UABDIVULGA
BARCELONA RECERCA I INNOVACIÓ

06/2012

Nuevo método de búsqueda de palabras en imágenes y manuscritos

Investigadores del Centro de Visión por Computador y del Departamento de Ciencias de la Computación de la UAB proponen un nuevo método que permite buscar palabras en colecciones digitales de documentos sin tener que utilizar un software de reconocimiento óptico de caracteres(OCR). Las principales ventajas del método son que se puede aplicar tanto en documentos manuscritos como mecanografiados, es robusto a degradaciones de los documentos, y no está limitado al alfabeto latino sino que se puede aplicar a cualquier sistema de escritura. Este trabajo obtuvo el premio al mejor artículo presentado en el undécimo congreso internacional sobre el análisis y el reconocimiento de documentos celebrado en Pekín en Septiembre de 2011.

En las últimas décadas, millones de documentos antiguos han sido digitalizados para poder preservar los contenidos. Sin embargo, consultar estos documentos, donde cada página está almacenada como una imagen, tiene ciertas limitaciones como por ejemplo la imposibilidad de realizar búsquedas. Un usuario que busque un contenido concreto, tendrá que "hojear" estas colecciones digitales hasta encontrarlo. Para incrementar la usabilidad de las colecciones digitales es imprescindible poder reconocer el texto escrito en las imágenes para luego poder

hacer búsquedas. Esto es lo que está haciendo Google dentro de su proyecto Google Books con su gran fondo de documentos digitalizados. Pero, los programas de OCR que "leen" automáticamente texto dentro de las imágenes, tienen sus limitaciones y no funcionan bien cuando se trata de documentos antiguos o manuscritos ya que transforman la imagen en cuestión en un formato de texto y en este realizan la búsqueda.

Para tratar este problema, surge lo que se llaman técnicas de word spotting, que permiten realizar búsquedas de palabras sin tener que reconocerlas explícitamente ya que la búsqueda se realiza directamente en la imagen. El usuario debe dar al sistema una muestra de la o las palabras que le interesa buscar y el sistema devuelve una lista de posiciones dentro de los documentos donde aparecen estas palabras.

Figura 1: Uno de los investigadores recogiendo el premio al mejor artículo presentado en el 11th International Conference on Document Analysis and Recognition..

Investigadores del Centro de Visión por Computador y del Departamento de Ciencias de la Computación de la UAB propusieron un nuevo método de word spotting que no necesita ningún tipo de pre-proceso de las imágenes y que funciona tanto para texto manuscrito como mecanografiado. Además. este nuevo método no segmenta las imágenes, lo que permite las búsquedas en alfabetos distintos del latín, pudiéndose utilizar pues con documentos escritos en cualquier idioma.

Dado un ejemplo de la palabra a buscar, el sistema describe la forma de la palabra y busca similares dentro de la colección. El método propuesto se ha probado con una colección de cartas manuscritas de George Washington, un libro mecanografiado editado en 1825 y otro libro del año 1848 escrito en persa.

Este trabajo obtuvo el premio al mejor artículo presentado en el 11th International Conference on Documento Analysis and Recognition, el principal congreso dentro del ámbito del análisis de documentos.

Marçal Rossinyol i David Aldavert
marcal@cvc.uab.es aldavert@cvc.uab.cat

Referencias

M. Rusiñol, D. Aldavert, R. Toledo and J. Lladós. Browsing Heterogeneous Document Collections by a Segmentation-free Word Spotting Method. In Proceedings of the Eleventh International Conference on Document Analysis and Recognition, ICDAR11, pages 63-67, Beijing, September 2011.

View low-bandwidth version