Generación de Documentos Nacionales de Identidad sintéticos mediante el uso de perfiles biográficos.

Victoria Amores Chaparro
Universidad de Extremadura
Escuela Politécnica de Cáceres
kamoresc@alumnos.unex.es

Fernando Broncano Morgado
Universidad de Extremadura
Escuela Politécnica de Cáceres
fbroncan@alumnos.unex.es

Álvaro Hernández Martín
Mobbeel Solutions S.L.
C/ Santa Cristina S/N, 10195
ahernandez@mobbeel.com

Óscar Mogollón Gutiérrez
Universidad de Extremadura
Escuela Politécnica de Cáceres
oscarmg@unex.es

José Carlos Sancho Núñez
Universidad de Extremadura
Centro Universitario de Mérida
jcsanchon@unex.es

Sebastião Pais
University of Beira Interior
Department of Computer Science
sebastiao@di.ubi.pt

***Resumen*—La generación de conjuntos de datos con información personal realista es compleja, aunque son múltiples sus ventajas, como la posibilidad de entrenar modelos para identificar manipulaciones. Este trabajo desarrolla un conjunto de técnicas avanzadas para generar perfiles biográficos sintéticos, tomando como base el Documento Nacional de Identidad de España (DNI). El perfil biográfico contiene una fotografía, datos personales y firma. Entre las técnicas, se plantea una metodología para la evaluación con métricas de imágenes de rostros, calificando en una escala (0-100) y verificando la validez de la fotografía con respecto a los estándares establecidos y estimando el sexo y la edad de la fotografía seleccionada. El resto de datos biográficos (nombre, apellidos, dirección, etc.) se genera con ayuda de datos obtenidos del Instituto Nacional de Estadística. La automatización del sistema permite obtener como resultado un amplio conjunto de datos completos y realistas sin comprometer la privacidad de los individuos.**

***Palabras clave*—Identidad digital, Privacidad, Seguridad, Generación sintética.**

Tipo de contribución: *Investigación en desarrollo*

I. Introducción

La protección de la identidad y de los datos personales se ha convertido en un asunto de suma importancia en la era digital. A diario, se producen multitud de filtraciones de datos personales que, en muchos casos, terminan en manos de ciberdelincuentes. Estas filtraciones terminan en abundantes ocasiones con la suplantación o usurpación de la identidad de una persona. Este delito consiste en apropiarse de forma ilegal de los datos personales de otra persona con fines fraudulentos. De esta forma, los ciberdelincuentes pueden acceder a trámites digitales como la creación de cuentas bancarias, realizar compras fraudulentas, abrir líneas de crédito y cometer otros delitos financieros en nombre de la víctima. Las consecuencias derivadas de sufrir una suplantación de identidad son graves desde el punto de vista económico, legal y emocional debido a los trastornos y preocupaciones que generan en los usuarios.

La usurpación de identidad digital y el fraude cibernético ha generado nuevas oportunidades y desafíos tecnológicos para combatir contra ellos. En los últimos años, se ha trabajado en propuestas y buenas prácticas que mejoren la autenticación, privacidad y anonimización de los usuarios. Para ello, algunas investigaciones [1][2] se han centrado en mejorar los procesos de autenticación con elementos biométricos a través del reconocimiento facial, de huellas dactilares, de iris, o de voz, que permiten autenticar de manera precisa la identidad de los usuarios. Otros como Liu et al. [3] o Nusantoro et al. [4] se han centrado en explorar el uso de tecnologías como blockchain para crear sistemas de identidad digital resistentes a la manipulación.

En España, la gestión de la identidad de una persona se lleva a cabo a través de varios mecanismos como son:

1. El Documento Nacional de Identidad (DNI) emitido por la Dirección General de la Policía con datos personales como nombre, apellidos, fecha y lugar de nacimiento, fotografía y firma manuscrita.
2. El DNI electrónico (DNIe) que permite la identificación digital y la firma electrónica de documentos.
3. Los certificados digitales emitidos por autoridades de certificación reconocidas, como la Fábrica Nacional de Moneda y Timbre (FNMT).

Aunque tanto el DNIe como los certificados digitales han proliferado en su utilización, el más utilizado hasta el momento es el DNI físico. Sin embargo, el uso del DNI físico en procesos de identificación virtual presenta una serie de problemáticas y carencias de seguridad que facilitan la suplantación de la identidad de los usuarios. Dicho documento nacional dispone de elementos de seguridad que verifican la autenticidad del documento y que solo se perciben con el contacto físico como son los grabados en relieve, los hologramas, los elementos ópticos, la tinta ultravioleta, etc. Estos elementos físicos son muy complicados de replicar con precisión y se consideran la clave de la protección del usuario. No obstante, pese a su seguridad, está claro que un riesgo inherente del DNI se produce cuando se utiliza en procesos virtuales, aumentando así la posibilidad de suplantación, falsificación o manipulación.

En la actualidad, es complejo investigar en técnicas de detección de patrones en usurpaciones de identidad debido a la falta de conjuntos de datos realistas con los que poder entrenar nuevos modelos por la sensibilidad que conlleva trabajar con datos personales realistas. Estas dificultades o hándicaps hacen que la investigación en este campo y en sus avances sea prometedora.

Para contribuir a ampliar las posibles investigaciones, este

trabajo se centra en el análisis de los elementos que componen un perfil biográfico y, tras ello, en la creación de técnicas avanzadas para la generación de perfiles biográficos sintéticos realistas, que toman como base los datos que aparecen en el DNI de España. Así, las contribuciones de este trabajo son las siguientes:

- El análisis de los elementos que componen el perfil biográfico de una persona y la importancia de su protección.
- La creación de una metodología de generación sintética de la fotografía del DNI. La propuesta evalúa en una escala de 0 a 100 la idoneidad de la fotografía con respecto el cumplimiento de los protocolos y estándares establecidos para que una foto sea válida para el DNI.
- Un proceso detallado de incrustación de datos personales simulados, extraídos de bases de datos reales, en los perfiles biográficos sintéticos que, a su vez, se complementa con la generación de firmas con diferentes trazos y grosores.

El objetivo de este trabajo es contribuir a la creación de una combinación de técnicas innovadoras para generar de forma automatizada conjuntos de datos completos y realistas, sin comprometer la privacidad de los individuos.

II. Trabajos relacionados

Actualmente, conseguir un conjunto de datos claramente definido y establecido es complejo, más si cabe, aquellos conjuntos de datos construidos por datos personales y documentos de identificación. Alguno de los conjuntos de datos que existen actualmente sobre documentos de identidad son MIDV-500 [5], MIDV-2019 [6] y MIDV-2020 [7]. En estos trabajos, se presentan diferentes conjuntos de datos de vídeos de documentos de identidad móvil, con un número de muestras diferente, aunque todos contienen muestras de diferentes países, en diferentes posiciones y entornos.

Otros autores exploran el campo de la construcción y la generación de documentos de identificación. Soares et al. [8] desarrollan un proceso de creación de documentos de identidad brasileños a partir de documentos reales y datos generados artificialmente. Benalcázar et al. [9] diseñaron un sistema de generación de documentos de identidad de Chile utilizando inteligencia artificial generativa, así como réplica de fraudes a través de impresiones o fotografía de pantallas utilizando documentos de identidad verídicos y transferencia de texturas de moiré o papel. Bothra et al. [10] optaron por un flujo de trabajo de generación aleatoria de elementos de identificación sobre documentos indios de todo tipo y un proceso de posprocesado completo incluyendo aplicación de perspectiva y colocación del documento generado en una escena. Attivissimo et al. [11] desarrollaron un sistema de lectura de varios tipos de documentos italianos que permite detectar varios tipos de documentos (pasaportes, carnets de conducir, etc.) y leer sus elementos, partiendo de un conjunto de documentos sintéticos generados por el mismo equipo a partir de plantillas. Zhao et al. [12] proponen una arquitectura que separa documentos de identidad en tres secciones: texto, imagen y fondo; siendo cada una procesada por una red neuronal específica, para después crear el documento falsificado completo. Sin embargo, por el momento estos autores no encuentran en la literatura conjuntos con datos personales de España.

III. El Perfil Biográfico

En el campo de la identificación y acreditación personal, un elemento de identificación se define como una característica que reconoce a un individuo y lo distingue de otro. En este sentido, los elementos de identificación van a permitir determinar ciertos atributos personales sobre un individuo de referencia. Y a modo de abstracción, se va a definir el perfil biográfico como el conjunto de elementos de identificación que permitan determinar a un único individuo y distinguirlo de otros.

Cuando se quiere identificar a una persona de forma inequívoca, se puede construir un perfil biográfico que englobe las diferentes características acerca de un individuo, y este conjunto tendría que contener un mínimo de variables que permitan distinguir un sujeto de cualquier otro. A este conjunto se le denomina perfil biográfico completo. Si el conjunto no pretende identificar de forma inequívoca a una persona, podría tratarse de un perfil biográfico parcial, puesto que contiene algunos elementos de identificación, pero no suficientes para garantizar su unicidad. Para identificar a una persona, o grupo de ellas, a través de un perfil biográfico, se debe construir una función que determine el individuo, o conjunto de ellos, en función de los elementos de identificación proporcionados.

Siendo P_B el conjunto del perfil biográfico y $x, y, ..., z$ los elementos de identificación presentes en el conjunto, se puede definir una función sobre el conjunto de elementos de identificación denominada $id(x, y, ..., z)$ que determine un individuo o conjunto de ellos.

$$P_B = \{x, y, ..., z\} \tag{1}$$

La función $id(x, y, ..., z)$ debe corresponder cualquier persona con un perfil biográfico, por lo que esta función necesariamente debe ser sobreyectiva, ya que todos los sujetos a identificar deben disponer de todos elementos de identificación que forman el perfil biográfico.

$$id(x, y, ..., z) = p \tag{2}$$

Para el caso específico del Documento Nacional de Identidad (DNI) de España, se debe construir un perfil biográfico completo en base de diferentes elementos de identificación. El DNI contiene un perfil biográfico que no solo identifica de forma inequívoca al sujeto que presenta, si no, que también identifica el soporte físico sobre el que se representa. Dentro de los elementos de identificación, se debe discernir entre aquellas características personales que son intrínsecas a las personas, que serán elementos de identificación personales, y las características que identifican al soporte físico sobre el que se representa los elementos de identificación personales.

El perfil biográfico para el DNI estará formado por los elementos de identificación personales y por los elementos de identificación del soporte. Por definición de los estándares de este documento, el perfil biográfico obtenido es completo, ya que, entre otros, se almacenan datos únicos para cada sujeto, como algunos de los atributos correspondientes al soporte y otros personales que se consideran únicos, como la digitalización de las huellas dactilares correspondientes a los dedos índices izquierdo y derecho, una fotografía del rostro, o la firma del sujeto representado.

A. *Elementos de identificación personales.*

Los elementos de identificación personales pretenden ser los datos personales del individuo representado en el documento. Comúnmente, estos datos tienen un carácter sensible visto que identifican personalmente, geográficamente e históricamente al sujeto identificado en el DNI.

Los nombres y apellidos son los datos más comunes de un DNI. Tradicionalmente, los nombres han sido útiles para referirse a las personas, mientras que los apellidos tenían la función de distinguirla de aquellas que tienen el mismo nombre y conocer la descendencia y procedencia familiar del individuo. Estos dos elementos de identificación personales no constituyen una identificación única, ya que el Estado español permite que dos personas puedan tener mismo nombre y apellidos que otro ciudadano del país.

La digitalización de una fotografía de la persona también se incluye en el DNI, y se muestra impresa junto al resto de datos. Esta permite comprobar si el rostro actual de una persona coincide con el individuo que posee el soporte, pero no es suficiente para la identificación inequívoca de una persona. Otro de los aspectos que se incluye, y van muy ligados a la fotografía, son el sexo de la persona y la fecha de nacimiento. Estos, por lo general, van ligados a la fotografía a causa de que visualizando una fotografía puede ser suficiente para estimar una edad y el género de la persona.

A los ciudadanos españoles, cuando obtienen la nacionalidad española, se les asigna un número de DNI, y este número, junto con la nacionalidad española, no suele variar a lo largo de la vida del sujeto. La nacionalidad es un dato complementario al perfil biográfico, pero que nunca puede determinar por sí misma un individuo concreto, si no, un conjunto de ellos. Por otro lado, debido a varios problemas existentes de la expedición de DNI que pueden verse en el estudio de García del Vello [14], el número de DNI tampoco se puede asegurar que sea único. Aunque dicho número del DNI deber ser personal, intransferible y perpetuo, por lo que dos personas no deberían tener nunca el mismo número de DNI.

El DNI también contiene el domicilio en el que se encuentra empadronado el sujeto, así como su lugar de nacimiento, en función a lo acordado en el registro civil. El domicilio contiene dirección, número e identificación adicional de la vivienda, municipio y provincia. Por otra parte, el lugar de nacimiento se conforma de un municipio y la provincia a la que pertenece. Estos datos tampoco identifican de forma inequívoca al sujeto, ya que en un domicilio pueden estar empadronados varias personas, y a partir del lugar de nacimiento tampoco se garantiza la exclusividad, ya que en una misma ciudad nace un conjunto de personas y no una sola.

El nombre de los progenitores y la firma también son datos correspondientes al perfil biográfico que soporta el DNI. La firma suele ser única, pero no se puede demostrar la unicidad de la misma. Y a un par de nombres, puede haber un conjunto de personas cuyos progenitores tengan dichos nombres.

El último elemento, aunque no visible, que figura entre los datos del Documento Nacional de Identidad son las huellas dactilares. Una digitalización de las huellas dactilares se guarda en el chip del DNI, que sirve para garantizar la originalidad del documento.

Tabla I: Elementos de identificación del perfil biográfico del DNI.

Elemento de identificación	Personal	Soporte
Número de DNI	X	
Nombre	X	
Apellidos	X	
Sexo	X	
Nacionalidad	X	
Fotografía	X	
Fecha de emisión		X
Fecha de validez		X
Fecha de nacimiento	X	
Número de soporte		X
Firma	X	
Domicilio	X	
Lugar de nacimiento	X	
Progenitores	X	
Equipo de expedición		X
Digitalización de huellas dactilares	X	

Este conjunto de elementos de identificación personales pueden no garantizar la biyectividad de la función de identificación, por lo que, se deben sumar los elementos de identificación del soporte como características adicionales al perfil biográfico que se define para el documento nacional de identidad.

B. *Elementos de identificación del soporte.*

Los elementos de identificación del soporte tienen el objetivo de aportar la unicidad al perfil biográfico del DNI junto con la función de identificar de forma inequívoca al soporte que contiene los elementos de identificación personales del perfil biográfico.

Este tipo de elementos lo conforman varias fechas relevantes para la validez del soporte y otras cadenas alfanuméricas que tienen relación con el soporte y quién lo expide. El número del soporte es una cadena con tres letras seguida de seis números, que identifica de forma inequívoca el soporte, y por ende, el perfil biográfico asociado a dicho soporte. Se garantiza que esta cadena alfanumérica no se repite entre ninguno de los DNI de España. También, se incluye el equipo de expedición, una cadena alfanumérica de ocho caracteres que identifica al equipo de Policía Nacional que ha expedido el DNI. En el DNI, la fecha en la que se expide el documento es relevante y forma parte del perfil biográfico, aunque esta no determina unicidad, ya que Policía Nacional emite DNI a diferentes sujetos en misma fecha. La fecha de expedición marca la fecha de validez en función de la edad que tiene el sujeto en el momento de la emisión del DNI.

A modo de clasificar los diferentes elementos de identificación que compone un perfil biográfico asociado a un DNI, se puede observar la Tabla I.

IV. Forma y elementos de seguridad del DNI

El Documento Nacional de Identidad (DNI) es la credencial que emite el Estado español para identificar a una persona de la misma nacionalidad. Consiste en una tarjeta de policarbonato de 85,60 mm de ancho y 53,98 mm de alto, que se denomina soporte. En él se encuentran diversos elementos de identificación tales como la fotografía de la persona y una serie de datos personales y de soporte. A efectos prácticos cumple los estándares relativos a los documentos de viaje

de tipo 1 marcados por la Organización de Aviación Civil Internacional [13].

Asimismo, el DNI consta de varios métodos de seguridad grabados físicamente que aseguran su autenticidad y son difíciles de replicar por máquinas no autorizadas. En el siguiente listado se explican brevemente los diferentes componentes de seguridad del DNI:

- *Kinegrama*. Impresión de difracción microscópica que, al mover el ángulo de incidencia de la fuente de luz y de observación, permite visualizar el escudo con diferentes coloraciones y animaciones.
- *Tinta ópticamente variable*. Tinta compuesta por finas partículas que otorgan el efecto óptico de variabilidad del color cuando se cambia el ángulo de incidencia de la luz y de observación.
- *Ventana transparente con grabado láser*. Es una zona del soporte transparente que tiene grabado el número de soporte.
- *«Changeable Laser Image» (CLI) o imagen láser cambiante en bajorrelieve*. Versión holográfica pequeña y en relieve de la foto de la persona grabada con láser.
- *Embozado efecto mate*. Dos franjas oblicuas con impresión ligeramente distinta.
- *Embozado en alto relieve*. Formas geométricas impresas en relieve en el documento.
- *Tintas ultravioleta en iris*. Tintas solo visibles mediante luz ultravioleta.
- *Tinta Oasis*. Tinta opaca que oculta las capas de debajo que, al verla a través de un filtro polarizado, deja ver dichas capas revelando el patrón de seguridad.
- *Fondos offset de seguridad*. Da forma al fondo de la imagen.
- *Microtexto*. Secuencias de caracteres impresos a tan bajo tamaño que los hace muy difíciles de fotocopiar.
- *Chip*. Un chip de interfaz dual y un chip NFC, ambos con certificados electrónicos.

V. Generación sintética de un perfil biográfico

La generación sintética de información es un recurso muy empleado cuando no se dispone de grandes volúmenes de ella asociados a un ámbito concreto, y los perfiles biográficos de un Documento Nacional de Identidad es uno de estos casos. Para generar los diferentes elementos de identificación asociados al perfil biográfico del DNI, es necesario conocer la forma que deben tener. Los elementos a generar son listados en la Tabla I, excepto la digitalización de las huellas dactilares, que no es un elemento observable en una fotografía delantera o trasera, solo como información relevante del chip.

La generación del perfil se va a realizar en cuatro iteraciones. En la primera se obtiene la fotografía, y, el sexo y la edad a partir del rostro del sujeto. En la segunda iteración, partiendo de los datos de sexo y edad, se obtienen el resto de elementos de identificación, excepto la firma manuscrita. Con todos los elementos de identificación, se genera la zona de lectura automática, que tiene la función de resumir los datos más relevantes del DNI. Por último, se genera la firma manuscrita.

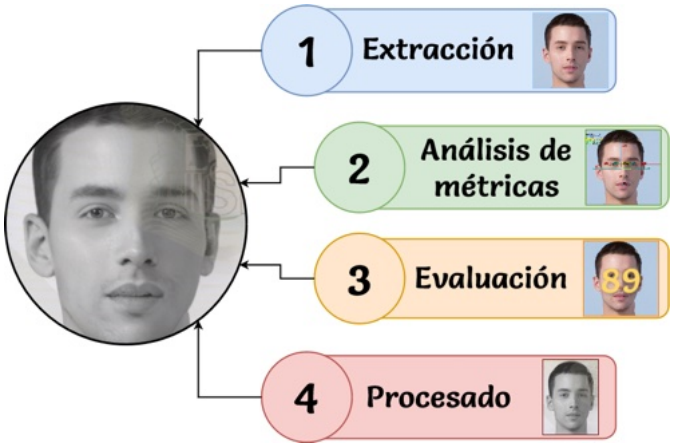


Figura 1: Proceso de obtención y evaluación de fotografía.

A. *Obtención de la fotografía, sexo y edad*

El primer elemento a obtener para formar un perfil biográfico del DNI es la fotografía del sujeto a representar. La fotografía, siguiendo las directrices del Real Decreto 1553/2005 [15], debe tener la relación 32:26, con el rostro a color, fondo uniforme blanco y liso, tomada de frente con la cabeza totalmente descubierta y sin gafas de cristales oscuros u objetos que dificulten la identificación de la persona.

Hay varias herramientas de redes adversativas generativas que son capaces de devolver imágenes de rostros que no existen. Se usará uno de esos servicios [16], aunque no todos los rostros generados por este servicio pueden ser empleados como fotografía del perfil biográfico, ya que deben estar en un plano frontal en la imagen. Las imágenes consideradas válidas en función de las características faciales se escalan a relación 32:26 y se les elimina el fondo, quedando a la persona con un fondo blanco uniforme.

El proceso para tratamiento de una imagen pretende adaptar el rostro generado por el servicio de la red adversativa generativa que devuelve imágenes de rostros sintéticas. Las imágenes generadas pueden ser de rostro completo o que tengan partes de la cara ocultas, todas disponen de un fondo y no siempre se obtienen imágenes de frente, sin gafas, con los ojos abiertos y que tengan el rostro completo descubierto. Para ello, se presenta un sistema de medición de la calidad de los rostros generados y poder conservar aquellos más adecuados, desechando los que no cumplen las condiciones mínimas recogidas en [15]. La Figura 1 resume en cuatro pasos el proceso de generación y evaluación de las fotografías: extracción (A1), análisis (A2), evaluación (A3) y procesado (A4).

A1. Extracción: Las imágenes son obtenidas del servicio web, pero puede intercambiarse a cualquier otro servicio, almacenadas en una base de datos y puestas a disposición del analizador. Este conjunto también podría ser formado por diferentes imágenes reales y no solo es compatible con imágenes generadas por redes generativas, aunque debe contener imágenes en las que se pueda reconocer un rostro, y estar mínimamente centrado en el contenido de la fotografía.

A2. Análisis de métricas: Cuando el rostro se somete a un análisis de métricas, se pretende obtener diferentes puntos relevantes de la fotografía que permitan medir la apertura de

los ojos, la apertura de la boca, la distancia a los bordes, la rotación de la cara o si el rostro de la fotografía porta o no gafas. También, será útil analizar otras características de la imagen, como la exposición y la nitidez.

A3. Evaluación: Una vez recogidas las diferentes métricas, se normaliza cada una de las características en una escala $[0-1]$ y ponderando cada característica con un peso para obtener la puntuación final. La escala de la puntuación final es $[0-100]$, permitiendo una lectura e interpretación más sencilla. En esta escala, cuanto mayor sea la puntuación, más idónea será la imagen para su uso en un documento de identidad.

A4. Procesado: Aquellas imágenes que cumplan el estándar mínimo para poder emplearse como elemento de identificación en el perfil biográfico del DNI se les aplicará el procesado de imagen. Este procesado consiste en el ajuste a la relación 32:26, la eliminación del fondo mediante técnicas de inteligencia artificial, una conversión a escala de grises, que incluye un ajuste de balance de grises, y un difuminado en los bordes del rostro para su mejor integración en el soporte base.

El Algoritmo 1 describe los pasos del proceso de evaluación de las fotografías. Tras obtener la imagen de la persona, se hace uso de una red neuronal [17] que determina la edad y sexo de la persona que aparece en la fotografía.

Algoritmo 1 Proceso de evaluación de imágenes

para $img \in imagenes$ **hacer**
  $puntosClave \leftarrow calcularPuntosClave(img)$
  $metricas \leftarrow calcularMetricas(img, puntosClave)$
  $puntuacion \leftarrow calcularPuntuación(metricas)$
  $puntosCorte \leftarrow calcularPuntosCorte(puntosClave)$

  $img \leftarrow recortar(img, puntosCorte)$
  $img \leftarrow eliminarFondo(img)$
  $img \leftarrow convertirGris(img)$
  $img \leftarrow modularNivelesColor(img)$
  $img \leftarrow desenfocarBordes(img)$
  $guardar(img, puntuacion)$
fin para

Para obtener la fecha de nacimiento, la edad es restada a la fecha de generación del perfil, así como un número de días aleatorio entre 0 y 364.

B. Resto de datos personales y del soporte

El sexo y la fecha de nacimiento determinan algunos elementos de identificación del DNI, como por ejemplo el nombre o la fecha de validez. Generar el nombre de un perfil biográfico de un DNI consiste en elegir de forma aleatoria un nombre de entre los todos los nombres de personas españolas, en función del sexo determinado en la primera iteración. La elección de este nombre va supeditada a la frecuencia de aparición del mismo. La base de datos se puede consultar en [18]. De misma forma, los dos apellidos del perfil biográfico del DNI son electos de forma aleatoria de entre el listado de todos los apellidos de personas españolas cuya frecuencia es mayor a veinte. La base de datos de apellidos se encuentra en el siguiente fichero [19]. Añadir, que en este caso, por defecto, la nacionalidad será española. Para generar el nombre de los progenitores se sigue el mismo método que con el nombre del sujeto, aunque en este respecto se obtiene un nombre de mujer y otro de hombre de forma aleatoria del listado de nombres.

Tabla II: Letra del DNI

Resto	0	1	2	3	4	5	6	7	8	9	10	11
Letra	T	R	W	A	G	M	Y	F	P	D	X	B

Resto	12	13	14	15	16	17	18	19	20	21	22
Letra	N	J	Z	S	Q	V	H	L	C	K	E

El elemento que se emplea para identificar físicamente y fiscalmente a los individuos en España es el DNI, construido por ocho dígitos, completando con ceros a la izquierda en caso de que el número tenga menos cifras. El número de DNI contiene una letra de control, que es calculada mediante la operación módulo 23 al número de DNI. En la Tabla II se observa la equivalencia entre el resultado de la operación y la letra correspondiente. En la generación sintética del perfil biográfico, se obtiene de forma aleatoria un número del rango permitido y se añade la letra de control correspondiente a la aplicación de la operación sobre el número obtenido.

El domicilio se compone del nombre de la vía y el número, así como la información adicional de identificación del domicilio, como portal, escalera o número. El domicilio y el lugar de nacimiento comparten formato y ambos deben incluir el municipio y la provincia dónde se ubica el municipio [20]. Para generar una vía e integrarla en el perfil biográfico, se dispone de una base de datos con el callejero completo de Madrid, y el número mínimo y máximo disponible en la calle. De forma aleatoria, se escoge una calle y se genera un número de domicilio en el rango obtenido. Una vez obtenida la vía, para el domicilio, se escoge un municipio de forma aleatoria de entre todos los municipios de España y se busca la provincia en la que se sitúa el municipio, de misma forma que para el lugar de nacimiento.

Una vez generados todos los elementos de identificación personales, se generan los elementos de identificación del soporte. Estos son cuatro: número de soporte, fecha de emisión, fecha de validez y equipo de expedición. El número de soporte es único para cada soporte físico de DNI, y vincula ese número de soporte a una única persona, la representada en ese DNI. Para generarlo, se obtienen tres letras y otros seis números, formando el número de soporte del perfil biográfico sintético. Por otro lado, la fecha de emisión será la misma que el día que fue generado el soporte, mientras que la de validez será cinco, diez años o permanente después de la fecha actual en función de la edad del rostro. Por último, el equipo de expedición está formado por nueve caracteres alfanuméricos, que se obtendrán de forma aleatoria.

El Algoritmo 2 resume el proceso de generación de datos siguiendo este procedimiento.

Concluyendo con estos elementos de identificación del soporte, solo falta por generar la firma, como elemento de identificación personal.

Algoritmo 2 Proceso de generación de datos

$foto \leftarrow imagenAleatoria()$
$edad, genero \leftarrow estimarEdadGenero(foto)$
$elementos \leftarrow \{\}$
$elementos[\text{"}fecha\_nac\text{"}] \leftarrow hoy() - edad$
$elementos[\text{"}sexo\text{"}] \leftarrow genero$
para $elem \in ePersonales$ **hacer**
  $elementos[elem] \leftarrow random(dataset[elem])$
fin para
para $elem \in eSoporte$ **hacer**
  $elementos[elem] \leftarrow random(patron[elem])$
fin para

C. Zona de lectura mecánica

La zona de lectura mecánica está definida por la norma ICAO-9303 [21]. En esta se explica que los documentos de viaje asociados producidos conforme al documento 9303 de la Organización de Aviación Civil Internacional incorporan una zona de lectura mecánica para facilitar la inspección de los documentos de viaje y reducir el tiempo de inspección de los documentos. Además, los elementos de esta zona permiten verificar la información que consta en la zona de inspección visual.

La proyección de esta zona en un soporte de viaje debe ser a través de una tipografía OCR-B, recogida en la norma ISO 1073-2. En esta zona se resumen los datos más relevantes de entre los elementos de identificación, entre ellos, el país expedidor en estándar ISO 3166-alfa-3, el número único de identificación del documento, que es el número de soporte en el perfil biográfico, otros números de identificación del estado expedidor, que sería el número del DNI, la fecha de nacimiento, el sexo y la fecha de caducidad. Por último, se incluyen los apellidos y nombres del sujeto representado en el soporte. Este elemento tiene el objeto de resumir los elementos de identificación más relevantes, y por ello, no forma parte del perfil biográfico.

D. Generación de firma

Generar diferentes firmas sintéticas puede llegar a ser una cuestión compleja de abordar. Varios autores [22], [23] han desarrollado diferentes métodos de generación de firma a través de la combinación de otras firmas de un mismo sujeto, o generación de firmas a través de complejas redes neuronales que construyen trazos y firmas a través de aportar diferentes elementos de identificación de entrada. Otros métodos más sencillos se basan en el uso de trazos predefinidos por cada letra a representar. La tipografía PWSignatureTwo es un ejemplo de ello.

El Algoritmo 3 resume el proceso de generación de firmas. Se opta por generar firmas que incluyan el nombre y el primer apellido del perfil biográfico, seguido de dos números aleatorios, ya que dicha fuente tiene algunos trazos asociados a ciertos números. Se dibujan dichos caracteres con esa tipografía y se almacena como digitalización de la firma. Algunos ejemplos de generación de firmas, de diferentes grosores, se pueden encontrar en la Tabla III.

Algoritmo 3 Proceso de generación de firmas

$cadena \leftarrow elementos[\text{"}nombre\text{"}]$
$cadena \leftarrow cadena + elementos[\text{"}primer\_apellido\text{"}]$
$cadena \leftarrow cadena + random(2, [0, 7])$
$dibujar(cadena, PWSignatureTwo)$

VI. Creación de perfil biográfico sobre el DNI

Obtenido un perfil biográfico sintético, se puede representar sobre un soporte base de un DNI para obtener una versión digitalizada del mismo, consiguiendo una imagen que represente el perfil biográfico sintético y obteniendo, de misma forma, un soporte sintético. Para representar el perfil biográfico sintético sobre un DNI se ha de obtener un soporte base de este documento sin la representación de elementos de identificación alguno.

Conseguir un soporte base lleva un trabajo costoso de procesamiento de imagen digital, permitiendo eliminar los diferentes elementos de identificación presentes en una copia de un soporte de alta calidad. De misma forma, la obtención de este soporte base viene ligada a una tabla de posiciones, que indica el punto inicial para comenzar la escritura de los diferentes elementos de identificación. Estos puntos deben considerar la tipografía a emplear, que además debe ser similar a la que figura en el DNI, puesto que la impresión de los elementos de identificación sobre el soporte no debe originar diferencias en los soportes bases originales, como por ejemplo, ocultar información del soporte base o no respetar las distancias de bordeado interior y exterior.

El proceso consistirá en estandarizar los elementos de identificación, escribir los elementos de identificación sobre la representación gráfica, y aplicar un posprocesado con diferentes efectos de difuminado y otros elementos.

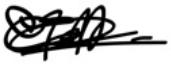
A. Estandarización de los elementos de identificación

El perfil biográfico puede almacenar los elementos de identificación de diferentes formas sin seguir un patrón o una plantilla común. En este caso, los elementos de identificación han de estandarizarse a diferentes reglas de estilo y representación sobre el soporte como cadenas de caracteres.

Los apellidos y nombres deberán escribirse en diferentes líneas y en mayúsculas, por lo que, el primer apellido deberá estar separado del segundo apellido, al igual que los nombres deberán estar separados en diferentes cadenas de caracteres. El número del DNI debe contener exactamente ocho números, por lo que si es un número de menos cifras deberá completarse por ceros por la izquierda, seguidos del dígito de control calculado en función de la tabla II, que del mismo modo deberá escribirse en mayúsculas.

La nacionalidad sigue el estándar ISO 3166-alfa-3, por lo que las siglas del país de nacionalidad se deberán representar siguiendo esa norma, y en mayúsculas, del mismo modo. El sexo deberá indicarse mediante una «M» si el género del perfil biográfico es masculino, o «F» si es de sexo femenino. Por último, todas las fechas deberán seguir el formato `DD MM YYYY`, siendo `DD` el día escrito con dos dígitos, `MM` el mes escrito en formato numérico con dos dígitos, y `YYYY` empleando cuatro dígitos para escribir el año. De misma forma, el número de soporte debe contener las tres letras en

Tabla III: Ejemplos de firmas con tipografía

Caracteres	Grosor 2	Grosor 5	Grosor 8	Grosor 10
daniel ganuza47				
Pacoc50				
CTM26				

mayúsculas, seguido de un número de seis cifras. Al igual que el número de soporte, el equipo de expedición deberá contener todas las letras en mayúsculas, siendo una cadena de caracteres, compuesta de números y letras, de longitud nueve.

El domicilio se compone de tres líneas, en la primera se escribe el nombre de la calle, el número y los complementos, en la segunda línea se escribe el municipio dónde se ubica el domicilio, y en la última línea debe aparecer la provincia que contiene al municipio, y del mismo modo, todas estas líneas deberán ir con las letras en mayúsculas. El lugar de nacimiento también se compondrá de dos líneas, la primera reservada para escribir el municipio, y la segunda para indicar la provincia, que deberán aparecer en mayúsculas. Asimismo, el nombre de los progenitores debe aparecer en mayúsculas, ocupando una sola línea y separados por el símbolo `/`.

La fotografía del rostro deberá convertirse a blanco y negro, eliminando el fondo y quedando solo el rostro de la persona para respetar el formato actual que tiene el DNI. Esta fotografía aparece en dos ocasiones en el DNI.

B. Escritura de elementos de identificación sobre el soporte

El proceso de escritura de elementos de identificación sobre la digitalización del soporte base es el más complejo de actuar. En este paso son requisitos tener la tabla de posiciones de elementos de identificación del soporte, que se debe indicar si debe ir en el anverso o reverso, y la posición dentro de la imagen, asimismo, se debe asignar la tipografía y un tamaño de esta para escribir los datos.

La fotografía deberá plasmarse sobre el soporte en blanco y negro, recortada, sin fondo y reescalada al tamaño que establece el soporte para albergar la imagen. Además, la imagen ha debido cumplir las características necesarias para fijar unos estándares mínimos fijados a la hora de procesar el rostro. En referencia a la firma, también ha de ocupar el espacio máximo permitido, y se eliminará cualquier fondo, resultando exclusivamente el trazo generado.

Una de las medidas que posee el DNI es una ventana transparente con la inscripción del número de soporte. Sobre esta zona se debe inscribir este dato, aunque el tamaño de la fuente debe ser más pequeño que la representación actual.

El otro sistema de seguridad que contiene datos personales visibles es la CLI «*Changeable Laser Image*» –imagen láser cambiante– en bajo relieve. En esta zona debe aparecer la misma imagen del rostro, recortada y sin fondo, pero en una escala menor. De forma superpuesta a esta, irá la fecha de expedición en el siguiente formato `DDMMYY`, siendo `DD` el día con dos dígitos, `MM` el mes formateado a dos dígitos, e `YY` los dos números finales del año.

Tabla IV: Posición y formato de los elementos de identificación del DNI.

Elemento	Posición	Formato	Rep.
Número de DNI	Anverso	8 números y 1 letra	2
Nombre	Anverso	Cadena de caracteres	1
Apellidos	Anverso	Cadena de caracteres	1
Sexo	Anverso	`M` o `F`	1
Nacionalidad	Anverso	Código del país en ISO 1073-2	1
Fotografía	Anverso	Fotografía del rostro	2
Fecha de emisión	Anverso	`DD MM YYYY`	2
Fecha de validez	Anverso	`DD MM YYYY`	1
Fecha de nacimiento	Anverso	`DD MM YYYY`	1
Número de soporte	Anverso	3 letras y 6 números	1
Firma	Anverso	Digitalización	1
Domicilio	Reverso	Cadena de caracteres	1
Lugar de nacimiento	Reverso	Municipio y provincia	1
Progenitores	Reverso	Cadena de caracteres	1
Equipo de expedición	Reverso	Cadena de caracteres	1
Zona de lectura automática	Reverso	Texto OCR-B	1

El último elemento a escribir sobre el DNI es el contenido de la Zona de Lectura Mecánica. Esta deberá digitalizarse con la fuente OCR-B, según lo indicado C, y ocupando los treinta caracteres y las tres filas, conteniendo la información que dicta [21]. Los diferentes elementos de identificación se dispondrán según la Tabla IV.

C. Posprocesado de la generación del soporte

Cuando se finaliza la fase de escritura de elementos de identificación sobre el soporte base, se aplican una serie de transformaciones en la imagen digital obtenida, compuestos por desenfoque y emborronado de los bordes de la representación de las letras grabadas y un ruido Gaussiano que permita homogeneizar el resultado. Tras ello, se posiciona el kinegrama sobre parte del rostro y los elementos de identificación, desplazado ligeramente a la izquierda, tomando como referencia la mitad. El kinegrama es un holograma generado por ordenador capaz de crear imágenes múltiples de alta resolución cuyo diseño puede variar para mostrar animaciones gráficas u otros efectos [24].

La Tabla V muestra el proceso y resultado final, plasmando un ejemplo que demuestra la generación de un documento nacional de identidad sintético y realista conforme al perfil biográfico generado.

Tabla V: Proceso de escritura sobre soporte base

VII. Conclusiones

Este trabajo presenta una metodología que genera de forma automatizada perfiles biográficos sintéticos de acuerdo a los Documentos Nacionales de Identidad de España. La metodología consigue analizar y evaluar la idoneidad de las fotografías de los rostros, generar firmas personalizadas en función de los datos generados y completar un perfil biográfico con diferentes elementos de identificación requeridos en el DNI. El sistema se integra con la representación de la información sobre soportes bases de DNI, obteniendo muestras sintéticas muy realistas.

La aplicación de esta metodología va a permitir crear amplios conjuntos de datos de documentos con muestras que implican información personal. El uso de este nuevo conjunto de datos puede tener finalidades como el entrenamiento de modelos para detección de falsificaciones.

Agradecimientos

Cabe destacar que esta iniciativa se realiza en el marco de los fondos del Plan de Recuperación, Transformación y Resiliencia, financiados por la Unión Europea (Next Generation) en el marco del proyecto con referencia C108/23 “Detección de Falsificación de Documentos de Identidad mediante Técnicas de Visión por Computador e Inteligencia Artificial”. Los autores agradecen la colaboración de la empresa Mobbeel Solutions S.L. por su generoso apoyo, su compromiso con la investigación y la asistencia técnica realizada en el desarrollo de este trabajo.

Referencias

[1] J. I. Agbinya, N. Mastali, R. Islam y J. Phiri: “Design and implementation of multimodal digital identity management system using fingerprint matching and face recognition,”, en *7th International Conference on Broadband Communications and Biomedical Applications*, Melbourne, VIC, Australia, pp. 272-278, 2021.

[2] Akitoshi Okumura, Takamichi Hoshino, Susumu Handa, Yugo Nishiyama y Masahiro Tabuchi: “Identity Verification of Ticket Holders at Large-scale Events Using Face Recognition”, en *Journal of Information Processing*, vol. 25, pp. 448-458, 2017.

[3] Yang Liu, Debiao He, Mohammad S. Obaidat, Neeraj Kumar, Muhammad Khurram Khan y Kim-Kwang Raymond Choo: “Blockchain-based identity management systems: A reviews” en *Journal of Network and Computer Application*, vol. 166, 2020.

[4] H. Nusantoro, R. Supriati, N. Azizah, N. P. Lestari Santoso y S. Maulana: “Blockchain Based Authentication for Identity Management,” 2021 en *9th International Conference on Cyber and IT Service Management (CITSM), Bengkulu, Indonesia*, pp. 1-8, 2021.

[5] V. V. Arlazarov, K. Bulatov, T. Chernov y V. L. Arlazarov: “MIDV-500: A Dataset for Identity Documents Analysis and Recognition on Mobile Devices in Video Stream”, en *Computer optics*, vol. 43, n. 5, pp. 818-824, 2019.

[6] K. Bulatov, D. Matalov y V. V. Arlazarov: “MIDV-2019: Challenges of the modern mobile-based document OCR”, en *Proc. SPIE 11433, Twelfth International Conference on Machine Vision*, 2020.

[7] K.B. Bulatov, E.V. Emelianova, D.V. Tropin, N.S. Skoryukina, Y.S. Chernyshova, A.V. Sheshkus, S.A. Usilin, Z. Ming, J.-C. Burie, M. M. Luqman y V.V. Arlazarov: “MIDV-2020: A Comprehensive Benchmark Dataset for Identity Document Analysis”, en *Computer Optics*, vol. 46, n. 2, pp. 252-270, 2022.

[8] Á. Soares, R. das Neves Junior y B. Bezerra: “BID Dataset: a challenge dataset for document processing tasks", en *Anais Estendidos da XXXIII Conference on Graphics, Patterns and Images*, pp. 143-146, 2020.

[9] D. Benalcazar, J. E. Tapia, S. Gonzalez, C. Busch: “Synthetic ID Card Image Generation for Improving Presentation Attack Detection”, en *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1814-1824, 2023.

[10] D. Bothra, S. Dixit, D. P. Mohanty, M. Haseeb, S. Tiwari y A. Chaulwar: “Synthetic Data Generation Pipeline for Private ID Cards Detection”, en *2023 IEEE Women in Technology Conference*, pp. 1-6, 2023.

[11] F. Attivissimo, N. Giaquinto, M. Scarpetta y M. Spadavecchia: “An Automatic Reader of Identity Documents”, en *2019 IEEE International Conference on Systems, Man an Cybernetics*, pp. 3525-3530, 2019.

[12] L. Zhao, C. Chen, J. Huang: “Deep Learning-Based Forgery Attack on Document Images”, en *IEEE Transactions on Image Processing*, vol. 30, pp. 7964-7979, 2021.

[13] Organización de Aviación Civil Internacional: “Especificaciones para documentos oficiales de viaje de lectura mecánica (MROTD) de tamaño DV1”, en *Documentos de viaje de lectura mecánica*, ed. 8ª, 2021.

[14] Justino García del Vello: “Estimación de los DNIs dDNIcados en España”, en *Estadística Española*, vol. 38, n. 141 y, pp. 219-235, 1996.

[15] Ministerio del Interior: “Real Decreto 1553/2005, de 23 de diciembre, por el que se regula la expedición del documento nacional de identidad y sus certificados de firma electrónica”, en *Boletín Oficial del Estado*, n. 307, 2005.

[16] This Person Does Not Exist, en *https://thispersondoesnotexist.com*

[17] Sawant, Mahesh, “Gender-and-Age-Detection”, en *https://github.com/smahesh29/Gender-and-Age-Detection*

[18] Instituto Nacional de Estadística: “Todos los nombres con frecuencia igual o mayor a 20 personas”, en *https://www.ine.es/daco/daco42/nombyapel/nombres_por_edad_media.xls*, 2023.

[19] Instituto Nacional de Estadística: “Lista de apellidos con frecuencia igual o mayor a 20 personas”, en *https://www.ine.es/daco/daco42/nombyapel/apellidos_frecuencia.xls*, 2023.

[20] Instituto Nacional de Estadística: “Relación de municipios y sus códigos por provincias”, en *https://www.ine.es/daco/inebase_mensual/febrero_2021/relacion_municipios.zip*, 2021.

[21] Organización de Aviación Civil Internacional: “Especificaciones comunes a todos los MRTD”, en *Documentos de viaje de lectura mecánica*, ed. 8ª, 2021.

[22] Ferrer, Miguel A. and Diaz-Cabrera, Moises y Morales, Aythami: “Synthetic off-line signature image generation”, en *2013 International Conference on Biometrics (ICB)*, pp. 1-7, 2013.

[23] Rabasse, Cedric and Guest, Richard M. y Fairhurst, Michael C.: “A New Method for the Synthesis of Signature Data With Natural Variability”, en *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 38, n. 3, pp. 691-699, 2008.

[24] Consejo de la Unión Europea: “Glosario de Términos Técnicos Relacionados con las medidas de seguridad y los documentos de seguridad en general”, en *Registro Público de Documentos Auténticos de Identidad y de Viaje en Red PRADO*, v. 12344/22, 2022.