

Sistema basado en conocimiento para la obtención de flujos de trabajo aplicables al realce de imágenes digitales de archivos históricos**Knowledge-based system for obtaining workflows applicable to the enhancement of digital images of historical archives****Sistema basado en conocimiento para realizar imágenes digitales**

Yuniel Guzmán Bazán. Ing ⁽¹⁾

Juan Valentín Lorenzo Ginori. Dr.C ⁽²⁾

Julio Cedeño Ferrín. MSc. ⁽³⁾

(1) Universidad de Granma (UDG), Departamento de Desarrollo de Software, Granma – Cuba,
yguzmanb@udg.co.cu

(2) Universidad Central de Las Villas (UCLV), Centro de Estudios Informáticos (CEI), Villa Clara
– Cuba, juanl@uclv.edu.cu

(3) Universidad Estatal del Sur de Manabí.

Contacto: yguzmanb@udg.co.cu

Receptado: 19/09/2017

Aceptado: 13/11/2017

Resumen

Hoy en día existe un creciente deterioro en los documentos custodiados en las instituciones archivísticas cubanas. La manipulación de los investigadores, las manchas, huecos y la humedad, han provocado que sea poco legible el contenido de estos documentos. Por tal razón, se trabaja sobre su vertimiento a formato digital como medida de conservación preventiva. Las imágenes resultado de la digitalización requieren una restauración que mejore su legibilidad. Con tal propósito se creó la aplicación DocLux. Actualmente el proceso de restauración no es efectivo. Los especialistas de los Archivos Históricos deben seleccionar el flujo de trabajo (transformaciones y parámetros) necesario para la restauración de cada documento digitalizado. La inexperiencia de estos especialistas en procesamiento de imágenes digitales afecta la productividad de la restauración. En este trabajo se presenta una herramienta de apoyo para los especialistas de los Archivos Históricos con el objetivo de asistirlos en el proceso de restauración de las imágenes digitales. La herramienta recomienda el flujo de trabajo necesario para restaurar las imágenes digitalizadas. La parte central de la herramienta la constituye un sistema de razonamiento basado en casos y como técnica de extracción de

características de imágenes el histograma de color. La recuperación de los casos más similares se realizó aplicando el coeficiente de correlación de Pearson. Las pruebas realizadas mostraron que con la herramienta propuesta se disminuye el tiempo necesario para restaurar las imágenes.

Palabras Claves: Realce de imágenes digitales, razonamiento basado en casos, histogramas de color, instituciones archivísticas, flujo de trabajo

Abstract: Enhancement of digital images, case-based reasoning, color histograms, archival institutions, workflow

Nowadays, there is an increasing deterioration in documents held in Cuban archival institutions. The handling of the researchers, stains, holes and humidity have turned the content of this documents to be unreadable. For this reason, it has been working on their migration into digital format as a measure of preventive conservation. The result of the scan images requires a restoration to improve readability. To this end, the Doclux application was created. Currently the restoration process is not effective. The specialists of the Archives must select the workflow (transformations and parameters) necessary for the restoration of each scanned document. The inexperience of these specialists on digital image processing affects the productivity of the restoration. This paper presents a support tool for specialists of the Archives, in order to assist in the restoration process of digital images. The tool recommends the workflow necessary to restore the scanned images. The central part of the tool is constituted by a case based reasoning system and the color histogram as feature images extraction technique. The recovery of the most similar cases was performed using the Pearson correlation coefficient. Tests showed that with the proposed tool, the time to restore the images decreases.

Introducción

Los archivos históricos (AH) son depósitos donde se guardan los testimonios escritos, gráficos o audiovisuales, conservados con el fin de servir de fuentes para la investigación (Moreno, 2004). En los documentos que en estas instituciones son custodiados, la antigüedad, la tinta empleada en su confección y la manipulación de los investigadores, han incidido de manera negativa sobre su conservación (Marín, 1992). La creación de repositorios digitales ha sido una de las medidas adoptadas para salvaguardar la información que contienen estos documentos. Con la ayuda de las nuevas tecnologías, una de las alternativas para evitar el deterioro de los documentos es la entrega para la consulta de los investigadores de reproducciones digitales de los documentos (P. González, 1994). Por ello, las instituciones archivísticas son unas de las entidades en las cuales se trabaja con especial interés la preservación digital de la memoria histórica (Esposito et al., 2004; Orera Orera, 2008; Térmens, 2009). A lo largo del proceso se han realizado distintas investigaciones que han permitido el desarrollo de varios algoritmos para mejorar la calidad de la imagen digital: eliminación de manchas o de tintas transparentadas, realce de tintas desvaídas, entre otros (Bescos, 1989; Bescos & Jaque, 1989; Bescos, Secilla, & Navarro, 1990).

En Cuba, las instituciones archivísticas custodian un gran volumen de documentos que constituyen una valiosa fuente para el conocimiento histórico de la nación. La falta de organización, las deficiencias de conservación en los depósitos y la carencia de métodos de prevención en los archivos propician el deterioro del material documental (Borrego, Marta, Reyes, & Cisneros, 2009). Se prevé que la memoria histórica podría desaparecer, ya sea por la manipulación, deterioro causados por manchas, huecos, humedad y otros factores destructivos que perturban la calidad, comprensión y perdurabilidad del contenido. La necesidad de difundir dicho patrimonio, unido a los inconvenientes de los niveles de acceso y las complejidades de su lectura a partir de las condiciones actuales, justifican y hacen necesario, después de su digitalización, una restauración que facilite en las imágenes la lectura del texto que contienen (M. J. A.-C. González, 1996).

DocLux es una aplicación desarrollada en la Facultad de Ciencias Informáticas de la Universidad de Granma, destinada específicamente a realzar las imágenes digitales de los AH. Esta aplicación, única de su tipo en Cuba, consta de funcionalidades básicas para el procesamiento digital de imágenes y otras funciones avanzadas para la limpieza del fondo y para definir el trazo de los textos contenidos en las imágenes. Actualmente se dan los primeros pasos en aras de transcribir el texto. El nivel de deterioro entre las imágenes de los AH cubanos es variable, por tanto, para que el proceso de realce mediante esta aplicación sea efectivo, es necesario ajustar correctamente los parámetros de los filtros

que serán aplicados en cada imagen, además de realizar las transformaciones en un orden que garantice una mejora visual. Para esta tarea, es necesario tener conocimientos teóricos y prácticos sobre tratamiento de imágenes digitales. Los AH cuentan con un personal especializado en la archivística, no así en tratamiento de imágenes digitales. Sus conocimientos sobre este tema carecen de pericia, lo que provoca que a la hora de realzar una imagen dañada desconozcan el flujo de trabajo (transformaciones y parámetros) que deben realizar para mejorarla visualmente. Esto trae consigo que las imágenes no tengan buena calidad perceptual; además de que el proceso se realiza a un ritmo muy lento.

En la actualidad, el campo del procesamiento de imágenes digitales está continuamente evolucionando. El uso de técnicas de Inteligencia Artificial (IA) para la toma de decisiones se encuentra muy generalizado. Esto se debe a que, en este tipo de dominio del conocimiento, la información que se maneja resulta difícil de formalizar por el alto grado de subjetividad asociado (Gallego, 2006). Dentro de las herramientas de IA más utilizadas están los Sistemas Basados en Conocimiento (SBC) y entre estos pueden mencionarse los Sistemas Basados en Reglas, las Redes Neuronales Artificiales, los Sistemas de Inferencia Borrosos y los Sistemas Basados en Casos (Aamodt, 1994; Dubitzky, 1997; Kolodner, 1992; Mariel, 1994).

El objetivo del trabajo es desarrollar un sistema que recomiende el flujo necesario para realzar las imágenes digitales de los AH. El sistema debe garantizar que el proceso de restauración de las imágenes sea más eficiente.

Materiales y métodos

Dentro de la inteligencia artificial (IA) existe una disciplina denominada Ingeniería de conocimiento (IC) que proporciona los métodos y técnicas para construir sistemas computacionales denominados Sistemas Basados en Conocimiento (SBC), estos sistemas se diferencian de otros en su manejo de grandes volúmenes de conocimiento de un dominio (Benítez, Escudero, Kanaan, & Rodó, 2014). El desarrollo de un SBC está condicionado por dos aspectos: el primero es que se cumpla con un conjunto de requisitos necesarios para poder construirlo y el segundo es que se justifique su construcción. Algunos de los aspectos que justifican la construcción de un SBC son:

- La pericia humana puede perderse, ó
- La pericia humana escasea y se necesita en muchos lugares, ó

- La pericia humana se necesita en ambientes hostiles, ó
- Se requiere mejorar la calidad del conocimiento de los expertos humanos.

Ventaja que aportan los SBC:

- Amplia distribución de pericia escasa.
- Fácil modificación (conocimiento explícito y accesible).
- Consistencia en las respuestas (los expertos humanos pueden diferir en sus explicaciones, incluso un mismo experto puede responder de forma diferente en momentos diferentes).
- Gran accesibilidad (los SBC trabajan todos los días).
- Preservación de la pericia (constituye una memoria institucional y poseen la capacidad para adquirir nuevo conocimiento y perfeccionar el que poseen).
- Permiten evaluar el efecto de nuevas estrategias añadiendo o modificando conocimiento.
- Constituyen un entrenador en el dominio de aplicación.

Al analizar los aspectos que justifican la implementación de un SBC y cada una de las ventajas que aportan se puede constatar que, a través de su utilización, se tributa a la solución del problema que presentan los AH planteado anteriormente. Es por ello que se propone utilizar la IC en este trabajo.

Descripción de la técnica de IA a emplear

Una de las tecnologías actuales para construir SBC, son los Sistemas con Razonamiento Basado en Caso (CBS). El Razonamiento Basado en Casos es un enfoque que aborda nuevos problemas tomando como referencia problemas similares resueltos en el pasado. Al razonar basado en casos, el solucionador de problemas recuerda situaciones previas similares a la actual y las usa para ayudar a resolver el nuevo problema. La idea básica es recuperar, adaptar y validar las soluciones encontradas en experiencias previas en un intento de relacionarlas con un problema actual. La arquitectura básica de un CBS consiste de una base de casos (BC), un procedimiento para buscar casos similares y un procedimiento de adaptación para ajustar las soluciones de los problemas similares a los requerimientos del nuevo problema(Althoff, 2001) . Lo anteriormente expuesto es considerado el ciclo de razonamiento basado en casos como se muestra en la figura 1

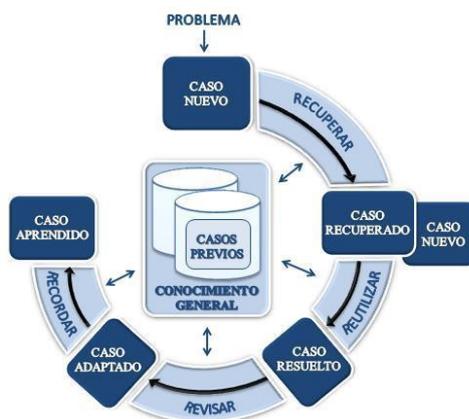


Figura 1. Ciclo de vida del razonamiento basado en casos.

Extracción de características

Algunos de los métodos utilizados para extraer características de una imagen son: directo, formas, texturas y color (Pina, 2010). Los métodos basados en formas, texturas y directo no son adecuados para buscar similitud en las imágenes de los AH. Estas imágenes poseen diferentes tamaños, lo cual descarta la utilización del método directo, y por tratarse de documentos manuscritos su textura sería similar en todas las imágenes, al igual que las formas por predominar el texto sobre hojas de papel. El método utilizado en el presente trabajo es el basado en color, ya que las imágenes de los AH poseen características similares (texto manuscrito, tinta empleada, nivel de deterioro) las cuales afectan directamente su color.

Medidas de similitud de imágenes mediante el histograma de color

Existen medidas de similitud que se pueden utilizar con el fin de comparar imágenes a través de sus histogramas de color, ya que al tratarse de una señal digital el histograma de color extraído de cada imagen se representa como una variable cuantitativa, la cual sigue una distribución normal.

El coeficiente de correlación de Pearson es una medida de similitud entre dos variables que resuelve los problemas de la similitud euclídea y de cómo las dos variables se relacionan entre sí. Este método toma un valor en el rango $[-1, 1]$. Si su valor es 1 indica que las dos variables están perfectamente relacionadas; si es 0, existe una relación débil y si es negativo es que existe una correlación negativa. La medida de correlación de Pearson está dada por la siguiente función de similitud (Benítez et al., 2014; Fernández & Díaz, 2001):

$$d(H1, H2) = \frac{\sum_I (H1(I) - \bar{H} 1)(H2(I) - \bar{H} 2)}{\sqrt{\sum_I (H1(I) - \bar{H} 1)^2 \sum_I (H2(I) - \bar{H} 2)^2}}$$

donde H1 y H2 son los histogramas a comparar, I es cada valor dentro de dicho histograma, $\bar{H} k = \frac{1}{N} \sum_j Hk (J)$ y N es el número total de intervalos del histograma.

Clasificación por indexación de los casos

El sistema cuenta con 213 imágenes en su BC inicial, las cuales poseen dimensiones y características diferentes, lo cual resulta inviable comparar cada nueva consulta con todas las imágenes. Para evitar esta comparación completa (búsqueda lineal) se han indexado los casos, en el cual se clasifica cada imagen en una serie de clases disjuntas, donde cada clase representa condiciones similares.

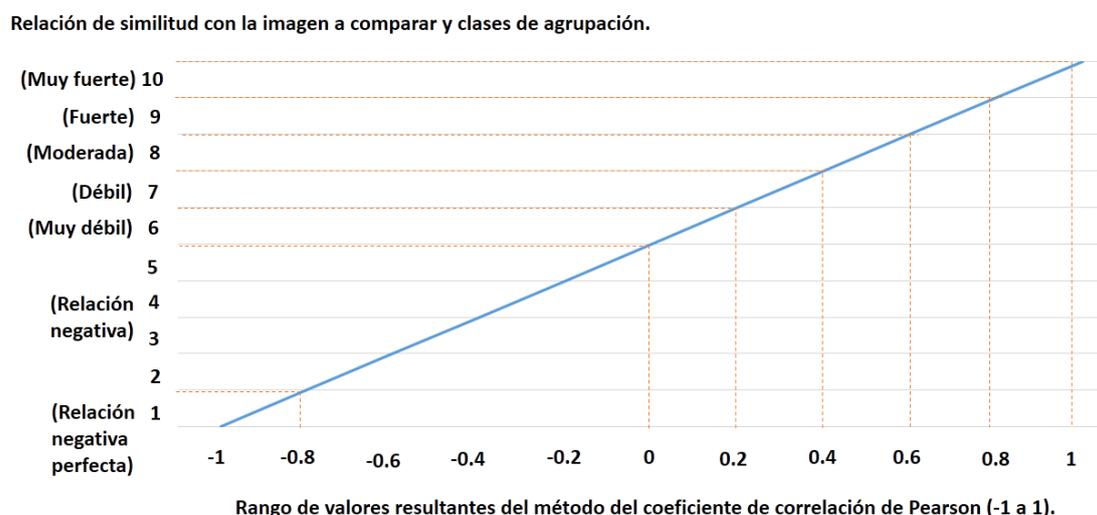


Figura 2. Rango de valores escogidos para la indexación de los casos

En esta clasificación se tiene en cuenta que el método utilizado para la comparación de similitud entre imágenes arroja un valor entre -1 y 1. Teniéndose en cuenta estos valores se crearon 10 clases en las cuales se almacenarán las imágenes correspondientes. Primeramente se toma una imagen como pivote con la cual se comparan las demás (esto es con el apoyo del principio de transitividad que plantea: si, $A \approx B$ y $B \approx C$ entonces $A \approx C$), en dependencia del valor entre -1 y 1 que arroje cada comparación será la clase que se le asigne (índice), además ese mismo valor devuelto por el método de comparación es utilizado para hacer la búsqueda más eficiente. Las clases que facilitarán y disminuirán la complejidad de búsqueda para la posterior fase de recuperación se pueden observar en la siguiente Figura 2.

Recuperación

Una vez entrado un nuevo caso, el sistema debe decidir qué casos recuperar de los que pertenecen a su misma clase. Una vez obtenida la clase de pertenencia (índice) y el valor del nuevo caso, se procede a realizar una búsqueda binaria para recuperar los valores más cercanos al caso de entrada. Luego de obtener un grupo de aproximadamente 6 casos similares, se les aplica el algoritmo de coeficiente de correlación de Pearson con la imagen de entrada para devolver según estos resultados, los casos más similares (los valores de similitud más altos). El sistema recupera de 1 a 3 casos en función de la cantidad de casos almacenados en la BC. Un esquema simplificado del algoritmo que ejecuta el SE para recuperar los casos semejantes de la BC (en pseudocódigo) es el siguiente:

Algoritmo 1. Algoritmo para recuperar los casos más similares

Hp = histograma pivote

Entrada:

N0: nuevo caso (imagen seleccionada).

Método:

- Se obtiene el histograma (Hn) de N0.
- Aplicar función de similitud entre Hn y el histograma de la imagen pivote Hp ($V = f1(Hn, Hp)$), donde $(-1 \leq V \leq 1)$.
- Se calcula la clase de pertenencia (C) del valor V.
- Se obtienen todos los casos (Cc) contenidos en C, la mitad superior de los casos de la clase anterior a (C) y la mitad inferior de la clase siguiente a (C).
- Se aplica una búsqueda binaria en Cc para encontrar la posición (P) del caso que tiene el valor más cercano a V.
- Luego de encontrar P se extraen los 3 casos por encima y los 3 por debajo de dicha posición, siempre que sea posible.

- Se aplica la función de similitud entre los histogramas (H_{ccj}) recuperados en el paso anterior y el histograma del caso a recomendar H_n ($V_{2j} = f_2(H_{ccj}, H_n)$), donde $1 \leq j \leq 6$.
- Se seleccionan los t casos en V_{2j} más altos, es decir, los que mayor grado de similitud con H_n tengan, donde ($1 \leq t \leq 3$).

Salida:

S: Los t casos más semejantes a H_n .

Reutilización

La reutilización implementada es sin adaptación, por lo que solamente se transfieren las soluciones recuperadas al nuevo problema directamente. Sin embargo, el usuario a parte de los flujos de trabajo recomendados por el sistema puede, o bien modificar los parámetros de las transformaciones o aplicar nuevas transformaciones con nuevos parámetros.

Validación- aprendizaje

La eficiencia de un sistema CBS se puede degradar cuando el número de casos crece excesivamente y por tanto, se debe evitar incluir casos que no aporten información nueva al sistema. El rango de posibilidades va desde los sistemas que, de forma autónoma deciden qué casos deben incluir hasta los que delegan esta posibilidad en el mismo usuario. Como puede existir una modificación por parte del usuario en el nuevo caso, es él quien decide si la imagen resultante con sus transformaciones asociadas debe agregarse como nuevo conocimiento. Esto sucede ya que este conocimiento es muy intuitivo y de sentido común, y por tanto difícil de codificar. Hasta el momento no existe ninguna forma de que el sistema verifique si lo que se le ha modificado a la imagen realmente mejora la legibilidad de su contenido. Por tanto, la validación del sistema se centra en verificar si existe un flujo de trabajo en la BC similar al que el usuario desea guardar. El sistema guardará solo aquel flujo de trabajo que constituya un nuevo conocimiento.

Criterios de validación de los nuevos casos

Los criterios de validación utilizados para verificar si el flujo de trabajo que el usuario desea guardar es un nuevo conocimiento son los siguientes:

1. Se verifica el número de transformaciones del flujo de trabajo que se desea almacenar.
2. Se tiene en cuenta el orden de las transformaciones en el flujo.
3. En el caso de las transformaciones con parámetro es necesario comprobar el rango de los valores.



Figura 3. Validación para almacenar los nuevos casos.

En resumen, si en la BC existe un flujo de trabajo con igual número de transformaciones del que se desea almacenar, si las transformaciones que posee están en el mismo orden y si el rango de los valores de las transformaciones con parámetros que contiene está entre ± 8 , el sistema no guardará ese flujo como conocimiento nuevo. Se tiene en cuenta este rango porque se comprobó que en DocLux cuando se varía el brillo o el contraste en una imagen (son las transformaciones con parámetros), los cambios comienzan a ser perceptibles cuando el parámetro para esa transformación se aumenta o disminuye en 8 unidades. La secuencia de validaciones que realiza el sistema para comprobar si el flujo de trabajo aporta información nueva se puede observar en la Figura 3.

Descripción del sistema experto propuesto

El esquema en forma de mapa conceptual que ilustra la Figura 4, muestra la estructura general del modelo que se describe en este trabajo. Los casos en la base de conocimiento (BC) representan las imágenes modelos, es decir, aquellas imágenes que se utilizarán como referencia para que el sistema, a partir de ellas, recomiende posibles transformaciones y parámetros a otras imágenes con características similares. El rasgo distintivo de las imágenes en la BC es el histograma de color. Este rasgo se utilizará para comprobar la semejanza entre una imagen modelo y otra que se desea procesar. Cada imagen modelo tiene asociada una o más transformaciones las cuales serán recomendadas a todas las nuevas imágenes similares que se deseen procesar. Las transformaciones almacenadas estarán representadas por su nombre y los parámetros (en casos requeridos). Ejemplo: (Nombre<Brillo>, Parámetro<50>); (Nombre<Limpiar Fondo>, Parámetro<>).

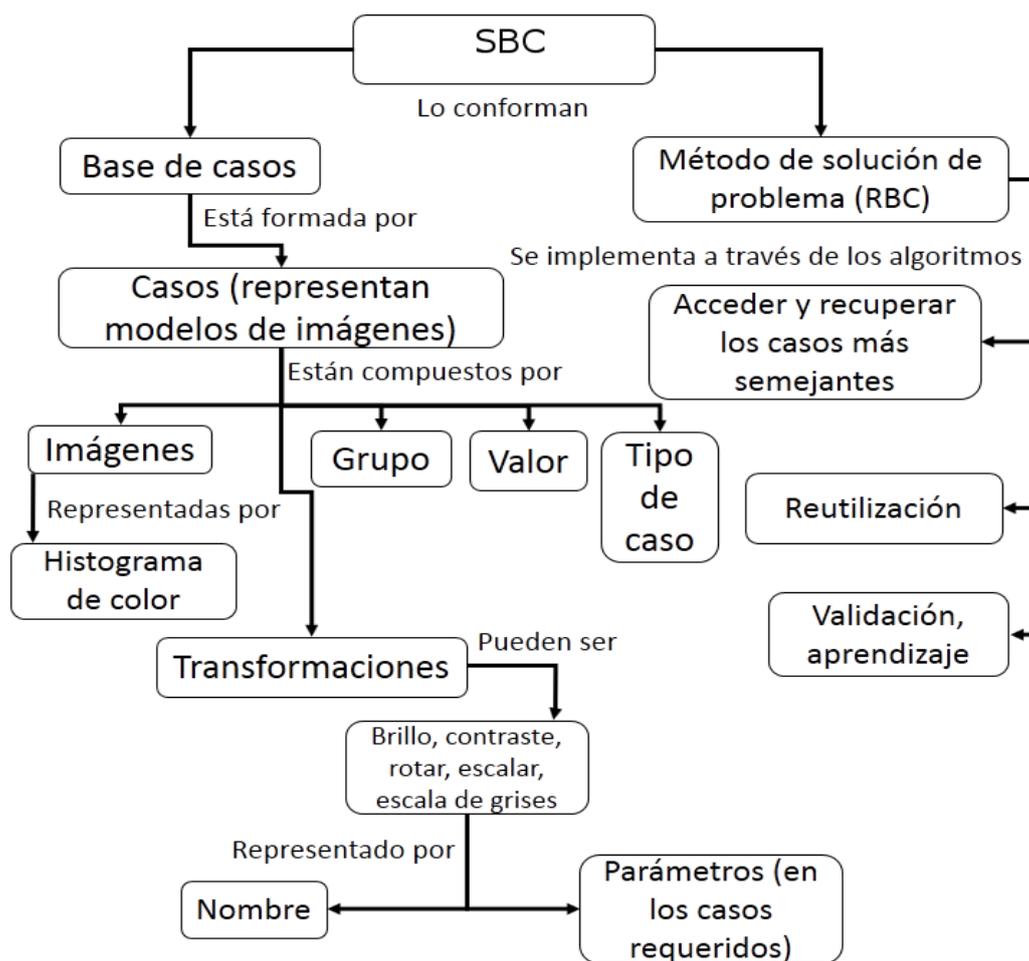


Figura 4. Mapa conceptual del sistema propuesto.

Exportar/Anexar base de conocimiento

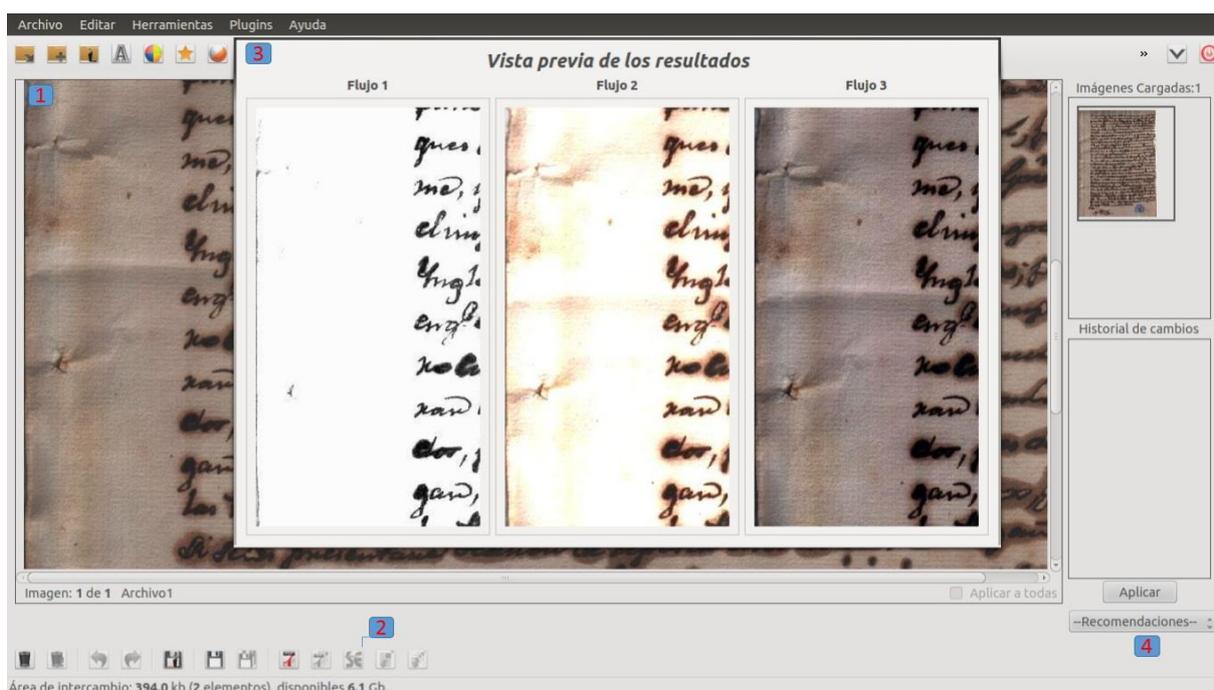
En 1991, se planteó que la construcción de SBC podría orientarse a combinar y ensamblar componentes de conocimiento lo que permitiría reutilizar y compartir conocimiento entre diferentes sistemas (Neches et al., 1991). El sistema propuesto almacena experiencias en realce de imágenes y permite el intercambio de las mismas lo que posibilita que el conocimiento construido de forma colectiva pueda ser aprovechado por especialistas de los AH del país. El sistema tiene implementada la funcionalidad Exportar BC, la cual permite almacenar en un fichero las nuevas experiencias, a partir de la BC inicial, que se han tenido en un AH determinado. Cuenta además con la funcionalidad de Anexar BC, la cual permite incluir a la BC las experiencias de otros AH siempre y cuando no sean las que ya conoce el sistema. Con estas funcionalidades los AH del país pueden compartir sus experiencias y lograr que el sistema instalado en su entidad cuente con una BC más actualizada y por tanto más efectiva.

El fichero en el cual se almacenan las experiencias para ser exportadas/anexadas es un archivo de texto plano de tipo TXT el cual no tiene formato lo que permite que ocupe poco espacio. En un intento de verificar la portabilidad de este archivo y su fácil distribución mediante las redes, se exportó la BC que actualmente contiene 213 casos. El tamaño total del archivo exportado es de 98.4 kb.

Resultados y discusión

Uno de los requisitos fundamentales del sistema propuesto es la visualización de los resultados. A través de esta funcionalidad el usuario podrá visualizar cómo quedaría la imagen si adopta una de las recomendaciones que propone el sistema. El sistema le muestra al usuario el resultado de aplicar los

3



flujos que le propone como se muestra en la figura 5.

Figura 5. Muestra de los resultados de la propuesta del sistema.

En la figura 5 se muestra la interfaz de la aplicación DocLux. El visor en el área 1 es el encargado de mostrar la imagen que se desea realzar. La etiqueta número 2 representa la funcionalidad mediante la cual el usuario solicita que la aplicación le recomiende el flujo de trabajo para realzar la imagen mostrada en el visor. Cuando se selecciona la funcionalidad con la etiqueta número 2 se muestra una interfaz (etiqueta 3) donde el sistema le muestra al usuario los 3 mejores resultados, según su BC, que se pueden obtener al realzar la imagen en el visor. El usuario puede seleccionar un flujo determinado dándole clic a la imagen que para él representa el mejor resultado o simplemente ignorar la recomendación. Si el usuario selecciona un flujo, el visor se actualiza automáticamente con la imagen

seleccionada. En el área con etiqueta 4 el usuario puede ver las transformaciones y los parámetros que componen a cada uno de los 3 flujos que el sistema le recomienda.

Eficiencia en cuanto a tiempo

Para comprobar la eficiencia de la solución propuesta se realizó una prueba para comparar el tiempo que demora un usuario en realizar 5 imágenes a través del proceso actual (con la aplicación DocLux) y luego utilizando DocLux con el SBC incorporado. Naturalmente, la prueba realizada no demuestra rigurosamente la eficiencia del proceso, pero sí da evidencia empírica de esta eficiencia.

DocLux		Propuesta solución	
Imágenes	Tiempo	Imágenes	Tiempo
5	24 min.	5	4.2 min
100	8 hr. (1 jornada laboral)	571	8 hr. (1 jornada laboral)

Tabla 1. Prueba para medir eficiencia.

Como se puede apreciar, el usuario necesitó 24 min a través del proceso actual (proceso de prueba y error) para realizar 5 imágenes de AH, mientras que con la solución propuesta solo necesitó 4 min y 20 seg. Estimando este resultado, en una jornada laboral de 8 hrs el personal del AH realizará aproximadamente 100 imágenes mientras que con la solución propuesta realizará aproximadamente 571. Este resultado exhibe la relevancia de la solución propuesta, toda vez que las categorías productivas de eficiencia y eficacia se evalúan teniendo como variable esencial el tiempo y su empleo. Es válido destacar que para medir el tiempo de la solución propuesta se tuvo en cuenta que el usuario en 4 de las 5 imágenes seleccionadas para el experimento hizo ajustes, según su percepción, al resultado final que propuso el sistema. El tiempo de respuesta del sistema para una imagen depende de la cantidad de transformaciones que se le vaya a realizar.

Evaluación de la calidad subjetiva de las imágenes realizadas

A través de la prueba descrita anteriormente se pudo constatar que la solución propuesta agiliza el realce en las imágenes de los AH otorgando al proceso mayor eficiencia ya que se elimina el proceso de prueba y error por parte del personal. No obstante, es necesario comprobar mediante la solución propuesta si existe un equilibrio entre eficiencia y eficacia en el proceso. Para comprobar que la

eficiencia demostrada no afecta la calidad del realce de las imágenes se realizaron pruebas subjetivas encaminadas a evaluar la calidad de las imágenes que han pasado por el proceso de realce por un método totalmente manual y a través de la solución propuesta.

Para la selección del método a utilizar en el experimento se analizaron los resultados descritos en (Mantiuk, Tomaszewska, & Mantiuk, 2012) y se llegó a la conclusión que el método de selección forzada es idóneo para este experimento, a partir de los resultados de su comparación con otros prominentes métodos subjetivos de evaluación de calidad: de un solo estímulo, de doble estímulo, y juicios de similitud.

Hoy en día se pueden encontrar investigaciones enfocadas a mitigar condiciones puntuales en las imágenes digitales de los AH que afectan la legibilidad del contenido (manchas, traspaso de la tinta, iluminación no uniforme, etc. Aunque se han obtenido buenos resultados, un documento puede mejorar en cierta condición con el procesamiento que se le aplica y al mismo tiempo empeorar en otra (Boussellaa, Alimi, & Zahour, 2007; Mello & Lins, 2002). Por tanto, el presente experimento no estará dirigido a decidir cuál imagen es mejor o peor en sentido general, sino que se tendrán en cuenta 4 condiciones puntuales: eliminación de manchas, contraste entre el texto y el fondo de la imagen, distribución uniforme de los colores y la inteligibilidad del texto. Mediante este sistema se evita el efecto de halo (destacar una condición más sobresaliente sobre las otras) o de tendencia central (evaluar a todas las condiciones de manera conservadora en la media). Para la selección de las condiciones puntuales que serán analizadas en el experimento se tuvo en cuenta que los filtros presentes en DocLux afectan las áreas representadas en las condiciones. Teniendo en cuenta que se utilizaron variables dicotómicas en el experimento, se utiliza la prueba binomial. En esta prueba no paramétrica, la categoría de referencia (referirse a la propuesta) es Automático con el código 1 para las cuatro condiciones evaluadas y el nivel de significación seleccionado fue 0,05.

		Categoría	N	Proporción observada	Proporción de prueba	Significación asintótica. (unilateral)
Eliminación de manchas	Grupo 1	Automatizado	51	,34	,42	,03
	Grupo 2	Manual	99	,36		
	Total		150	1,00		

Tabla 2. Prueba binomial para eliminación de manchas.

Contraste texto vs fondo		Categoría	N	Proporción observada	Proporción de prueba	Significación Asintótica (unilateral)
	Grupo 1	Automático	102	,68	,60	,03
	Grupo 2	Manual	48	,32		
	Total		150	1.00		

Tabla 3. Prueba binomial para contraste entre el texto y el fondo de la imagen

Distribución uniforme de los colores		Categoría	N	Proporción observada	Proporción de prueba	Significación Asintótica (unilateral)
	Grupo 1	Automático	46	,31	,38	,04
	Grupo 2	Manual	104	,69		
	Total		150	1,00		

Tabla 4. Prueba binomial para la distribución uniforme de los colores

Inteligibilidad del texto		Categoría	N	Proporción observada	Proporción de prueba	Significación Asintótica (unilateral)
	Grupo 1	Automático	106	,71	,63	,03
	Grupo 2	Manual	44	,29		
	Total		150			

Tabla 5 Prueba binomial para la inteligibilidad del texto

Como se puede apreciar en el caso de la primera condición "Eliminación de manchas" se selecciona el método automático en el 42 % de las veces con un nivel de confianza de 0.95. Para la segunda condición evaluada "Contraste del texto vs fondo" se obtuvo que a nivel de población en al menos el 60% de los casos, se selecciona el método automático como mejor resultado con un nivel de confianza de 0.95. Para la condición "Distribución uniforme de los colores" se obtiene un por ciento bajo al

resultar que en el 38% de las veces se selecciona el método propuesto como mejor resultado con una confianza de 0.95. En el caso de la condición "Inteligibilidad del texto" el comportamiento a nivel de población es favorable a la solución propuesta, pues se obtuvo, con un nivel de confianza de 0.95, que en al menos el 63% de los casos, el método propuesto arroja mejores resultados.

Los resultados muestran que cuando se aumenta el contraste se logra mayor legibilidad del texto en las imágenes de los AH. Sin embargo, cuando se aplica esta transformación aumenta el ruido contenido en el fondo de las imágenes y no existe un equilibrio entre los colores. Esta situación provoca que aumente el impacto negativo que producen las manchas y que existan colores excesivamente predominantes en las imágenes. La razón de este comportamiento es que la información de interés en las imágenes de los AH puede contener niveles de intensidad iguales a los del ruido y como esta transformación no toma en cuenta la ubicación de los píxeles estos sufrirán la misma transformación que los píxeles pertenecientes al ruido.

En 184 casos de los 213 almacenados se utilizó la transformación contraste. Esto demuestra que es una transformación muy utilizada para lograr mayor legibilidad del texto en las imágenes de AH, sin embargo, aunque el balance de colores fue empleado en un alto por ciento, en solo 16 ocasiones fue aplicado después del contraste. Aplicar el balance de colores después de aumentar el contraste puede ser factible para garantizar que exista una distribución uniforme de los colores presentes en las imágenes. Aunque la aplicación DocLux no cuenta con un filtro dirigido específicamente a eliminar las manchas, los filtros para la limpieza del fondo pueden mitigar su impacto. Al analizar en la base de casos se puede apreciar que en solo 22 ocasiones estos filtros fueron aplicados. Este puede ser el motivo de la baja aceptación que tuvo la condición "Eliminación de manchas". Todo indica que para los especialistas de los AH el objetivo fundamental en el realce de las imágenes es lograr mayor legibilidad del texto aunque para lograrlo se afecten otras condiciones en las imágenes.

Conclusiones

La solución desarrollada constituye un sistema capaz de hacer más efectivo el tratamiento de imágenes digitales de AH, pues a partir de experiencias previas se pueden sugerir transformaciones y parámetros. Para ello se implementaron las etapas del razonamiento basado en casos y tras el análisis de los métodos para la comparación de imágenes, se seleccionó el método basado en color y el coeficiente de correlación de Pearson como idóneos para imágenes digitales de AH. El SBC desarrollado logra, en al menos el 63% de los casos, el objetivo fundamental del procesamiento de imágenes digitales de AH. La solución propuesta consigue también resultados satisfactorios en cuanto

al contraste entre el texto y el fondo de las imágenes en al menos el 60% de los casos. Sin embargo, se obtienen bajos resultados en las condiciones "Eliminación de manchas" y "Distribución uniforme de los colores" con valores de aceptación del 42 % y 38 % respectivamente.

Referencias

1. Aamodt, A. (1994). Case-Based Reasoning: Foundational Issues , Methodological Variations , and System Approaches, 7, 39-59.
2. Althoff, K. (2001). Case-Based Reasoning. Handbook of Software Engineering and Knowledge Engineering. Fraunhofer Institute for Experimental Software. Fraunhofer Institute for Experimental Software Engineering (IESE), Kaiserslautern, 1, 549-587.
3. Benítez, R., Escudero, G., Kanaan, S., & Rodó, D. M. (2014). Inteligencia artificial avanzada. (E. UOC, Ed.).
4. Bescos, J. (1989). Image processing algorithms for readability enhancement of old manuscripts. Proc. Int'l Electronic Imaging Exposition and Conf, 392-397.
5. Bescos, J., & Jaque, F. (1989). Reflectance and Optical Contrast of Old Manuscripts: Wavelength Dependence. Intl Congress on Optical Science and Engineering. International Society for Optics and Photonics, 258-262.
6. Bescos, J., Secilla, J. P., & Navarro, J. (1990). Filtering and compression of old manuscripts by adaptive processing techniques. Proceedings of the Society for Information Display Symposium (Vol. 21, pp. 384-387).
7. Borrego, S., Marta, A., Reyes, C., & Cisneros, G. G. (2009). Diagnóstico del estado de conservación de colecciones fotográficas del Archivo Nacional de la República de Cuba. Códice, 5(1), 67-82.
8. Boussellaa, W., Alimi, A., & Zahour, A. (2007). A Methodology for the Separation of Foreground / Background in Arabic Historical Manuscripts using Hybrid Methodss. En Proceedings of the 2007 ACM symposium on Applied computing. ACM (Vol. 14, pp. 605-609).
9. Dubitzky, W. (1997). Knowledge Integration in Case-Based Reasoning: A Concept-Centred Approach. University of Ulster.
10. Esposito, F., Malerba, D., Semeraro, G., Ferilli, S., Altamura, O., Basile, T. M. A., Berrardi, M., et al. (2004). Machine Learning methods for automatically processing historical documents: from paper acquisition to XML transformation. In C. Society (Ed.), First International Workshop on Document Image Analysis for Libraries (DIAL'04) (pp. 328-335). IEEE. Retrieved from http://www.di.uniba.it/~ceci/Papers/Pubblicazioni/International_Collections/IC.18__20880328.pdf
11. Fernández, S. P., & Díaz, S. P. (2001). Investigación: Relación entre variables cuantitativas, 4, 1-8.

12. Gallego, A. (2006). Modelos visuales en el análisis de la calidad de imagen. Universidad Politécnica de Madrid. Retrieved from http://optica.csic.es/papers/pfc_ana.pdf
13. González, M. J. A.-C. (1996). El Departamento de Coordinación y Normalización: Organización y funcionamiento. Boletín de la ANABAD, 46(1), 261-280.
14. González, P. (1994). El Archivo General de Indias y su proyecto de informatización: nuevas posibilidades para la investigación. Cuadernos de historia moderna, (15), 231-249. Retrieved from <http://revistas.ucm.es/index.php/CHMO/article/download/CHMO9494110231A/23708>
15. Kolodner, J. L. (1992). An Introduction to Case-Based Reasoning, 6(1), 3-34.
16. Mantiuk, R. K., Tomaszewska, A., & Mantiuk, R. (2012). Comparison of Four Subjective Methods. Computer Graphics Forum, 31(8), 2478-2491. doi:10.1111/j.1467-8659.2012.03188.x
17. Mariel, F. (1994). "CBR: A categorized bibliography". The Knowledge Engineering Review, 9, 1-36.
18. Marín, F. M. (1992). La conservación y utilización de textos en el futuro inmediato, 75(4), 1010-1023. Retrieved from <http://www.illf.uam.es/~fmarcos/articulo/92Hadmyte.pdf>
19. Mello, C. A. B., & Lins, R. D. (2002). Generation of Images of Historical Documents by Composition. In Proceedings of the 2002 ACM symposium on Document engineering (pp. 127-133). Brazil.
20. Moreno, A. V. (2004). El tiempo de la archivística : un estudio de sus espacios de racionalidad histórica, II(15), 76-96. Retrieved from <http://www.scielo.br/pdf/ci/v33n3/a10v33n3>
21. Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., & Swartout, W. R. (1991). Enabling Technology for Knowledge Sharing. AI magazine, 12(3).
22. Orera Orera, L. (2008). Preservación digital y bibliotecas: un nuevo escenario. Revista general de información y documentación, 18, 9-24.
23. Pina, A. N. (2010). Clasificación y búsqueda de imágenes usando características visuales. Universidad de Murcia.
24. Térmens, M. (2009). Investigación y desarrollo en preservación digital: un balance internacional. El profesional de la información, 18(6), 613-624. doi:10.3145/epi.2009.nov.04