Nou mètode de cerca de paraules en imatges i manuscrits

06/2012 - **Telecomunicacions, Electrònica i Informàtica.** **#Investigadors del Centre de Visió per Computador i del Departament de Ciències de la Computació de la UAB proposen un nou mètode que permet cercar paraules en col·leccions digitals de documents sense haver d'utilitzar un programari de reconeixement òptic de caràcters(OCR). Els principals avantatges del mètode són que es pot aplicar tan en documents manuscrits com mecanografiats, és robust a degradacions dels documents, i no està limitat a l'alfabet llatí sinó que es pot aplicar a qualsevol sistema d'escriptura. Aquest treball va obtenir el premi al millor article presentat a l'onzè congrés internacional sobre l'anàlisi i el reconeixement de documents celebrat a Pequín al Setembre del 2011.**

#En les últimes dècades, milions de documents antics han estat digitalitzats per poder-ne preservar els continguts. Tanmateix, consultar aquests documents, on cada pàgina està emmagatzemada com una imatge, té certes limitacions com per exemple la impossibilitat de realitzar cerques. Un usuari que busqui un contingut concret, haurà de col·leccions digitals fins a trobar-lo. Per tal d'incrementar l'usabilitat de les col·leccions digitals és imprescindible poder reconèixer el text escrit en les imatges per després poder fer-ne cerques. Això és el que està fent Google dins del seu projecte Google Books amb el seu gran fons de documents digitalitzats. Però, els programes d'OCR que "llegeixen" automàticament text dins de les imatges, tenen les seves limitacions i no funcionen bé quant es tracta de documents antics o manuscrits ja que transformen l'imatge en qüestió en un format de text i en aquest realitzen la cerca.
Per tal de tractar aquest problema, sorgeix el que s'anomenen tècniques de word spotting, que permeten fer cerques de paraules sense haver-les de reconèixer explícitament ja que la cerca es fa directament a l'imatge. L'usuari ha de donar al sistema una mostra de la o les paraules que li interessa buscar i el sistema retorna una llista de posicions dins dels documents on hi apareixen aquestes paraules.

Un dels investigadors recollint el premi al millor article presentat en el 11th International Conference on Document Analysis and Recognition.

Investigadors del Centre de Visió per Computador i del Departament de Ciències de la Computació de la UAB van proposar un nou mètode de word spotting que no necessita cap mena de pre-proces de les imatges i que funciona tant per text manuscrit com mecanografiat. A més. aquest nou mètode no segmenta les imatges, fet que permet les cerques en alfabets diferents del llatí, podent-se utilitzar doncs amb documents escrits en qualsevol idioma.
Donat un retall d'exemple de la paraula a cercar, el sistema descriu la forma de la paraula i en busca de similars a dins de la col·lecció. El mètode proposat s'ha provat amb una col·lecció de cartes manuscrites d'en George Washington, un llibre mecanografiat editat en 1825 i un altre llibre de l'any 1848 escrit en persa.
Aquest treball va obtenir el premi al millor article presentat al 11th International Conference on Document Analysis and Recognition, el principal congrés dins de l'àmbit de l'analisi de documents.

Marçal Rossinyol i David Aldavert

Centre de Visió per Computador

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.