PRESERVACIÓN DEL PATRIMONIO CULTURAL – DIGITALIZACIÓN Y RECONOCIMIENTO DE DOCUMENTOS MANUSCRITOS

Marisa R. De Giusti (marisa.degiusti@ing.unlp.edu.ar)
A.C. Maria Marta Vila
A.C. Gonzalo Luján Villarreal
Universidad Nacional de La Plata, Argentina.

Resumen:

Las nuevas tecnologías han permitido la digitalización de una gran cantidad de piezas documentales que se encuentran en bibliotecas y archivos de todo el mundo, lo cual ha facilitado el acceso a un mayor número de personas a dichos documentos sin alterar o deteriorar los documentos originales.

Al ser digitalizados, los documentos son mejor preservados, pero la búsqueda y el acceso a la información allí contenida es un proceso lento, secuencial y por consiguiente, altamente ineficiente. Este problema de acceso a tales documentos presenta dos posibles soluciones:

– la catalogación de los documentos

– su indexación automática

La primera es una solución factible pero que demanda una gran cantidad de tiempo, ya que el proceso de catalogación requiere gran esfuerzo y dedicación. Sumado a esto, es necesaria la intervención de seres humanos durante todo el proceso de catalogación, lo que se traduce en altos costos económicos.

La segunda solución propone el desarrollo de una aplicación capaz de interpretar esos documentos, procesándolos y reconociendo automáticamente los textos que cada documento contiene, que posteriormente pueden ser indexados utilizando base de datos, y permitiendo un rapido acceso a la información. Esto implica la utilización de sistemas de reconocimiento de patrones, asi como también una gran cantidad de técnicas de procesamiento de imagenes digitales. Esta solución no solo aumentará los tiempos de acceso a los datos, sino que se traduce en una notoria reducciónde costos y de tiempo.

Este proyecto, llevado a cabo en PrEBi – UNLP, esta centrado en el reconocimiento de la escritura manuscrita, lo cual agrega un nivel de complejidad mucho mayor, debido a las características propias de los textos manuscritos (irregularidades, deformaciones, etcétera), lo que se le suma al deterioro que han sufrido las piezas documentales con las que se trabaja.

El recococimiento de los textos manuscritos es un proceso que debe realizarse en etapas, que van desde la digitalización de los documentos, pasando por el preprocesamiento y limpieza de las imágenes, segmentación en objetos lógicos, extracción de características, reconocimiento, corrección o refinamiento mediante el uso de diccionarios, hasta llegar finalmente a la indexación de dichos documentos.

El presente trabajo abarca la etapa de extracción de características, en la cual se toman los objetos lógicos obtenidos en la etapa de segmentación y se realizan un gran número de operaciones, tanto morfológicas como espaciales, obteniendo así una gran cantidad de información que permitirá a la etapa posterior – etapa de reconocimiento – hacer coincidir los objetos lógicos con los objetos reales. Todas las características se almacenan en un vector, el cual relaciona directamente a cada una con un valor (peso), que denota su importancia, y que indica al programa "cuanta importancia debe darle a la misma". Dentro del vector se encuentran características "simples" (alto y ancho, cantidad de agujeros, posicion respecto de la línea base), y características complejas, (código de cadenas, características basadas en técnicas con ventana de muestreo)