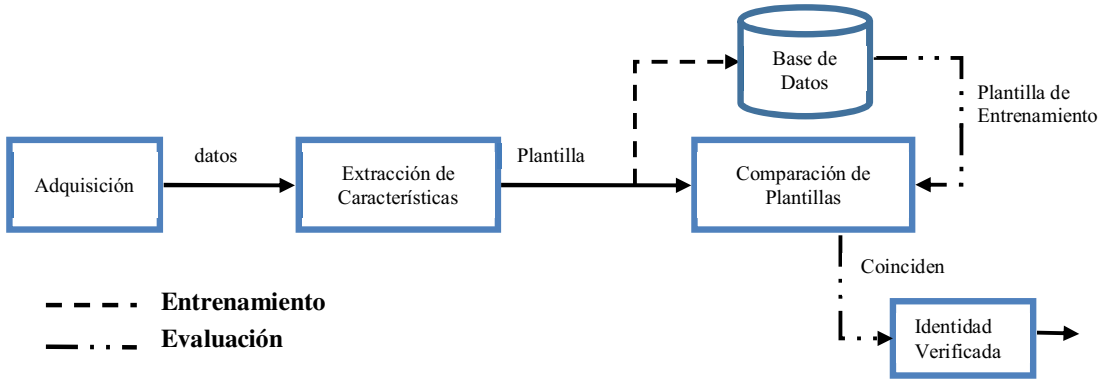

Figura 3: Diagrama de bloques de un sistema de reconocimiento biométrico

B. Modelado de la dinámica de tecleo

El modelado de los usuarios se hace a través de pequeñas cadenas de caracteres conocidas como dígrafos y trígrafos según tengan 2 o 3 caracteres respectivamente (Peacock et al., 2004). Para cada uno de los dígrafos se construye un vector de características que incluya los dos Tiempos de Espera y la Latencia Alzado-Presionado. En el caso de los trígrafos se añade un tercer Tiempo de Espera y una segunda Latencia Alzado-Presionado. La dinámica de pulsación de teclas de un estudiante vendrá definida por estos vectores de características o plantillas.

Cuando se quiere contrastar la identidad de una nueva secuencia de caracteres, se compara la plantilla almacenada de los estudiantes (plantilla de entrenamiento) con la plantilla obtenida de la nueva secuencia, ver Figura 3. En este trabajo se analizan dos sistemas de reconocimiento de dinámicas de tecleo: Distancia de Mahalanobis + Vecino Más Cercano (NN por sus siglas en inglés) y Distancia de Manhattan Modificada. Asuma $\mathbf{f} = \{f_1, f_2, \dots, f_M\}$ una plantilla de características (con M características) de una muestra de test y $\mathbf{g}^k = \{g_1^k, g_2^k, \dots, g_M^k\}\ k \in 1, \dots T$ un conjunto de muestras de entrenamiento con T plantillas de un mismo usuario. Las dos distancias estudiadas en este trabajo se calculan de la siguiente forma:

- **Mahalannobis + NN:** este clasificador fue propuesto en (Cho, Han y Kim, 2000). La distancia entre una plantilla de test $\mathbf{f}$ y cada una de las plantillas de entrenamiento $\{\mathbf{g}^k\}_{k=1}^T$ se calcula como:

$$d_1{}^k = (\mathbf{f} - \mathbf{g}^k)\mathbf{S}^{-1}(\mathbf{f} - \mathbf{g}^k)^{\mathrm{T}} \qquad (1)$$

  donde se introduce la matriz de covarianza $\mathbf{S}$ para ponderar aquellas características con menor varianza y donde $(\cdot)^{\mathrm{T}}$ es la traspuesta. La distancia final d_1 se obtiene como la mínima distancia para todo k.

- **Distancia de Manhattan Modificada:** esta distancia es una modificación de la distancia de Manhattan normalizada propuesta en (Araujo, Sucupira, Lizarraga, Ling y Yabuuti, 2005). La distancia entre una plantilla de test $\mathbf{f}$ y las plantillas de entrenamiento $\{\mathbf{g}^k\}_{k=1}^T$ se calcula como:

$$d_2 = \sum_{i=1}^{M} |f_i - \bar{g}_i| / \sigma_i' \qquad (2)$$

  donde $\boldsymbol{\sigma}' = [\sigma_1', \sigma_2', \dots, \sigma_M']$ es la desviación estándar modificada:

$$\sigma_i' = \begin{cases} \frac{0.2}{M}\sum_{j=1}^{M}\sigma_j & si\ \ \sigma_i < \frac{0.2}{M}\sum_{j=1}^{M}\sigma_j \\ \sigma_i & resto \end{cases} \qquad (3)$$

  Y σ_i es la desviación estándar calculado sobre el conjunto de entrenamiento $\{\mathbf{g}^k\}_{k=1}^T$. Esta simple modificación trata de mitigar los efectos de muestras con muy poca variabilidad.

Ambas distancias se han seleccionado por sus altos rendimientos en el entorno de validación público CMU (Killourhy y Maxion 2009). Destacan sus prestaciones incluso cuando son comparadas con algoritmos de aprendizaje automático más complejos, ver Tabla 1. Como puede apreciarse, el rendimiento de ambas distancias es similar con errores entre 8.84% y 9.96% (EER o Tasa de Igual Error promedio de los usuarios).

C. Base de datos

Los experimentos incluidos en este trabajo se realizan sobre la base de datos OhKBIC (Monaco et al., 2015). La base de datos incluye la respuesta de 64 estudiantes a 3 exámenes desarrollados a través de internet (utilizando la plataforma educativa Moodle de la asignatura). Para formar la base de datos se almacenó la dinámica de pulsación de al menos 1500 caracteres de cada alumno (500 por cada examen), lo cual se corresponde a tres párrafos de extensión moderada. Las características de la base de datos se pueden resumir en

Tabla 1. Rendimiento (EER) de las dos distancias analizadas en el entorno de validación CMU (Killourhy y Maxion, 2009)

Clasificadores	EER promedio
Mahalannobis + Nearest Neighbor (d_1)	**0.0996**
Modified Scaled Manhattan distance (d_2)	**0.0884**
z-score	0.1022
SVM	0.1025
Mahalanobis	0.1101
Mahalanobis norm.	0.1101

- **Escenario de autenticación independiente de texto:** las muestras correspondientes a cada uno de los tres exámenes pertenecen a diferentes preguntas con diferentes respuestas que dependen de cada alumno.
- **Experimento multisesión:** existe un lapso temporal (aproximadamente dos meses) entre cada una de las tres muestras adquiridas por cada estudiante.

D. Metodología de Experimentación

El objetivo del experimento es conocer el rendimiento de los sistemas de autenticación biométrica basados en la dinámica de tecleo de los estudiantes. Para ello se diseñó un protocolo de experimentación dividido en las siguientes 4 fases:

Fase 1: Para cada usuario, se divide su patrón de tecleo en una plantilla de entrenamiento/modelado (formado por los 500 caracteres disponibles del primer examen) y un conjunto de 203 plantillas de evaluación (con 500 caracteres cada una correspondientes a las respuestas del segundo y tercer examen).

Fase 2: Se buscan dígrafos (cadenas de dos caracteres) y trígrafos (cadenas de tres caracteres) coincidentes entre las plantillas de entrenamiento y evaluación. Al ser un escenario de autenticación independiente de texto, es necesario buscar este tipo de cadenas comunes.

Fase 3: Se aplica un protocolo de evaluación cruzada (todos los dígrafos y trigrafos de todos los usuarios se cruzan) para calcular las distancias entre las plantillas de entrenamiento y verificación de cada usuario. Para calcular las distancias entre plantillas se utilizan los sistemas expuestos en el apartado 3.B. El objetivo es reconocer a los alumnos de las plantillas de evaluación a partir de sus modelos o plantillas de entrenamiento.

Fase 4: El rendimiento del sistema se mide en base a la Falsa Aceptación, el Falso Rechazo y la Tasa de Igual Error por usuario. Las métricas de rendimiento finales son promedios de las métricas obtenidas para todos los usuarios.

4. RESULTADOS

Se realizan dos tipos de comparaciones según las plantillas de evaluación y de entrenamientos utilizadas:

- **Comparación genuina:** las plantillas de evaluación y de entrenamiento pertenecen al mismo usuario. Esta comparación simula el acceso de un usuario normal que intenta acceder al servicio de forma regular.
- **Comparación impostora:** las plantillas de evaluación y de entrenamiento pertenecen a usuarios diferentes. Esta comparación simula la suplantación de identidad de un usuario que intenta acceder al servicio haciéndose pasar por otro.

La Figura 4 muestra las Distribuciones de Densidades de Probabilidad de los 4 experimentos propuestos. Se observa el alto grado de solapamiento entre muestras genuinas e impostoras lo que da una idea de la difícil tarea de separación de clases (genuinas e impostoras). Aun así, se observan diferentes comportamientos y se intuye un mejor comportamiento de los trígrafos respecto a los dígrafos así como de la distancia modificada de Manhattan respecto a la de Mahalanobis +NN.

La Tabla 2 muestra el rendimiento de los 4 sistemas en términos de Tasa de Igual Error (EER). Los resultados muestran tasas de igual error desde el 9.05% (distancia de Mahalanobis + NN y trígrafos) hasta el 25.12% (distancia modificada de Manhattan y trígrafos). El mejor comportamiento de los trígrafos respecto a los dígrafos se puede deber a la mayor cantidad de información y patrones más distintivos de los primeros. La distancia de Manhattan modificada muestra el mejor rendimiento debido fundamentalmente al buen comportamiento de ésta cuando se dispone de pocas muestras de entrenamiento con las que poder ponderar el peso de cada característica (algo fundamental en la normalización de Mahalanobis).

Tabla 2. Rendimiento (EER) de los cuatro sistemas estudiados usando la base de datos OhKBIC

Clasificadores	EER promedio	
	Dígrafo	Trígrafo
Mahalanobis + Nearest Neighbour (d_1)	0.1543	0.1409
Distancia Manhattan Modificada (d_2)	0.2512	0.0905

5. CONCLUSIONES

En este trabajo se ha realizado un análisis de la potencial aplicación de tecnologías de reconocimiento biométrico para la autenticación web en entornos docentes. Se ha abarcado todo el proceso desde el diseño hasta el desarrollo y experimentación incluyendo un caso de estudio con datos de 64 estudiantes. Los resultados obtenidos muestran tasas de reconocimiento por encima del 90% lo que anima a seguir investigando en esta línea.

Trabajos futuros deberán incluir una evaluación más extensa con mayor número de estudiantes. También se propone trabajar con modelos de entrenamiento más potentes que permitan incorporar la mayor cantidad de información posible a esta etapa fundamental.

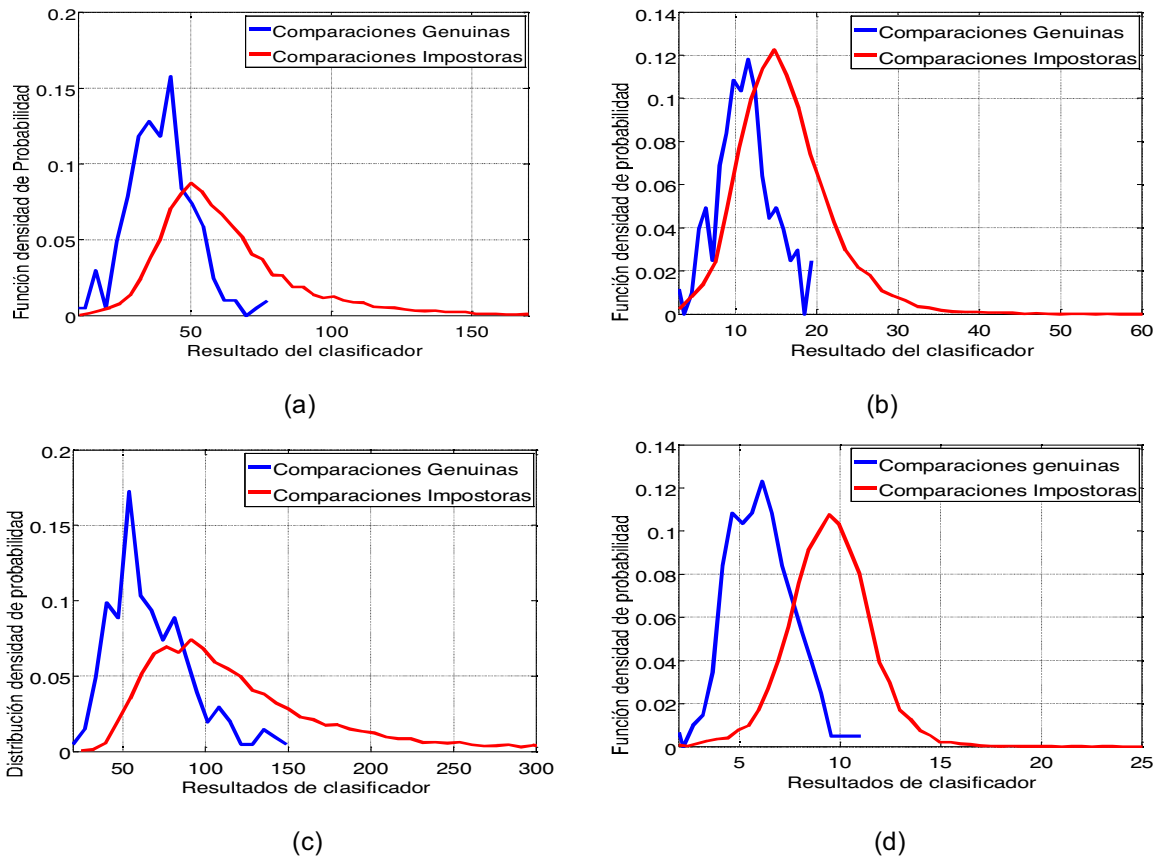

Figura 4: Distribuciones de Densidades de Probabilidad para: Mahalanobis + NN y dígrafos (a); Distancia de Manhattan Modificada y dígrafos (b); Mahalanobis + NN y trígrafos (c); Distancia de Manhattan Modificada y trígrafos (d);

AGRADECIMIENTOS

Aythami Morales está financiado por el MECD español a partir de un contrato Juan de la Cierva (JCI-2012-12357). Este trabajo ha sido parcialmente financiado a través de los proyectos: Bio-Shield (TEC2012-34881) from Spanish MINECO, BEAT (FP7-SEC-284989) from EU, CECABANK and Cátedra UAM Telefónica.

REFERENCIAS

Araujo, L. C. F., Sucupira, L. H. R., Lizarraga, M. G., Ling, L. L., Yabuuti, J. B. T., (2005). User Authentication Through Typing Biometrics Features. *IEEE Trans. On Signal Processing*, 53(2), pp. 851-855.

Cho, S., Han, C., Han, D. H., and Kim, H., (2000). Web-based keystroke dynamics identity verification using neural network. *Journal of Organizational Computing and Electronic Commerce*, 10(4), pp. 295–307.

Gunetti D., and Picardi, C., (2005). Keystroke analysis of free text. *ACM Transactions on Information and System Security*, 8(3), pp. 312–347.

Killourhy, K. S., and Maxion, R. A., (2009). Comparing Anomaly Detectors for Keystroke Dynamics. *In Proceedings of the 39th Annual Int. Conf. on Dependable Systems and Networks (DSN-2009)*, Estoril, Lisbon, Portugal. IEEE Computer Society Press, Los Alamitos, California, pp. 125- 134.

Miguel, J., Caballe, S., and Prieto, J., (2013). Providing information security to MOOC: Towards effective student authentication. *In Proc. of the Int. Conf. on Intelligent Networking and Collaborative Systems* (Xian, China), IEEE Press, pp. 289–292.

Monaco, J. V., Perez, G., Tappert, C. C., Bours, P., Mondal, S., Rajkumar, S., Morales, A., Fierrez J., and Ortega-Garcia, J., (2015). One-handed Keystroke Biometric Identification Competition. *In Proc. IEEE/IAPR Int. Conf. on Biometrics (ICB15)*, Phuket (Thailand), May 2015, pp. 1-7.

Morales, A., Fierrez, J., and Ortega-Garcia, J., (2014). Towards predicting good users for biometric recognition based on keystroke dynamics. *In Proc. of European Conf. on Computer Vision Workshops*, Springer LNCS-8926, Zurich, Switzerland, pp. 711-724.

Peacock, A., Ke, X., Wilkerson, M., (2004). Typing patterns: A key to user identification. *IEEE Security and Privacy*, 2(5), pp. 40–47.

Stanford, Education's digital future (2013). Coursera announces details for selling certificates and verifying identities. Available at: http://edf.stanford.edu/readings/coursera-announces-details-selling-certificates-and-verifying-identities.