DOI 10.14672/0.2018.1432

C. HERNÁNDEZ TORNERO, V. ROMERO GÓMEZ, J. A. SÁNCHEZ PEIRÓ, A. H. TOSELLI ROSSI Y E. VIDAL RUIZ

INDEXACIÓN Y RECONOCIMIENTO AUTOMÁTICO DE TEXTO MANUSCRITO

Universitat Autònoma de Barcelona
Universitat Politècnica de València
Universitat Politècnica de València
Universitat Politècnica de València
Universitat Politècnica de València

Resumen

Se especula que la cantidad de texto manuscrito acumulado en documentos custodiados por bibliotecas y archivos alrededor del mundo, supera ampliamente a la cantidad de texto (original) impreso o mecanografiado existente hasta la actualidad. Solo una pequeñísima fracción de esta ingente cantidad de documentos ha sido digitalizada hasta el momento, y de ella solo una parte infinitesimal ha sido transcrita. Así pues, la información de mayor interés contenida en la inmensa mayoría de imágenes digitales (es decir, la información transmitida por el texto), continúa siendo inaccesible para su fácil lectura, edición, indexación y búsqueda. En este artículo se introducen proyectos, y soluciones efectivas recientemente desarrolladas en ellos, para la búsqueda de información y para la transcripción completa de imágenes de documentos manuscritos históricos.

palabras clave: Reconocimiento de texto manuscrito, indexación, búsqueda de palabras clave, transcripción asistida

Abstract

Indexing and automatic recognition of handwritten text

It is speculated that the amount of manuscripts accumulated in libraries and archives around the world far exceeds the amount of (original) text printed or typed to the present. Just a small amount of these documents has been digitized so far, and only part of it has been transcribed. Therefore, the most interesting information contained in the vast majority of digital images (i.e., the information transmitted by the text), remains inaccessible for easy reading, editing, indexing and search. In this article, projects and effective solutions recently developed within their frameworks are presented, both for the search of information and for the complete transcription of historical handwritten documents.

keywords: handwritten text recognition, indexing, keyword spotting, assisted transcription

Recibido: 1 de diciembre 2017
Aceptación definitiva: 1 de junio 2018

1. Introducción

El desarrollo de tecnologías capaces de reconocer e interpretar automáticamente texto escrito ha tenido desde hace varias décadas un gran interés académico, social e industrial. Esto es particularmente cierto cuando el texto está *manuscrito* en papel o en algún otro soporte.

Ante la aparición de las nuevas tecnologías informáticas, el interés por el reconocimiento automático de texto manuscrito decayó durante algún tiempo, bajo la asunción de que estas tecnologías pronto harían desaparecer los documentos de papel y con ellos la necesidad de procesar texto manuscrito. Sin embargo, más recientemente, el reconocimiento de documentos de texto manuscrito ha vuelto a ser un tema candente de investigación y desarrollo, al constatar la ingente cantidad de manuscritos históricos que se conservan en archivos y bibliotecas de todo el mundo, muchos de los cuales se han estado digitalizando a lo largo de las últimas décadas para dejarlos al alcance del público en general.

El documento escrito se usa desde la antigüedad como medio para almacenar y transmitir información. Enormes cantidades de dichos documentos se han conservado hasta la actualidad en museos, archivos y bibliotecas, tanto públicos como privados, en condiciones razonables que permiten abordar el estudio de su contenido. Se especula que la cantidad de texto manuscrito actualmente existente en el mundo es superior a la del texto producido mecánicamente, incluyendo el texto (original) producido de forma digital en las últimas décadas. Solo una pequeñísima fracción de esta ingente cantidad de documentos ha sido digitalizada, y de ella solo una parte infinitesimal ha sido transcrita. Así pues, nuestro conocimiento de la historia de la humanidad está basado en una insignificante muestra de la enorme cantidad de información existente, pero inaccesible de facto. No es de extrañar, por tanto, el gran interés social y comercial en hacer posible el acceso simple y ágil al inmenso legado de información histórica, política, económica, demográfica, y cultural en general, contenido en dichos textos.

Sin embargo, para que las imágenes de texto manuscrito sean realmente útiles deben ser anotadas con información acerca de su contenido. Lógicamente la información más rica acerca del contenido de una imagen de texto es precisamente su *transcripción*. Dado el inmenso volumen de documentos de interés involucrados, dicha transcripción no puede obtenerse de manera manual, por lo que el proceso pasa necesariamente por el uso de métodos automatizados.

En el caso de documentos impresos, la transcripción automática se ha venido abordando desde hace algunas décadas con técnicas de reconocimiento óptico de caracteres (más conocidas como OCR por su expresión en inglés *Optical*

CUADERNOS AISPI 11 (2018): 131-144
ISSN 2283-981X

Character Recognition). Los resultados obtenidos mediante estas técnicas son bastante variables y dependientes de la calidad de los documentos. Pero para documentos en buenas condiciones, las tasas de acierto de caracteres pueden estar por encima de 99% (lo que significa alrededor de un 5% de palabras con algún error). Es importante destacar que las herramientas de OCR actuales se basan en una segmentación explícita de los caracteres que aparecen en el documento digitalizado. Si la segmentación en caracteres (o incluso palabras) no funciona correctamente, entonces la precisión de reconocimiento se reduce drásticamente, hasta hacer prácticamente inservible el resultado.

Cuando los documentos digitalizados son textos manuscritos, la segmentación explícita de los caracteres es simplemente imposible y las técnicas de OCR no son de ninguna utilidad. En la mayoría de documentos manuscritos históricos, ni tan sólo la segmentación en palabras es posible, y es necesario recurrir a técnicas holísticas que no requieren segmentación previa en palabras ni en caracteres y reconocen de forma integrada líneas de texto completas. Estas técnicas, que con frecuencia se denominan simplemente Reconocimiento de Texto Manuscrito (RTM) –en inglés *Handwritten Text Recognition* (HTR)–, utilizan en la actualidad conceptos de Reconocimiento de Formas, Aprendizaje Automático y Lingüística Computacional, tales como Modelos de Markov, Modelos de Lenguaje, Redes Neuronales y Aprendizaje Profundo.

En los últimos años ha habido notables avances en el campo de RTM. Sin embargo, las transcripciones que se obtienen con estos sistemas aún están lejos de ser perfectas. El proyecto europeo *tranScriptorium*[1] (2013-2015) estudió como mejorar las prestaciones de las técnicas de RTM y también como aprovechar las prestaciones de los sistemas existentes para facilitar los trabajos de transcripción e indexación requeridos por archivos y bibliotecas. Actualmente, dicho estudio se continúa en el ambicioso proyecto READ[2].

Dada la complejidad de la mayoría de los documentos manuscritos de interés, para obtener transcripciones sin (demasiados) errores es preciso revisar las transcripciones producidas por estos sistemas. Trabajos recientes han estudiado técnicas interactivas que integran al usuario en el proceso de RTM (Romero, Toselli, Vidal 2012), pero es todavía necesario avanzar más en este campo para que los resultados obtenidos sean suficientemente satisfactorios.

Por otra parte, es importante destacar que con la tecnología de RTM actualmente disponible es posible desarrollar sistemas de indexación y búsqueda

1 <http://transcriptorium.eu> [21/05/2018]

2 <https://read.transkribus.eu> [21/05/2018]

de contenidos en imágenes de documentos manuscritos. Estas técnicas de localización de términos o *palabras clave*, o *key word spotting* (KWS) en inglés, permiten explorar colecciones de imágenes de texto sin transcribir para encontrar aquellas imágenes en las que una determinada palabra o frase puede aparecer con un grado de confianza dado (Fischer *et al.* 2010; Frinken *et al.* 2012; Toselli *et al.* 2016).

Los avances en RTM y KWS se espera que van a tener un fuerte impacto en diversos ámbitos. En el cultural y social, estas tecnologías proporcionarán fácil acceso a la enorme y valiosa información encerrada en documentos cuyos contenidos hoy por hoy permanecen prácticamente inaccesibles. En el ámbito económico, sin duda surgirán nuevos segmentos de negocios asociados a estas tecnologías y sus aplicaciones.

2. Reconocimiento de texto manuscrito

El RTM es una tarea de gran desafío en el reconocimiento de formas. Aunque el texto está básicamente compuesto de caracteres, las aproximaciones tradicionales de reconocimiento de caracteres aislados, tal y como hemos comentado anteriormente, generalmente fracasan en la tarea de RTM. Esto se debe a la imposibilidad material de segmentar de manera fiable un texto continuo en sus caracteres individuales. Sin embargo, los seres humanos realizan estas tareas de segmentación y reconocimiento de una manera natural y sin aparente esfuerzo. La precisión se alcanza gracias a una fuerte cooperación entre diferentes niveles de conocimiento: visual, morfológico, léxico, sintáctico y semántico. En este campo, las técnicas existentes de mayor éxito están basadas en la cooperación de las mencionadas fuentes de conocimiento para conseguir un reconocimiento global.

Los primeros desarrollos en RTM aparecieron hacia finales de los años 60 con aplicaciones restringidas que implicaban vocabularios limitados, tales como el reconocimiento de direcciones postales o cheques bancarios. Sin embargo, no fue hasta varias décadas después cuando estos desarrollos recibieron un fuerte impulso, gracias al uso de tecnologías heredadas del Reconocimiento Automático del Habla (Makhoul *et al.* 1998; Kim 1999; Plamondon, Srihari 2000; Steinherz, Rivlin, Intrator 1999), como los bien conocidos Modelos de Lenguaje (N -gramas) o los modelos ocultos de Markov (Jelinek 1998). Recientemente, esta tecnología ha conseguido una considerable mejora al introducirse el uso de las redes neuronales recurrentes (Bluche 2015; Graves *et al.* 2009) para el modelado morfológico de los caracteres, obteniéndose tasas de error cercanas al 5% al nivel

CUADERNOS AISPI 11 (2018): 131-144
ISSN 2283-981X

de caracteres (Sánchez *et al.* 2014; Sánchez *et al.* 2015). Un aspecto muy destacable de estos modelos estadísticos es que pueden aprenderse automáticamente a partir de ejemplos (Dempster, Laird, Rubin 1977). Esto es, dado un conjunto de imágenes de líneas, párrafos o páginas con su correspondiente transcripción (no necesariamente alineada a nivel de carácter ni de palabra), existen algoritmos robustos que estiman automáticamente los parámetros de estos modelos. En el caso de modelos ocultos de Markov, la estimación se basa principalmente en el llamado algoritmo Baum-Welch o *forward-backward*, mientras los parámetros de las redes recurrentes se estiman mediante una versión *forward-backward* del principio general de descenso por gradiente conocida como *connectionist temporal classification* (CTC) (Bluche 2015; Graves *et al.* 2009). Posteriormente, estos modelos pueden utilizarse para transcribir imágenes que no han sido vistas con anterioridad. Por tanto, esta tecnología se puede aplicar fácilmente a cualquier idioma o sistema de escritura, lo que reduce notablemente los costes de desarrollo de sistemas de RTM.

El RTM se divide en dos etapas fundamentales. La primera etapa consiste en el análisis de la imagen de un documento con el fin de localizar y extraer las diferentes partes a transcribir. En la actualidad la unidad de transcripción más pequeña con la que se trabaja suele ser la línea, aunque idealmente se debería trabajar con unidades de mayor tamaño (párrafos o páginas) puesto que el contexto[3] en el proceso de transcripción es sumamente importante. Para esta etapa se utilizan técnicas de Análisis de Imágenes de Documentos (Pastor i Gadea 2007; Fiel *et al.* 2017), las cuales se encargan de limpiar la imagen, normalizar su geometría, detectar las zonas o bloques de texto, y detectar y extraer las líneas (o unidades mínimas de transcripción). En la segunda etapa, se realiza el propio proceso de transcripción, mencionado en el párrafo anterior.

Aunque el RTM puede aplicarse a documentos de muy diversa naturaleza, una de las aplicaciones más destacables es la transcripción de documentos manuscritos antiguos. Es posiblemente también la aplicación con mayor interés económico a corto y medio plazo. Como se ha comentado anteriormente, esto es debido a la inmensa cantidad de documentos manuscritos antiguos que existen en archivos y bibliotecas, cuyos contenidos permanecen prácticamente inaccesibles al no disponer de su transcripción. Cualquier navegador o buscador de Internet que permitiera búsquedas en este tipo de documentos tendría un gran valor añadido frente a otros productos que carezcan de esta facilidad. Pero también es conveniente destacar otro tipo de aplicaciones a documentos más recientes, tales como transcripciones de censos, transcripciones de historiales o partes médicos, o

3 Por contexto se entiende aquí las palabras cercanas a la palabra que se está transcribiendo.

transcripciones de procesos judiciales, entre muchas otras.

3. Transcripción asistida de texto manuscrito

Aunque el RTM ha avanzado notablemente en la última década hasta obtener prototipos con resultados muy satisfactorios (precisión en algunos casos por encima del 80% de palabras correctas), los resultados conseguidos están lejos de ser transcripciones perfectas. Estos sistemas son de gran utilidad en tareas restringidas con vocabularios pequeños y con restricciones en el estilo de escritura. Sin embargo, en tareas reales sin ningún tipo de restricción, la tecnología actual proporciona resultados erróneos.

Si bien es cierto que la tecnología RTM disponible actualmente puede ser útil para indexar y realizar búsquedas en documentos, si lo que realmente se desea es obtener una transcripción de calidad o sin errores, es necesario que un experto humano (típicamente un paleógrafo en el caso de documentos antiguos) realice un trabajo de revisión y corrección. El escenario más usual que se contempla para este proceso consiste en realizar dicha corrección después de obtener los resultados de reconocimiento; esto es, como un proceso de post-edición. Sin embargo, la post-edición puede resultar ineficiente y poco cómoda, además de económicamente costosa (téngase en cuenta que para este proceso se suelen necesitar expertos en paleografía, y también en la materia de los documentos a transcribir). Una solución alternativa o complementaria es utilizar técnicas interactivo-predictivas (Toselli *et al.* 2010, Toselli *et al.* 2011).

La aproximación interactivo-predictiva supone un marco más cómodo y eficiente para el transcriptor humano, donde se combina la eficiencia de los sistemas de RTM con la precisión de los expertos, permitiendo una transcripción perfecta de las imágenes con un coste admisible. La interacción se basa en la retroalimentación (*feedback*) proporcionada por el usuario. Durante el proceso de transcripción, el sistema tiene en cuenta tanto la imagen de texto que se está transcribiendo como una porción de la transcripción validada por el usuario. La nueva salida proporcionada por el sistema es una nueva hipótesis que tiene en cuenta ambas informaciones. Este proceso continua iterativamente hasta obtener una transcripción satisfactoria para el usuario (Toselli *et al.* 2010; Toselli et al 2017). La tecnología interactivo-predictiva actual se basa en lo que se conoce como grafo de palabras (Romero, Toselli, Vidal 2012).

Cabe destacar que, además de ser un marco más cómodo para el transcriptor, la transcripción asistida puede ahorrar trabajo al corrector humano en el proceso

CUADERNOS AISPI 11 (2018): 131-144
ISSN 2283-981X

de corrección de errores con respecto a utilizar la corrección tradicional de post-edición (Toselli *et al.* 2017).

La aproximación interactivo-predictiva para transcripción fue considerada recientemente en el marco de los proyectos *Multimodal Interaction in Pattern Recognition and Computer Vision* (MIPRCV)[4] y *tranScriptorium*. Aunque dichos proyectos mejoraron significativamente la tecnología y las diferentes formas de interacción entre el usuario y la máquina, dejaron abiertas diversas líneas de trabajo que deberían ser continuadas en el futuro. Algunas de estas líneas ya se están estudiando en el proyecto europeo READ.

4. Búsqueda de términos en imágenes de texto

En las secciones anteriores se ha comentado la gran cantidad de información, potencialmente de gran utilidad, que se "esconde" tras las imágenes digitales de texto manuscrito histórico, así como la inviabilidad de transcribir automáticamente estas imágenes con suficiente precisión. La precisión necesaria puede obtenerse mediante la cooperación persona-máquina, en una aproximación interactivo-predictiva al RTM. No obstante, incluso con la reducción de esfuerzo humano que se logra con estas aproximaciones, dicho esfuerzo aún es completamente prohibitivo cuando se trata de transcribir los centenares de miles o incluso millones de páginas que constituyen muchas de las colecciones de manuscritos de interés.

En esta sección se presenta un planteamiento alternativo en el que la tecnología de RTM actualmente disponible se utiliza, no para transcribir, sino para indexar colecciones de imágenes de texto manuscrito de una forma adecuada que permita búsquedas flexibles de términos o "palabras clave". La idea es anotar cada imagen con información de palabras que probablemente pueden aparecer en ella, junto con las probabilidades y posiciones correspondientes. A esta representación del contenido léxico de una imagen nos referiremos como su *índice probabilístico*.

Para sacar el mejor provecho de este tipo de indexación, se debe permitir al usuario especificar un umbral de confianza como parte de su consulta. Esto le permite decidir el compromiso entre precisión y cobertura (*recall* en inglés) que considera más adecuado en cada consulta.

Por ejemplo, si el usuario está interesado en pasajes de texto en español acerca festividades, quizás intente buscar el termino *fiesta*, con un umbral de confianza medio. Probablemente el sistema detectará una serie fragmentos de imágenes,

[4] <http://miprcv.iti.upv.es> [21/05/2018]

muchos de los cuales serán imágenes de pasajes de texto relevante, aunque algunas ocurrencias relevantes sin duda se podrán perder. Si estas omisiones son problemáticas para las necesidades del usuario, entonces este tratará de incrementar la cobertura disminuyendo el umbral de confianza. Aunque esto sea necesariamente a costa de aumentar los *falsos positivos*, es decir, los fragmentos de imágenes incorrectamente presentados como si fueran relevantes (es decir, a costa de una disminución en la precisión). Por el contrario, si el sistema proporciona demasiadas coincidencias falsas, entonces el usuario puede aumentar la precisión incrementando el umbral de confianza; quizás a expensas de perder más pasajes relevantes (es decir, a expensas de una disminución de la cobertura). Es evidente que tal modelo de consulta basado en umbral de confianza no puede implementarse sólo mediante el uso de métodos convencionales de recuperación de la información textual obtenida como salida ruidosa de un sistema de RTM.

Este modelo de búsqueda basada en confianza está fuertemente emparentado con tecnologías que en la literatura se suelen denominar *key-word spotting* (KWS), como se ha comentado en la introducción. Los métodos tradicionales de KWS (Ahmed *et al.* 2017; Giotis 2017; Puigcerver, Toselli, Vidal 2015; Pratiakis *et al.* 2016), tales como Fischer (2010: 8), entre muchos otros, no distinguen explícitamente la indexación de la consulta propiamente dicha. Así, para cada nueva consulta realizan "sobre la marcha" la búsqueda y el correspondiente cálculo de la confianza de que el término buscado se encuentre en cualquier posición de cualquier imagen de cualquier documento de la colección considerada. Si bien esto es una idea atractiva y conceptualmente aceptable, lamentablemente no es realmente factible en la práctica, debido a la alta carga computacional que entraña el cálculo de la confianza directamente en las imágenes de texto. Esto es así incluso para colecciones relativamente pequeñas. No obstante, en un trabajo posterior, Toselli y Vidal (2013: 4) reducen este coste drásticamente gracias a una representación de las imágenes de texto mediante grafos de caracteres, manteniendo las prestaciones de KWS originales. Aunque técnicas como esta pueden ser útiles para pequeñas colecciones de imágenes de texto, los costes computacionales siguen siendo claramente prohibitivos para grandes colecciones de imágenes, que es exactamente el caso en que el uso de consultas basadas en umbral de confianza es realmente importante en la práctica. Obviamente, los usuarios esperan respuestas a sus consultas en fracciones de segundo, pero KWS "sobre la marcha" puede requerir horas para una sola consulta.

Por este motivo, como se ha comentado anteriormente, las tendencias actuales en KWS adoptan enfoques orientados a la indexación, es decir, a la creación de estructuras de datos (tal vez mediante procesos computacionalmente intensivos)

CUADERNOS AISPI 11 (2018): 131-144
ISSN 2283-981X

que permitan consultas basadas en umbral con bajo tiempo de respuesta. Entre los muchos trabajos recientes que siguen esta tendencia más o menos explícitamente, citaremos Frinken *et al.* 2012 y Toselli, Vidal 2013.

Para incrementar las prestaciones de KWS del método propuesto en Fischer *et al.* (2010), en Toselli *et al.* (2016) se propone el uso de información de niveles lingüísticos superiores al carácter. Concretamente se propone el uso de un sistema completo de RTM estadístico, basado en modelado de caracteres mediante HMMs, modelado léxico mediante redes de estados finitos estocásticas de caracteres y modelado sintáctico mediante n-gramas. El coste de entrenamiento de un sistema de este tipo es el mismo que el del método propuesto en Fischer *et al.* (2010), pero las prestaciones de KWS son muy superiores gracias al modelado del contexto de las palabras proporcionado por los nuevos niveles léxico y sintáctico. Finalmente, el coste de indexación puede mantenerse en niveles moderados similares a los presentados en Toselli *et al.* (2016), gracias al uso de grafos de palabras obtenidos como subproducto de la decodificación por Viterbi de las imágenes de texto a indexar.

Este planteamiento conlleva varias ventajas importantes: a) En contraste con otros enfoques tradicionales de KWS, proporciona un marco bien conocido, computacionalmente y matemáticamente tratable, que necesita pocos heurísticos para llegar a ser útil en la práctica; b) Las medidas de confianza para KWS se basan en probabilidades *a posteriori* de palabras, que están bien definidas y debidamente normalizadas y por tanto permiten alcanzar cómodos compromisos de precisión y cobertura en el uso práctico; c) Se utiliza tecnología holística de RTM, que no requiere ningún tipo de (generalmente difícil o imposible) pre-segmentación en palabras ni en caracteres; y d) lo más importante, permite el aprovechamiento del contexto de palabra, lo que redunda en un comportamiento de KWS significativamente mejor que el de otros métodos que ignoran el léxico y la sintaxis.

Un resumen de resultados obtenidos mediante esta tecnología en imágenes de diversas colecciones de texto manuscrito histórico puede consultarse en Vidal (2017: 16).

Los trabajos más recientes de KWS en imágenes de texto asumen la línea como el nivel de búsqueda más bajo. Esta es una asunción muy conveniente ya que, en la mayoría de los casos de interés, las imágenes de texto pueden ser fácilmente segmentadas automáticamente en líneas con suficiente precisión, y las líneas son posiciones objetivo suficientemente precisas en la práctica para la mayoría de tareas de indexación y consulta en documentos manuscritos. Esta asunción también se contempla en el trabajo presentado en (Toselli, Vidal 2013), donde la medida

de confianza de KWS se basa en la probabilidad *a posteriori* de que una palabra aparezca en cada posición horizontal de una imagen de una línea manuscrita. En base a estas probabilidades se calcula una probabilidad de aparición de la palabra en la imagen completa de la línea. Esta idea se aplica recursivamente para obtener probabilidades *a posteriori* de que la palabra considerada aparezca en la imagen completa de una página, o en el conjunto de imágenes de un libro, o incluso en el conjunto de libros de una colección. Con ello se consigue implementar un indexado jerárquico que posibilita consultas a distintos niveles de granularidad, lo que constituye uno de los aspectos más destacables del trabajo de Toselli, Vidal (2013).

Recientemente, en el proyecto europeo HIMANIS[5] se ha indexado la colección completa de documentos históricos conocida como Chancery, que consta de unas 83.000 imágenes de texto medieval densamente manuscrito en francés y en latín con muy buenos resultados (Bluche *et al.* 2017). El sistema de búsqueda para la colección completa está accesible públicamente[6]. Tanto en este proyecto, como en el proyecto READ anteriormente mencionado, se están utilizando las técnicas de indexación y búsqueda anteriormente comentadas (Toselli *et al.* 2016). En el marco de READ, en colaboración con la Biblioteca Nacional y el Grupo Prolope, se está actualmente realizando un trabajo similar al de HIMANIS para la colección completa de manuscritos del Siglo de Oro del Teatro Español que consta de más de 50.000 páginas. Hasta el momento se han indexado alrededor de 2.300 páginas de unas 20 comedias de Lope de Vega, en las que ya es posible encontrar información mediante búsquedas textuales[7].

5. Conclusiones

RTM y KWS son áreas de investigación que han cobrado un creciente interés en los últimos años. Su progresión ha sido realmente rápida hasta alcanzar prestaciones notables. Gracias a estos desarrollos, actualmente se puede afirmar que la inmensa cantidad de documentos manuscritos que residen en archivos y bibliotecas podrá estar accesible al público en general en un futuro próximo, algo impensable hace tan solo un par de décadas. A pesar de estos enormes avances, todavía queda mucho trabajo pendiente para alcanzar el objetivo final de hacer

5 <www.himanis.org> [21/05/2018]

6 <http://prhlt-kws.prhlt.upv.es/himanis> [21/05/2018]

7 <http://prhlt-carabela.prhlt.upv.es/tso> [21/05/2018]

CUADERNOS AISPI 11 (2018): 131-144
ISSN 2283-981X

cualquier tipo de documento plenamente accesible. Solo por citar algunos de los trabajos pendientes para un futuro inmediato: a) el formato de los documentos antiguos es tan variable que hace que los resultados alcanzados con algún tipo de documentos todavía sean claramente mejorables, como ocurre con los documentos en forma de tablas, la mayoría de los cuales no siguen ningún estándar; b) muchos documentos antiguos contienen infinidad de abreviaturas que no siguen ningún estándar y que en numerosas ocasionas son tan ambiguas que solo se pueden interpretar por el contexto; c) un problema notable en la preparación de un sistema RTM es la necesidad de contar inicialmente con datos de entrenamiento. Esta labor es costosa y cara, puesto que se realiza manualmente por expertos paleógrafos. Un problema que habrá que aliviar en el futuro es la preparación de estos datos de entrenamiento iniciales; d) el proceso de digitalización de los documentos se ha realizado en el pasado de diferentes formatos y con parámetros bastante dispares incluso para una misma colección. Los sistemas RTM y KWS que se desarrollen en el futuro deberán ser robustos a estas variaciones. Esta lista de trabajos para el futuro todavía es incompleta, pero trata de dar una visión de los problemas más inmediatos. En conclusión, este artículo ha pretendido dar a conocer algunas de las característica de RTM y KWS, así como algunas líneas de trabajo futuro que permitirían un rápido avance de este campo de estudio.

Bibliografía citada

Ahmed, Rashad; Al-Khatif, Wasfi G.; Mahmoud, Sabri (2017), "A survey on handwritten documents word spotting", *International Journal of Multimedia Information Retrieval*, 6 (1): 31-47.

Bluche, Théodore (2015), *Deep Neural Networks for Large Vocabulary Handwritten Text Recognition*, Tesis doctoral, Université Paris Sud - Paris XI.

Bluche, Théodore; Hamel, Sebastien; Kermovant, Christopher; Puigcerver, Joan; Stutzmann, Dominique; Toselli, Alejandro; Vidal, Enrique (2017), "Preparatory KWS experiments for large-scale indexing of a vast medieval manuscript collection in the HIMANIS project", *In Proceedings of the International Conference on Document Analysis and Recognition*, 311-18.

Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977) "Maximum likelihood from incomplete data via the EM algorithm (with discussion)", *Journal of the Royal Statistical*

Society, ser. B. 39 (1): 1-38.

Fiel, Stefan; Grüning, Tobias; Gatos, Basilis; Dien, Markus; Kleber, Florian (2017), "cBAD: ICDAR 2017 competition on baseline detection", *Proceedings of the International Conference on Document Analysis and Recognition*.

Fischer, A.; Keller, A.; Frinken, V; Bunke, H. (2010), "Lexicon-free handwritten word spotting using character HMMs", *Pattern Recognition Letters*, 33 (7): 934-42.

Frinken, V; Fischer, A; Manmatha, R; Bunke, H. (2012), "A Novel Word Spotting Method Based on Recurrent Neural Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34 (2): 211-24.

Giotis, Angelos P.; Sfikas, Giorgos; Gatos, Basilis; Nikou, Christophoros (2017), "A survey of document image word spotting techniques", *Pattern Recognition*, 68: 310-32.

Graves, A.; Liwicki, M.; Fernández, S.; Bertolami, R.; Bunke, H.; Schmidhuber, J. (2009), "A novel connectionist system for unconstrained handwriting recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (5): 855-68.

Jelinek, Frederick (1998), *Statistical Methods for Speech Recognition*, Cambridge (Mass.), MIT Press.

Kim, G.; Govindaraju, V.; Srihari, S.N. (1999), "An architecture for handwritten text recognition systems", *International Journal on Document Analysis and Recognition*, 2 (1): 37-44.

Makhoul, J.; Schwartz, R.; Lapre, C.; Bazzi, I. (1998), "A script-independent methodology for optical character recognition", *Pattern Recognition*, 31: 1285-94.

Pastor i Gadea, Moisés (2007), *Aportaciones al reconocimiento automático de texto manuscrito*, Tesis doctoral, Universitat Politècnica de València.

Plamondon, R.; Srihari, S.N. (2000), "On-line and off-line handwriting recognition: a comprehensive survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (1): 63-84.

Pratikakis, I.; Zagoris, K.; Gatos, B.; Puigcerver, Joan; Toselli, Alejandro H.; Vidal, Enrique (2016), "ICFHR2016 handwritten keyword spotting competition (h-kws 2016)", *15th International Conference on Frontiers in Handwriting Recognition*, IEEE: 613-18.

Puigcerver, Joan; Toselli, Alejandro H.; Vidal, Enrique (2015), "ICDAR2015 competition on keyword spotting for handwritten documents", *Document Analysis and Recognition (ICDAR)*, IEEE: 1176-80.

Romero, Verónica; Toselli, Alejandro H.; Vidal, Enrique (2012), *Multimodal Interactive Handwritten Text Transcription*, Machine Perception and Artificial Intelligence (volume 80), Singapore, World Scientific Publishing.

Sánchez, Joan Andreu; Romero, Verónica; Toselli, Alejandro H.; Vidal,

Enrique (2014), "ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (HTRtS)", *15th International Conference on Frontiers in Handwriting Recognition*, IEEE: 181-6.

—, (2015), "ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset", *13th International Conference on Document Analysis and Recognition*, IEEE: 1166-70.

Steinherz, T; Rivlin, E.; Intrator, N. (1999), "Off-line cursive script word recognition-a survey", *International Journal on Document Analysis and Recognition*, 2: 90-110.

Toselli, Alejandro H; Romero, Verónica; Pastor i Gadea, M.; Vidal, E (2010), "Multimodal interactive transcription of text images", *Pattern Recognition*, 43 (5): 1814-25.

Toselli, Alejandro H; Vidal, Enrique; Casacuberta, Francisco (2011), *Multimodal Interactive Pattern Recognition and Applications*, Springer.

Toselli, Alejandro H; Vidal, Enrique; Romero, Verónica; Frinken, Volkmar (2016), "HMM word graph based keyword spotting in handwritten document images", *Information Sciences*, 370-371: 497-518.

Toselli, Alejandro H; Leiva, Luis A.; Bordes-Cabrera, Isabel; Hernández-Tornero, Celio; Bosch, Vicent; Vidal, Enrique (2017), "Transcribing a 17th-century botanical manuscript: Longitudinal evaluation of document layout detection and interactive transcription", *Digital Scholarship in the Humanities*, 33 (1): 173-202.

Toselli, Alejandro H; Vidal, Enrique (2013), "Fast HMM-Filler approach for Key Word Spotting in Handwritten Documents", *12th International Conference on Document Analysis and Recognition*: 501-5.

Vidal, Enrique (2017), "Advances in handwritten keyword indexing and search technologies", *Codicology and Palaeography in the Digital Age 4*, eds. Patrick Sahle; Hannah Busch; Franz Fischer. Norderstedt, Books on Demand: 103-19.

Celio Hernández Tornero es Investigador predoctoral de Filología Española, subvencionado por el MINECO en 2017. Licenciado en Historia y especializado en Patrimonio Bibliográfico y Documental, centra su investigación en la aplicación del desarrollo de tecnologías al estudio de manuscritos antiguos, actualmente a los relacionados con el teatro del Siglo de Oro y, más concretamente, a los vinculados con Lope de Vega. Interesado además en el espectro denominado como Humanidades Digitales aplicadas al Patrimonio Cultural.
celio.hernandez@uab.cat

Verónica Romero Gómez es ingeniera en informática por la UPV y doctora en Informática por

la misma Universidad desde 2010. En 2005 se unió al centro de investigación PRHLT de la UPV. Sus campos de interés incluyen el reconocimiento de formas, la interacción multimodal y las aplicaciones de reconocimiento de texto manuscrito. En estos campos ha publicado más de 60 artículos en revistas, congresos y libros de alto impacto. Actualmente trabaja en el proyecto europeo READ y además es profesora asociada en el Departamento de estadística, investigación operativa y calidad de la UPV.
vromero@prhlt.upv.es

Joan Andreu Sánchez Peiró es Profesor Titular en la UPV y es miembro del PRHLT. Los temas de investigación en los que está interesado incluyen Reconocimiento de Patrones, Aprendizaje Automático y las aplicaciones al Reconocimiento de Texto Manuscrito. El Dr. Sánchez ha participado en diferentes proyectos europeos y nacionales relacionados con estas temáticas y ha coordinado el proyecto europeo TRANSCRIPTORIUM. Es co-autor de más de noventa artículos publicados en revistas y en actas de congresos internacionales.
jandreu@prhlt.upv.es

Alejandro Héctor Toselli Rossi es ingeniero eléctrico por la Universidad Nacional de Tucumán (Argentina) en 1997 y doctor en Informática por la UPV en 2004. Varias estancias posdoctarales cuentan en su haber, como la del Institut de Recherche en Informatique et Systémes Alatoires (IRISA, Rennes, Francia, 2008), con el grupo de investigación Recognition and interpretation of Images and Documents (IMADOC). Actualmente trabaja como investigador a tiempo completo en el PRHLT de la UPV, participando activamente en diferentes proyectos europeos (READ, Carabela, etc.) en temas de reconocimiento en indexación de documentos manuscritos.
ahector@prhlt.upv.es

Enrique Vidal Ruiz es Catedrático de la Universitat Politècnica de València (UPV) y ha sido durante varias décadas co-director del grupo de investigación Pattern Recognition and Human Language Technologies (PRHLT), que actualmente es un centro propio de esta Universidad. Es co-autor de más de doscientos cincuenta publicaciones científicas en las áreas de Reconocimiento de Formas, Interacción Multimodal y aplicaciones en tratamiento automatizado del lenguaje, el habla y la escritura. En estas áreas y aplicaciones ha dirigido diversos proyectos de gran envergadura, incluyendo varios internacionales y uno español del programa Consolider Ingenio 2010. Es miembro de IEEE y *fellow* de la International Association for Pattern Recognition (IAPR) y su índice h según Google Scholar es 47.
evidal@prhlt.upv.es

CUADERNOS AISPI 11 (2018): 131-144
ISSN 2283-981X