¿PUEDEN LA PALEOGRAFÍA Y LA CODICOLOGÍA BENEFICIARSE DE LOS AVANCES EN LA INTELIGENCIA ARTIFICIAL?

Ana Belén SÁNCHEZ PRIETO
Universidad Complutense de Madrid

Correría el año 1994 más o menos cuando el profesor Riesco (para mí "don Ángel") me pidió que le ayudase con una publicación con la que andaba ocupado. Yo era por aquel entonces extremadamente joven y me pareció una oportunidad perfecta para aprender directamente del maestro, así que acepté encantada. Mi tarea era "pasar a ordenador" el texto que él me daba mecanografiado y hacer dibujos de letras y otros signos gráficos que luego escaneaba e insertaba en el texto. A don Ángel siempre le fascinó el progreso tecnológico en la medida en que podía ayudar al rendimiento académico e investigador y me contaba cómo él, de joven, fue de los primeros de su entorno en adoptar el uso de la máquina de escribir, y me animaba a continuar aprendiendo sobre "ordenadores" y cómo usarlos. Él, a su edad, ya no se sentía con ánimo para hacerlo. Poco después Internet empezó a popularizarse con la aparición de los navegadores y en pocas semanas me hice adicta a ella. También recuerdo que algunas búsquedas hice para don Ángel, aunque por aquellos años lo que podía encontrarse era una pequeña fracción de lo que está disponible hoy en día.

Con el tiempo mi gusto por los ordenadores y las tecnologías no hizo más que aumentar, y por supuesto se convirtieron en una parte integrante de mi vida personal y académica, pero básicamente se limitaban al uso de Internet (que tanto nos ha facilitado el acceso a bibliotecas y archivos) y las aplicaciones de Microsoft Office. Siempre me gustaron los videojuegos y aprendí a programar un poco, pero solo a nivel de "hobby", porque, después de todo, la informática y las Ciencias Historiográficas no parecen tener mucho en común. O eso me creía yo, hasta que descubrí los "Big Data", el "Aprendizaje Automático" o "Aprendizaje Máquina" (*Machine Learning*, ML), la "Visión por ordenador" (*Computer Vision*), y la Inteligencia Artificial. Cuando esto ocurrió (hace solo unos 6 o 7 años ahora) don Ángel ya se había jubilado y mudado a Málaga, así que nunca tuve la oportunidad de intercambiar opiniones acerca de ello con él,

pero estoy segura de que me hubiese animado a explorar estos terrenos prácticamente vírgenes para los paleógrafos, en cuanto que pueden abrir el camino para una auténtica "paleografía cuantitativa" a gran escala. A medida que estas ideas iban cuajando en mi cabeza, me figuraba que este sería un terreno virgen, porque cómo me iba a imaginar yo que había científicos computacionales con interés por el campo de los manuscritos medievales.

En los párrafos que siguen es mi intención explicar en los términos más sencillos posibles (y por tanto de forma muy incompleta y hasta inexacta) algunos avances que se han producido en el terreno del Aprendizaje Automático (ML) y que pueden ser de interés para paleógrafos y codicólogos, porque los ordenadores y los humanos se complementan casi a la perfección: los humanos somos capaces de percibir patrones y relaciones entre los distintos fenómenos de forma instintivamente inmediata, cosa que no está al alcance de los ordenadores, que en cambio son capaces de realizar sencillas operaciones miles e incluso millones de veces más rápidamente que los humanos. Una advertencia previa: no existe un consenso entre los científicos de la computación acerca del significado y contenido exacto de estos términos (por ejemplo, mientras que algunos consideran que el Aprendizaje Automático es un campo de la Inteligencia Artificial, otros consideran que Aprendizaje Automático e Inteligencia Artificial son disciplinas independientes), y en buena medida la relación entre ellos depende de lo que se quiera entender por cada uno.

"Big Data" es la "administración de grandes volúmenes de datos provenientes de diferentes fuentes y que se generan con rapidez" (Hernández Leal, Duque-Méndez y Moreno-Cadavid, 2017) utilizando una serie de tecnologías bastante complejas que no es el caso referir aquí. Cada vez más a menudo encontramos el término "Ciencia de los Datos" (*Data Science*) utilizado como alternativa a "Big Data", aunque de hecho "Ciencia de los Datos" es un término paraguas que también abarca el Aprendizaje Automático y la Visión por ordenador, y no siempre los datos que analiza son tan "big".

El "Aprendizaje Automático" es la tecnología que permite extraer patrones y relaciones en los datos, de modo que se puedan predecir ciertos comportamientos. En general se considera que es un campo dentro de la Inteligencia Artificial (aunque hay autores que diferencian entre los dos conceptos). Hay dos tipos fundamentales de Aprendizaje Automático: supervisado y no-supervisado. Ambos tipos requieren que el sistema sea alimentado con los conjuntos de datos a analizar, que en la mayor parte de los casos (pero no siempre) consistirán en bases de datos "rectangulares", donde cada entrada está compuesta de una serie de elementos o características que son objeto de observación.

En el aprendizaje no-supervisado, el ordenador "entiende" los datos y produce una evaluación cualitativa que en la práctica suele traducirse en el agru-

pamiento de las observaciones en distintos grupos (*clusters*). Este tipo de algoritmos puede ser útil para descubrir relaciones entre los individuos observados.

Por su parte, el aprendizaje supervisado produce modelos predictivos y resuelve problemas de clasificación y regresión (predicción de una cantidad). El científico de datos trabaja sobre un conjunto de datos (*dataset*) "supervisado" o "etiquetado" por humanos expertos en la materia, y la máquina "aprende" explícitamente las relaciones entre los distintos elementos observados, de modo que pueda predecir el elemento que se desea conocer en nuevas observaciones. Por ejemplo, a partir de datos históricos sobre el abandono estudiantil en una carrera universitaria determinada, se podría predecir qué estudiantes actuales tienen mayores probabilidades de abandono, con la finalidad, acaso, de prestarles una atención más personalizada. En este caso, la variable "objeto" es el abandono, y el algoritmo predice la pertenencia a una de las dos clases posibles: "sí" y "no".

Los algoritmos de aprendizaje automático supervisado se dividen en dos grandes categorías: los más antiguos, basados principalmente en la probabilidad y la estadística, y los más recientes basados en las "redes neuronales artificiales" (*Artificial Neural Networks*) y el "aprendizaje profundo" (*Deep Learning*). Estos últimos producen modelos con menos margen de error pero son más difíciles de interpretar y necesitan mayor potencia de computación.

Aunque sea difícil de creer, no es especialmente difícil producir modelos de aprendizaje automático, porque existen librerías y aplicaciones, algunas de uso libre y gratis, que pueden ser adaptadas fácilmente (relativamente). Otra cosa es, por supuesto, comprender los fundamentos matemáticos subyacentes, pero esto último no es realmente necesario, salvo para los que quieran desarrollar nuevos algoritmos, para lo cual se necesita un más que notable dominio sobre los conceptos matemáticos subyacentes.

1. CONJUNTOS DE DATOS (DATASETS) DISPONIBLES

De hecho, el principal cuello de botella en el aprendizaje supervisado es precisamente la supervisión, ya que se necesitan muchas horas de trabajo por parte de expertos para preparar los conjuntos de datos (lo que se denomina "ground truth") a partir de los cuales el ordenador pueda aprender (Valveny, 2014: 985-987). Lo ideal es que estos conjuntos de datos sean públicos, ya que ello permite a los investigadores comparar los distintos algoritmos en circunstancias similares.

En el caso particular del tratamiento automatizado de manuscritos (entendidos en sentido amplio), el cuello de botella no es la carencia de conjuntos de datos (*datasets*) rectangulares, porque el aprendizaje no se realiza a partir de ellos. En su lugar, es preciso contar con manuscritos "anotados". Básicamente

se trata de análisis de la página y transcripción realizada por expertos paleógrafos que se superponen a las líneas escritas del manuscrito en cuestión. Existen algunas herramientas para "anotar" manuscritos, siendo los pioneros DEBORA (Bourgeois y Emptoz, 2007), DMOS (Coüason *et al.*, 2007), STATE (Gordo *et al.*, 2008), CATTI (Romero *et al.*, 2007) y GIDOC (Serrano *et al.* 2010). Más recientes es tranSriptorium (2013-2015, http://transcriptorium.eu/), Transkribus (https://readcoop.eu/, Kahle *et al*., 2017) y más recientemente GraphManuscribble (Garz, 2021), que utilizan plataformas interactivas.

En la actualidad existen varios conjuntos de datos (*datasets*) que han sido y están siendo utilizados para aplicaciones de análisis de manuscritos medievales:

– IAM-HistDB (diuf.unifr.ch/main/hisdoc/iam-histdb), creada en 2012 por Michael Stolz, es una base de datos de investigación con imágenes de manuscritos históricos que contiene a su vez tres bases de datos distintas (Fischer, 2021, 28 ss):

- Saint Gall, con digitalizaciones de Cod. Sang. 562 (*Vita sancti Galli* de Walafrido Estrabón) copiado en el siglo IX en escritura carolina, en latín.
- Parzifal, con imágenes del poema épico *Parzival* de Wolfram von Eschenbach, Cod. Sang. 857, en escritura gótica del siglo XIII, en alemán.
- Washington, con correspondencia de George Washington, s. XVIII, en inglés, cuyos originales se conservan en la Biblioteca del Congreso de Washington.

– DIVA-HisDB es una colección de imágenes procedentes de tres manuscritos glosados: Cod. Sang. 18, Cod. Sang. 863 (ambos procedentes de San Gal) y Cod. Bodmer 55 (Fondation Martin Bodmer, Cologny-Geneve) (Liwicki, 2021).

– Medieval Paleographical Scale (MPS) es un conjunto de documentos municipales de cuatro ciudades del área de lengua neerlandesa (Leiden, Arnhem, Lovaina y Groninga) de los siglos XIII al 16 (https://research.rug. nl / en / datasets / mps- data- set- with- images- of-medieval-charters-for-handwriting-sty).

– ESPOSALLES, es una colección de 173 páginas con 1747 licencias matrimoniales de entre 1451 y 1905, más 29 páginas de índices, procedentes del Archivo de la Catedral de Barcelona (Romero *et al.*, 2013).

– RODRIGO contiene imágenes de un manuscrito de la *Historia de España* de Ximénez de Rada, copiado en 1545 (Serrano *et al.*, 2010).

– Alcaraz contiene documentos del proceso inquisitorial contra Pedro Ruiz de Alcaraz (1534-1539), con algunas páginas escritas en escritura cortesana y procesal encadenada y muchas abreviaturas; sin embargo, solo se

proveyó de una transcripción profesional para 44 páginas (Villegas *et al.*, 2015).
– WienStUlrich Dataset es un conjunto de 52 páginas del registro de nacimientos de la iglesia de San Ulrich de Viena, del siglo XVI, en alemán y latín y en escritura gótica e itálica (Romero *et al*., 2016).
– *Cantus* (http://cantus.uwaterloo.ca) es la base de datos más importante para notación musical medieval, con digitalizaciones "anotadas" de manuscritos musicales de los siglos XII al XVII (Lacoste *et al*., 2011).

Otros conjuntos de datos que pueden ser interesantes son:

– *Scripta Qumranica Electronica* (https://www.qumranica.org/) combina dos bases de datos principales: la de imágenes digitalizadas de todos los fragmentos recuperados en Qumran, y sus correspondientes transcripciones. Se complementa con el proyecto "The Hands that Wrote the Bible: Digital Palaeography and Scribal Culture of the Dead Sea Scrolls" (https://cordis.europa.eu/project/id/640497).
– GERMANA es resultado de digitalizar y anotar un manuscrito español de 764 páginas con la vida de Germana de Foix, copiado 1891 y conservado en la Colección Nicolau Primitiu, en la Biblioteca Valenciana (Pérez *et al*., 2009).
– Plantas contiene 1035 páginas sacadas de siete volúmenes de la "Historia de las plantas" de Bernardo de Cienfuegos, del siglo XVII, la mayor parte escritas en español, pero con numerosas citas en latín (Toselli *et al.*, 2017).
– GRPOLY-DB agrupa cuatro subconjuntos de datos anotados de forma semiautomática, con un total de 399 digitalizaciones de páginas con escritura manuscrita, mecanografiada e impresa desde mediados del siglo XIX hasta mediados del XX (Gatos *et al.*, 2020).
– Bentham contiene 433 páginas digitalizadas con escritos del filósofo inglés Jeremy Bentham (1748-1832) y su entorno (Causer y Wallace, 2012).
– Novel2Vec agrupa una serie de datasets con digitalizaciones de páginas manuscritas de varios novelistas famosos del siglo XIX (Grayson *et al.*, 2016).
– Pinkas es el primer conjunto de datos construido específicamente para el estudio de manuscritos hebreos medievales (Barakat y Rabaek, 2019).
– Vml-hd (Kassis *et al.*, 2017) y Wahd (Abdelhaleem *et al.*, 2017) reúnen digitalizaciones de textos en árabe, completamente anotados.

Como se puede apreciar, los conjuntos de datos (*datasets*) son bastante pequeños y homogéneos, lo que no sorprende demasiado, ya que en el fondo han sido realizados para hacer las pruebas piloto en el diseño de los algoritmos o

servir de punto de referencia a diversas competiciones en Aprendizaje Automático, pero acaso lo más digno de señalar es que los materiales hispanos medievales son muy escasos y mayoritariamente de la Corona de Aragón. Dado que los conjuntos de datos son imprescindibles para entrenar los algoritmos, sería muy conveniente la formación de equipos de paleógrafos y científicos computacionales para la creación de datasets anotados que permitieran el procesamiento de documentos y códices de todas las áreas geográficas españolas y todos los tipos de escritura.

2. Principales líneas de investigación

El resto de la contribución está destinado a pergeñar las principales líneas de investigación dentro del Aprendizaje Automático que pueden resultar de interés para la Paleografía y la Codicología. En líneas generales, se pueden dividir en las siguientes grandes áreas: cuestiones relacionadas con el formato de la página, clasificación del tipo de letra, transcripción (reconocimiento óptico de caracteres), identificación de imágenes, datación, y mejora de la calidad de la imagen. Aunque en muchos de los proyectos han participado en alguna medida paleógrafos, codicólogos y bibliotecarios, en algunas ocasiones es evidente un cierto desfase de vocabulario.

2.1. Impaginación

La primera línea de investigación es la relacionada con la impaginación. De hecho, el primer paso para prácticamente cualquier otra operación es separar la escritura del fondo. Esta operación, que es trivial para los humanos hasta el punto de que ni siquiera somos conscientes de que la realizamos, es sin embargo muy compleja para un ordenador. Cuando se refieren a las tareas de reconocer las diferentes partes de la página, los científicos de la computación utilizan el término "segmentación". Primero se separa el fondo de la página, luego los elementos gráficos no-textuales y la notación musical, si la hubiere, para después diferenciar las líneas de escritura (a su vez discriminando entre el texto principal y las glosas) y finalmente las letras individuales. Cada uno de estos pasos representa un auténtico reto (Seuret, 2021, 67-75).

Aunque en los últimos años se ha avanzado mucho en esta dirección, sin embargo los algoritmos desarrollados no son todavía capaces de distinguir correctamente las partes constituyentes de impaginaciones muy complicadas. Precisamente para avanzar en esta dirección en 2017 la Universidad de Friburgo, con financiación de la Fundación Nacional Suiza para la Ciencia, organizó la Competición para el Análisis de la Impaginación de Manuscritos Medievales Complicados (*Competition on Layout Analysis for Challenging Medieval Manuscritps, https://diuf.unifr.ch/main/hisdoc/icdar2017-hisdoc-layout-comp*) para la cual, además, se creó la base de datos DIVA-HisDB, reseñada más arriba,

que sirvió como banco de pruebas (Simistira *et al.*, 2017). En ella participaron ocho grupos de investigación de seis países distintos, además del grupo de la universidad convocante (que no participó en la competición), que compitieron en tres tareas distintas: análisis de la impaginación, detección de líneas de escritura, y segmentación de las líneas de texto.

Hasta donde hemos podido encontrar, no existen trabajos que aborden la automatización del estudio de la impaginación en el sentido que lo hacen los codicólogos, reconociendo las medidas y las proporciones de la caja de escritura, los márgenes, etc, aunque esto no en principio tiene mucha menos complicación.

2.2. Clasificación del tipo de escritura

La segunda área de investigación a destacar es la identificación de los tipos de letra. En este sentido las primeras investigaciones se remontan a 1999, cuando un grupo de científicos computacionales de la universidad de Pisa coordinados por los profesores Alessandro Sperduti y Antonina Starita trabajaron en el desarrollo del SPI (System for Palaeographic Inspections). Desgraciadamente, los resultados no alcanzaron los niveles deseados y el proyecto acabó por abandonarse (Ciula, 2005). Quizá la causa del fracaso se debiera a un exceso de ambición de los investigadores, ya que el sistema implicaba desarrollar procedimientos de adquisición (digitalización de los manuscritos), segmentación (reconocimiento de la escritura y las letras individuales dentro de la página), identificación de las letras más relevantes en cuanto a morfología, e identificación del tipo de escritura, cuando ninguno de los campos había recibido todavía atención especializada por parte de los científicos computacionales.

Los subsiguientes intentos, más modestos, para automatizar la identificación de los tipos escriturarios se han realizado independientemente en función de las escrituras. Los primeros de los que tengo noticias son los de Schomaker, Franke y Bulaku (2007) e Izady y Suen (2008), que han trabajado sobre la identificación los tipos escriturarios latinos; Abuhaiba, Mahmud y Green (1994), Gazzah y Amara (2008), Abandah y Khedher (2009), Mohammed *et al.* (2010), para la escritura árabe en lengua árabe, Shahabi y Rahmati (2006, 2009) para la escritura árabe en lengua farsi; Yosef *et al.* (2004) para la escritura hebrea (con participación de Malachi Ben-Arie), y Joshi, Garg, Sivaswamy (2007) y Padma y Vijaya (2007) para la india. Sin embargo, aunque el foco de los mencionados autores está centrado en torno a un área cultural determinada, muchos de ellos intentan aplicar sus métodos a la escritura latina, aunque sea para alguna lengua moderna (el inglés en la mayor parte de los casos). En algunos de los proyectos mencionados la identificación del tipo de escritura se mezcla con la identificación de la "mano".

De todos modos, los resultados son muy incompletos en lo que se refiere a los tipos de escritura medievales, sobre todo de la Baja Edad Media, debido a la enorme diversidad de las escrituras góticas.

2.3. Transcripción

El Reconocimiento Óptico de Caracteres (OCR) ha sido objeto de atención por parte de los científicos computacionales desde los años 70 del siglo XX y en la actualidad ya puede considerarse una tecnología madura en lo que se refiere a la lectura de textos impresos, así que el interés dentro de este campo se ha trasladado a la escritura manuscrita (Fischer, Liwicki e Ingold, 2021, "Introduction"), tanto moderna como "antigua". Disponer de algún sistema capaz de transcribir escrituras paleográficas dotaría a las numerosas digitalizaciones de manuscritos que se están haciendo accesibles a través de Internet de una nueva dimensión, ya que permitiría realizar búsquedas textuales. Desgraciadamente, el "estado del arte" del reconocimiento de textos manuscritos (*Handwritten Text Recognition*, HTR) está muy distante de poder hacer realidad este objetivo, aunque en las últimas dos décadas se han producido importantes avances que nos proporcionan cierta esperanza (Puigcerver *et al.*, 2021, 226), pero de todos modos hay que señalar que se está trabajando en este terreno.

En este caso, la principal dificultad estriba en el hecho de que en la escritura manuscrita, sobre todo en la escritura cursiva, las letras individuales rara vez se presentan aisladas, de modo que para aislar las letras es preciso reconocerlas previamente, pero para reconocerlas es preciso haberlas aislado. El primero en percatarse de este círculo vicioso fue Sayre, en 1973. Esta es la razón principal por la que la lectura automática de escrituras manuscritas requiera una técnica totalmente distinta que el reconocimiento óptico de caracteres (OCR) de textos impresos a que estamos acostumbrados.

Para las escrituras cursivas los modelos ocultos de Markov (*Hidden Markov Models,* HMM), propuestos por primera vez en este campo por Ploetz y Fink (2009) y las redes de memoria a corto largo-plazo (*Long Short-Term Memory Networks*, LSTM) (Graves *et al.*, 2006) son las tecnologías más prometedoras, porque de alguna manera imitan el proceder del paleógrafo humano, al considerar la probabilidad de que aparezca una letra concreta inmediatamente detrás de otra, así como la probabilidad de que aparezca una determinada palabra en la secuencia textual. Y una ventaja adicional es que para "aprender" el sistema solo necesita "ser entrenado" con la transcripción de un experto (humano, por supuesto) (Fischer, 2021b, 95).

Hasta donde alcanza mi conocimiento, todos los intentos de identificación de las letras se han basado en los conceptos "mallonianos" de morfología y el estilo, y ninguno ha utilizado el *ductus*. El reconocimiento del *ductus* es la apro-

ximación habitual en los paleógrafos "humanos" para leer las escrituras cursivas y quizás "enseñando" a la computadora a identificarlo podrían mejorarse los resultados.

Una alternativa al reconocimiento completo de textos manuscritos es la transcripción asistida por ordenador, que fue explorada por primera vez por Toselli *et al.* (2010) y luego por Romero *et al.* (2012) y Alabau *et al.* (2014). Más recientemente esta forma de interacción entre humano y computadora ha tomado forma en los proyectos TranScriptorium (http://transcriptorium.eu/) y Read (https://eadh.org/projects/read), este último desarrollado por la Asociación Europea de Humanidades Digitales (*European Association for Digital Humanities*, EADH).

2.4. Identificación de palabras clave

A veces una transcripción completa de una obra no es necesaria y basta con identificar algunas palabras clave para, por ejemplo, aislar una serie de manuscritos. Para responder a esta necesidad los científicos de datos han creado la "identificación de palabras clave" (*keyword spotting*) (Frinken y Palakodety, 2021). Gracias a esta técnica se pueden hacer búsquedas en páginas manuscritas digitalizadas pero no transcritas completamente, y además tiene la ventaja de ser mucho más tolerante a errores que la transcripción completa (Stauffer *et al.*, 2018, 1). El usuario formula su búsqueda y el sistema evalúa la similitud con los documentos almacenados y devuelve como una lista ordenada de los resultados más similares a los parámetros de búsqueda, y lo más interesante es que todo el proceso se basa en la correspondencia entre las representaciones de rasgos más comunes, como color, textura, forma geométrica, etc. (Giotis *et al.*, 2017, 311).

Los primeros intentos se deben a Wang *et al.* (2014), que utilizó entonces una aproximación basada en la teoría de grafos, y la competición del año 2016 supuso un buen acicate para el desarrollo de esta aproximación a la recuperación de información (Pratikakis *et al.*, 2016).

Existen varias tecnologías que permiten esta funcionalidad, bien basadas en la probabilidad, como los modelos ocultos de Markov (*Hidden Markov Models*), a veces en combinación con el modelo bolsa de palabras (*bag of features*) y Aprendizaje profundo (*Deep Neural Networks*) (Staufer *et al.*, 3).

2.5. Datación de manuscritos

A menudo las bibliotecas se enfrentan al problema de tener que datar manuscritos que carecen de fecha. Con el tiempo y el entrenamiento necesario, casi todos los paleógrafos desarrollan "el ojo" y la intuición necesaria para poder aventurar una fecha de copia con unos márgenes de error de unas pocas décadas.

Especialmente interesante desde nuestro punto de vista es el proyecto "Graphem" de la Universidad de Orleans, por utilizar una combinación de análisis paleográfico tradicional y análisis de patrones (https://www.univ-orleans.fr/lifo/action.php?id=24&lang=en), aunque este proyecto parece haber sido abandonado. En cualquier caso tenía el inconveniente de que parte del proceso debía hacerse manualmente, lo cual lo hacía inviable para grandes volúmenes de imágenes.

Hasta donde he podido averiguar, los primeros por interesarse en datar escrituras manuscritas de forma automática fueron por un lado Brink, Smit, Bulacu y Schomaker por un lado y Palermo, Hays y Efros por otro, ambos equipos en 2012. En el primer caso, el método utilizado fue la medida del grosor y la dirección de los trazos.

He, Samara, Burges y Schomaker (2014) fueron los primeros en interesarse por desarrollar algoritmos capaces de determinar, en función del estilo de la morfología de la escritura, la fecha de un documento dentro de un cuarto de siglo, para lo cual se sirvieron de una colección de 1706 imágenes de documentos municipales de Arnhem, Lovaina, Leiden y Groninga, datados entre 1300 y 1550. En la misma línea y con las mismas bases de datos, en 2019 Hamid *et al.* utilizaron redes neuronales convolucionales (*Convolutional Neural Networks*, CNN) capaces de analizar y comparar pequeños fragmentos de escritura.

Un año antes, el mismo Hamid y colaboradores habían utilizado un método algo distinto, en este caso utilizando también las redes neuronales convolucionales para analizar un vector de 345 elementos creado previamente a través de distintos procedimientos de extracción (Hamid *et al.*, 2018).

2.6. Restauración digital y mejora de imágenes

Finalmente, también existen algunas contribuciones de Aprendizaje Automático a la restauración digital. La mala calidad de las imágenes es frecuentemente un reto añadido a la dificultad de leer las "escrituras antiguas" por parte de los paleógrafos humanos, pero para los ordenadores una página degradada por el tiempo y las agresiones es un reto mayor todavía, ya que a menudo "confunden" las manchas con elementos gráficos.

Especialmente problemático, tanto para humanos como para computadoras es el traspaso de la tinta de una cara a la otra de la página. Este es probablemente uno de los pocos casos en los que es relativamente fácil implementar algoritmos que permitan corregir la degradación de la imagen, superponiendo las imágenes las dos caras de la misma hoja e identificando, mediante el análisis del color, los rasgos que se han traspasado de un lado al otro. Desgraciadamente, irregularidades en el proceso de adquisición de las imágenes (lo que

generalmente los legos denominamos "escaneo") impiden a menudo que esta operación se pueda llevar a cabo de modo inmediato.

La mayor parte de los investigadores que han trabajado en la mejora de imágenes se han basado en la técnica de binarización (*Binarization Technique*) o algún algoritmo derivado de ella, mapas adaptativos de contraste (*adaptive contrast map*), filtros de Gabor, etc., aunque todas estas técnicas han conseguido un éxito solo relativo (Alexander *et al*, 2020).

En 2020 Alexander, Kumar y Sowmya desarrollaron un algoritmo basado en la Lógica Difusa (*Fuzzy Logic*) con el que consiguieron mejorar los resultados de una manera significativa.

Otra posibilidad es utilizar descomposición con contraste con multirresolución (*Multiresolution Contrast, MC*) para estimar el contraste de los trazos de la escritura en diferentes escalas espaciales, que tiene la ventaja de no depender del tamaño de las imágenes ni de su resolución (Fornes *et al*., 2021, p. 256).

3. CONCLUSIONES

La posibilidad de una paleografía y una codicología cuantitativa depende en buena medida del acceso simultáneo a miles e incluso millones de imágenes, lo cual solo es posible recurriendo a las nuevas tecnologías, en especial la ciencia de los datos (*Data Science*), el aprendizaje automático (*Machine Learning*) y el reconocimiento automático de textos manuscritos (*Handwritten Text Recognition*). Todas estas tecnologías están siendo objeto de la atención de los científicos computacionales y (en combinación con otras tecnologías innovadoras como la computación distribuida y el continuo aumento de potencia y velocidad de los procesadores) parece que pueden alcanzar en los próximos años suficiente madurez como para que puedan convertirse en valiosos instrumentos para paleógrafos, codicólogos e historiadores de la cultura escrita en general. Pero esta madurez jamás podrá lograrse sin la colaboración de los paleógrafos y codicólogos con los científicos computacionales, porque solo paleógrafos y codicólogos profesionales son capaces de producir las "anotaciones" necesarias para que los algoritmos de aprendizaje automático puedan "aprender".

Pensamos, además, que algunas operaciones conducentes a mejorar los algoritmos para el reconocimiento de textos manuscritos podrían beneficiarse de la aplicación sistemática del método paleográfico clásico, por ejemplo, aplicando el análisis del *ductus* en lugar de considerar exclusivamente la morfología y el estilo de la escritura.

Faltan también colecciones de digitalizaciones "anotadas", sobre todo en el ámbito de la Corona de Castilla, que sirvan de referencia para entrenar los algoritmos y poder realizar automáticamente operaciones como datación, identificación de la tipología escriptoria, e incluso transcripción.

Llegados a este punto es legítimo preguntarse si los sistemas de aprendizaje automático llegarán a desplazar a los paleógrafos y codicólogos humanos en un futuro próximo. Predecir el futuro es una empresa arriesgada, pero es bastante difícil que esto suceda. Los ordenadores llevan ya décadas "conviviendo" con los humanos y por supuesto que los han desplazado de algunos trabajos, pero también han hecho posible muchos otros. En el terreno que nos ocupa, una vez que las tecnologías estén lo suficientemente maduras, harán posibles investigaciones que hoy por hoy no lo son.

4. BIBLIOGRAFÍA

Abandah, G. A., Khedher, M. Z. (2009). "Analysis of Handwritten Arabic Letters Using Selected Feature Extraction Techniques", *International Journal of Computer Processing of Languages*, 22: 49-73. https://www.worldscientific.com/doi/abs/10.1142/S1793840609001981

Abdelhaleem, A., Droby, A., Asi, A., Kassis, M., Al Asam, R., El-Sanaa, J. (2017). "Wahd: a database for writer identification of Arabic historical documents", en 207 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), IEEE"64-68. https://ieeexplore.ieee.org/xpl/conhome/8054539/proceeding

Abuhaiba, I. S. I., Mahmud, S. A., Green, R. J. (1994). "Recognition of Handwritten Cursive Arabic Characters", IEEE Transactions on Pattern Analysis and Machine Intelligence, 16: 664-672. https://ieeexplore.ieee.org/document/295912

Alabau, V., Martínez-Hinarejos, C., Romero, V., Lagarda, A. (2014). "An iterative multimodal framework for the transcription of handwritten historical documents", Pattern Recognition Letters 35: 195-2-3. https://www.sciencedirect.com/science/article/abs/pii/S0167865512003765

Alexander, T. J., Kumar, S. S., Sowmya, B. (2020). "Performance Analysis of Fuzzy based Restoriation technique for Ink Bleed-though Degraded Documents", en *Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)*, IEEE: 1429-1434.

Barakat, B. K., Rabaev, I. (2019). "The Pinkas Dataset", en *2019 International Conference on Document Analysis and Recognition (ICDAR)*, Los Alamitos, IEEE: 732-737. https: // www.computer.org / csdl / proceedings / icdar / 2019/1h81oza1jwY

Bourgeois, F. L, Emptoz, H. (2007), "DEBORA: Digital AccEss to BOoks of the RenAissance", *International Journal on Document Analysis and Recognition* 9:193-221.

Brink, A. A., Smit, J., Bulacu, M. L, Schomaker, L. R. B. (2012). "Writer identification using directional ink-trace width measurements", *Pattern Recognition*, 45/1: 162-171. https://www.sciencedirect.com/science/article/abs/pii/S0031320311002810

Causer, T., Wallace, V. (2012). "Building a volunteer community: results and findings from Transcribe Bentham", Digital Humanities Quarterly 6/2. http://www.digitalhumanities.org/dhq/vol/6/2/000125/000125.html

Ciula, A. (2005). "Digital palaeography: using the digital representation of medieval script to support palaeographic analysis", *Digital Medievalist* 1.1. https: // www.researchgate.net / publication / 268371736 Digital palaeography using the digital representation of medieval script to support palaeographic analysis

Coüason, B., Camillerapp, J., Leplumey, I. (2007). "Access by content to handwritten archive documents: Generic document recognition method and platform for annotations", *International Journal on Document Analysis and Recognition* 9: 223-242.

Fischer, A. (2021). "IAM-HistDB A Dataset of Handwritten Historical Documents", en Fischer, Liwicki e Ingold (2021): 25-41.

Fischer, A. (2021b). "Automatic Handwriting Recognition in Historical Documents", en Fischer, Liwicki e Ingold (2021): 94-111.

Fischer, A., Baechler, M., Garz, A., Liwicky, M., Ingold, R. (2015). "A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents", *11th IAPR International Workshop on Document Analysis Systems*, Los Alamitos, CA, IEEE. https://www.researchgate.net/publication / 267034380 A Combined System for Text Line Extraction and Handwriting Recognition in Historical Documents

Fischer, A., Liwicki, M., Ingold, R. (eds.) (2021). *Handwritten Historical Document Analysis, Recognition, and Retrieval – State of the Art and Future Trends*, Machine Perception Artificial Intelligence 89, Singapore, World Scientific Publishing.

Fornés, A., Lladós, J., Pujadas-Mora, J. M. (2021). "Browsing of the Social Network of the Past: Information Extraction from Population Manuscript Images", en Fischer, Liwicki e Ingold (2021): 248-278.

Frinken, V., Palakodety, S. (2021). "Handwritten Keyword Spotting in Historical Documents", en Fischer, Liwicki e Ingold (2021): 112-135.

Gatos, B., Louloudis, G., Stamatopoulos, N., Retsinas, G., Sfikas, G., Giotis, A., Liwicki, F. S., Papavassiliou, V., Katsouros, V. (2021). "OldDocPro: Old Greek Document Recognition", en Fischer, Liwicki e Ingold (2021): 201-223.

Gazzah, S., Amara, N. B. (2008). "Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script", *The International Arab Journal of Information Technology* 5: 92-1-1. https://www.researchgate. net / publication / 220413899 Neural Networks and Support Vector Machines Classifiers for Writer Identification Using Arabic Script

Garz, A. (2021). "GraphManuscribble: Interactive Annotation of Historical Manuscripts", en Fischer, Liwicki e Ingold (2021): 159-41.

Giotis, A. P., Sfikas, G., Gatos, B., Nikou, C. (2017). "A survey of document image Word spotting techniques", *Pattern Recognition* 68: 310-332. https://www.sciencedirect.com/science/article/abs/pii/S0031320317300870

Gordo, A., Llorens, D., Marzal, A., Prat, F., Vilar, J. (2008). "State: A multimodal assisted text-transcription system for ancient documents", *Eight IAPR International Workshop on Document Analysis Systems*, Los Alamitos, IEEE: 135-142.

Graves, A., Fernández, S., Gómez, F., Schmidhuber, J. (2006). "Connectionist temporal classification: Labelling unsegmented sequential data with recurrent neural networks", *Proceedings of the 23rd International Conference on Machine Learning*, New York, Association for Computing Machinery: 369-376. https://dl.acm.org/doi/proceedings/10.1145/1143844

Grayson, S., Mulvany, M., Wade, K., Meaney, G., Greene, D. (2016). "Novel2Vec: Characterising 19th Century Fiction via Word Embeddings", en *24th Conference on Artificial Intelligence and Cognitive Science*, Dublin: 1-12. https:// researchrepository.ucd.ie / bitstream / 10197 / 8360 /1/insight publication.pdf

Hamid, A., Bibi, M., Siddiqi, I., Moetesum, M. (2018). "Historical Manuscript Dating using Textural Measures", en *2018 International Conference on Frontiers of Information Technology (FIT)*, IEEE: 235-240. https://ieeexplore.ieee.org/document/8616997

Hamid, A., Bibi, M., Moetesum, M., Siddiqi, I. (2019). "Deep Learning based Approach for Historical Manuscript Dating", en *2019 International Conference on Document Analysis and Recognition (ICDAR)*, IEEE: 967-972. https://ieeexplore.ieee.org/document/8978080

He, S., Sammara, P., Burgers, J., Schomaker, L. (2014). "Towards style-based dating of historical documents", en *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference*, IEEE: 265-270. https://www.researchgate . net / publication / 263518853 Towards Style - Based Dating of Historical Documents

He, S., Samara, P., Burgers, J., Schomaker, L. (2016). "Image-based historical manuscript dating using contour and stroke fragments", *Pattern Recognition* 58: 159-171. https://www.ai.rug.nl/~sheng/PRdating.pdf

Hernández Leal, E. J., Duque-Méndez, N. D., Moreno-Cadavid, J. (2017). "Big Data: una exploración de investigaciones, tecnologías y casos de aplicación", TecnoLógicas, 20/39, mayo-agosto. http://www.scielo.org.co/pdf/teclo/v20n39/v20n39a02.pdf

Izady, S., Suen, C. Y. (2008). "Online Writer-Independent Character Recognition Using a Novel Relational Context Representation", 2008 Seventh International Conference on Machine Learning and Applications, San Diego, CA: 867-870.

Joshi, G. D., Garg, S., Sivaswamy (2007). "A generalized framework for script identification", International Journal on Document Analysis and Recognition (IJDAR), 10(2): 55-68. http://cdn.iiit.ac.in/cdn/cvit.iiit.ac.in/ images/ConferencePapers/2007/saurabh07Ageneralised.pdf

Kahle, P., Colutto, S., Hackl, G., Mühlberger, G. (2017). "Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents", en *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, Los Alamitos, IEEE: 19-24. https://ieeexplore.ieee.org/document/8270253

Kassis, M., Abdalhaleem, A., Droby, A., Alaasam, R., El-Sana, J. (2017). "Vml-hd: The historical Arabic documents dataset for recognition systems", en 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), IEEE: 11-14. https://ieeexplore.ieee.org/xpl/conhome/8054539/proceeding

Lacoste, D., Kolácek, J., Helsen, K. E. (2011). *Cantus: A database for Latin ecclesiastical chant* (sitio web). https://baobab.biblissima.fr/en/resource-729

Lacoste, D. (2012). "The Cantus Database: Mining for Medieval Chant Traditions", *Digital Medievalist* 7 (publicación electronica). https: // journal.digitalmedievalist.org/articles/10.16995/dm.42/

Liwicki, F. S. (2021). "DIVA-HisDB A Precisely Annotated Dataset of Challenging Medieval Manuscripts", en Fischer, Liwicki e Ingold (eds.) (2021): 43-65.

Moaalla, I, Alimi, M., Lebourgeois, F., Emptotz, H. (2006), "Image Analysis for Palaeography Inspection", *Second International Conference on Document Image Analysis for Libraries*, Los Álamos, CA, IEEE: 303-311. https://ieeexplore.ieee.org/document/1612972

Mohammed, K., Abdil, B., Zaiton, S., Hashim, M. (2010). "Swarm-Based Feature Selection for Handwriting Identification", *Journal of Computer Science*, 6: 80-86. https: // www.researchgate.net / publication / 47554367_Swarm-Based_Feature_Selection_for_Handwriting_Identification

Padma, M. C., Vijaya, P. A. (2007). "Identification of Telugu, Devanagari and English Scripts using Discriminating", *Journal of Computer Science*, 1: 64-

78. https: // citeseerx.ist.psu.edu / viewdoc / download?doi=10.1.1.212.7161&rep=rep1&type=pdf

Palermo, F., Hays, J., Efros, A. A. (2012). "Dating historical color images", in *European Conference on Computer Vision*, Springer: 499-512. https://link.springer.com/content/pdf/10.1007%2F978-3-642-33783-3.pdf

Pérez, D., Tarazón, L., Serrano, N., Castro, F. M., Ramos-Terrades, O., Juan, A.(2009). "The GERMANA database", en *Proceedings of the10th International Conference on Document Analysis and Recognition*, Los Alamitos, IEEE: 301-305. http://www.cvc.uab.es/icdar2009/papers/ 3725a301.pdf

Ploetz, T., Fink, G. A. (2009). "Markov models for offline handwriting recognition: A survey". International Journal on Document Analysis and Recognition 12,4: 269-298. https: // www.researchgate.net / publication / 220163405_ Markov_ models_ for_ offline_ handwriting_ recognition_ A_ survey

Pondenkandath, V., Seuret, M., Ingold, R., Azal, M. Z., Liwicki, M. (2017). "Exploiting State-of-the-art Deep Learning Methods for Document Image Analysis", *14th IAPR International Conference on Document Analysis and Recognition*, Los Alamitos, IEEE, 5:30-35.

Pratikakis, I., Zagoris, K., Gatos, B., Puigcerver, J., Toselli, A. H., and Vidal, E. (2016). "Icfhr2016 handwritten keyword spotting competition (h-kws 2016)", en *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Los Alamitos, IEEE: 613-618. https: // users. iit.demokritos.gr/~bgat/ICFHR_2016_KWS_Comp.pdf

Puigcerver, J., Toselli, A. H., Vidal, E. (2021). "Advances in Handwritten Keyword Indexing and Search Technologies", en Fischer, Liwicki e Ingold (2021): 225-247.

Romero, V., Fornés, A., Serrano, N., Sánchez, J. A., Toselli, A. H., Frinken, V., Vidal, E., Lladós, J. (2013). "The Esposalles database: An ancient marriage license corpus for off-line handwriting recognition", Pattern Recognition 46/6: 1658-1669. https: // riunet.upv.es / bitstream / handle / 10251 / 40254/ PRLesposalles.pdf?sequence=2&isAllowed=y

Romero, V., Toselli, A. H., Rodríguez, L., Vidal, E. (2007). "Computer assisted transcription for ancient text images", Computer Assisted Transcription for Ancient Text Images. 4th International Conference, ICIAR, Berlin-Heidelberg, Springer: 1182-1193. https: // link.springer.com / content / pdf/10.1007%2F978-3-540-74260-9.pdf

Romero, V., Toselli, A., Vidal, E. (2012). *Multimodal Interactive Handwritten Text Recognition, Machine Perception and Artificial Intelligence*, Series in Machine Perception and Artificial Intelligence, 80, Singapore: World Scientific Publishing Company.

Romero, V., Toselli, A. H., Sánchez, J. A., Vidal, E. (2016). "Handwriting transcription and keyword spotting in historical daily records documents", en *Document Analysis System (DAS), 2016 IAPR Workshop*, IEEE: 275-280. https://ieeexplore.ieee.org/document/7490130

Sayre, K. M. (1973). "Machine recognition of handwritten words: A Project report", Pattern Recognition 5/3: 213-228. https://www.sciencedirect.com/science/article/abs/pii/0031320373900447?via%3Dihub

Siministira, F., Bouillon, M., Seuret, M., Würsch, M., Alberti, M., Ingold, R., Liwicki, M. (2017). "ICDAR2017 Competition on Layout Analysis for Challenging Medieval Manuscripts", *14th IAPR International Conference on Document Analysis and Recognition*, Los Alamitos, IEEE, 1:1361-370.

Schomaker, L. (2016). "Design considerations for large-scale image-based text search engine in historical manuscrpt collections", *IT-Information Technology* 58/2: 80-88. https://www.degruyter.com/document/ doi/10.1515/itit-2015-0049/html

Schomaker, L., Franke, K., Bulaku, M. (2007). "Using codebooks of fragmented connected-component contours in forensic and historic writer identification", *Pattern Recognition Letters* 28: 719-727.

Serrano, N., Castro, F., Juan, A. (2010). "The Rodrigo database", en *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, European Language Resources Association (ELRA): 2709-2712. https: // www.researchgate.net / publication / 220746542_The_RODRIGO_Database

Serrano, N., Tarazón, L., Pérez, D., Ramos-Terrades, O., Juan, A. (2010). "The GIDOC prototype", Proceedings of the 10th International Workshop on Pattern Recognition in Information Systems, Funchal, Madeira. https://www.researchgate.net/publication/221383052_The_GIDOC_prototype

Seuret, M. (2021). "Layout Analysis in Handwritten Historical Documents", en Fischer, Liwicki e Ingold (2021): 67-93.

Shahabi, R., Rahmati, M. (2006). "Comparison of Gabor-Based Features for Writer Identification of Farsi / Arabic Handwriting", Proceedings of the Tenth International Workshop on Frontiers in Handwriting Recognition, La Baule, France, Suvisoft. https://www.researchgate.net/publication/228605921_Comparison_of_Gabor-Based_Features_for_Writer_Identification_of_FarsiArabic_Handwriting

Shahabi, F., Rahmati, M. (2009). "A New Method for Writer Identification of Handwritten Farsi Documents", *2009 10th International Conference on Document Analysis and Recognition*: 426-430. Los Alamitos, CA, IEEE. https://ieeexplore.ieee.org/document/5277644

Simistira, F., Bouillon, M., Seuret, M., Würsch, M., Alberti, M., Inglod, R., Liwicki, M. (2017). "ICDAR2017 Competition on Layout Analysis for Challenging

Medieval Manuscripts", en *2017 14th IAPR International Conference on Document Analysis and Recognition*, IEEE: 1361-1370.

Stauffer, M., Fischer, A., Riesen, K. (2018). "Keyword spotting in historical handwritten documents base don graph matching", *Pattern Recognition* 81: 240-253. https: // www.researchgate.net / publication / 324269910 Keyword Spotting in Historical Handwritten Documents based on Graph Matching, https: // icosys. ch / wp-content / papercite-data / pdf /stauffer18keyword.pdf

Toselli, A., Romero, V., Pastor, M., Vidal, E. (2010). "Multimodal interactive transcription of text images", *Pattern Recognition* 43/5: 1814-1825. https://www.sciencedirect.com/science/article/abs/pii/S0031320309004385

Toselli, A. H., Vidal, E., Puigcerver, J., Noya-García, E. (2019). "Probabilistic multi-word spotting in handwritten text images", *Pattern Analysis and Applications* 22: 23-32. https: // www. researchgate. net / publication / 326815301 Probabilistic multi-word spotting in handwritten text images

Vanveny, E. (2014). "Datasets and Annotations for Document Analysis and Recognition", en D. Doermann y Tombre, K. (eds.) Handbook of Document Image Processing and Recognition, London Springer: 983-1009. https://link.springer.com/content/pdf/10.1007%2F978-0-85729-859-1_32.pdf

Villegas, M., Sánchez, J. A., Vidal, E. (2015). "Optical modelling and language modelling trade-off for handwritten text recognition", in *Document Analysis and Recognition (ICDAR) 2015 13th International Conference*, IEEE: 831-835. https://ieeexplore.ieee.org/document/7333878

Yosef, I. B., Kedem, K., Dinstein, I., Beit-Arie, M., Engel, E. (2004). "Classification of Hebrew Calligraphic Handwriting Styles: Preliminary Results", *International Workshop on Document Image Analysis for Libraries*, Palo Alto, CA, IEEE. https: // www.researchgate.net / publication / 4054888 Classification of Hebrew Calligraphic Handwriting Styles Preliminary Results

Wang, P., Eglin, V., Garcia, C., Largeron, C., Lladós, J., Fornés, A., "A Novel Learning-Free Word Spotting Approach based on Graph Representation", en *International Workshop on Document Analysis Systems*, Los Alamitos, IEEE: 2017-211. https://ieeexplore.ieee.org/document/6830999