Resum

Habitualment, els sistemes de reconeixement de patrons consisteixen en dos grans apartats. D'una banda, l'adquisició de les dades i, de l'altre, la classificació d'aquestes dades dins d'una certa categoria. Per tal de reconèixer a quina categoria pertany un cert element, un conjunt de patrons models han d'haver estat proporcionats per anticipat. Per tal d'entrenar el classificador i que pugui oferir una classificació de patrons robusta, es necessita una etapa d'aprenentatge que es pot dur a terme fora de línia. Dins del camp de reconeixement de patrons, estem interessats en el reconeixement d'elements gràfics i, en particular, en l'anàlisi de documents rics en informació gràfica. En el cas particular del reconeixement de símbols, certs descriptors s'extreuen del símbol a reconèixer i s'emparellen posteriorment amb el conjunt de símbols model. En aquest context, una de les principals inquietuds és assegurar-se que els sistemes que es proposen romanen escalables respecte al volum de dades i que poden fer front a quantitats creixents de models. Per tal d'evitar el fet de treballar amb una base de dades de símbols de referència, s'han proposat en els últims anys els sistemes de reconeixement de símbols al vol, o els sistemes de localització aproximada de símbols gràfics coneguts com a mètodes de *symbol spotting*.

En termes generals, es pot definir el problema de *symbol spotting* com la identificació d'un conjunt de regions d'interès d'un document que puguin contenir una instància d'un determinat símbol sense haver d'aplicar explícitament tot el procés de reconeixement de patrons. El nostre marc d'aplicació consisteix en la indexació d'una col·lecció de documents gràfics. Aquesta col·lecció es consulta donant un exemple del que es vol trobar, és a dir, amb una única instància del símbol a cercar i, gràcies als mètodes de *spotting* es retornen les regions d'interès dels documents on és probable trobar el símbol en qüestió. Aquest tipus d'aplicacions es coneixen com a recuperació d'informació dirigida.

Per tal que els sistemes de recuperació d'informació puguin gestionar grans col·leccions de documents necessitem proporcionar un accés eficient als grans volums d'informació que es poden emmagatzemar. Farem ús d'estratègies d'indexació que siguin capaces de retornar localitzacions on apareguin parts similars del símbol consultat. En aquest escenari, els patrons gràfics s'hauran d'emprar com a índexs afavorint l'accés i la navegació de la col·lecció de documents. Aquests mecanismes d'indexació permeten a l'usuari buscar elements semblants utilitzant informació gràfica en lloc de formular consultes textuals.

Al llarg d'aquesta tesi es presenten una arquitectura i diferents mètodes de localització aproximada de símbols, per tal de construir una aplicació de recuperació dirigida fent front a una col·leccions de documents gràfics.

S'han proposat diferents descriptors de símbols que codifiquen informació geomètrica i estructural. L'objectiu d'aquests descriptors és descriure les parts dels símbols de manera molt compacta i eficient. Signatures vectorials, cadenes amb atributs i descriptors de forma genèrics s'han fet servir per agrupar símbols per semblança.

S'han utilitzat diverses estratègies de cerca d'informació gràfica per semblança. Per tal de recuperar les localitzacions dins de la col·lecció de documents on apareixen les parts dels símbols, hem utilitzat *lookup tables* i *grid files* indexades a partir de patrons gràfics. S'ha introduït una fase final de verificació per validar les hipotètiques localitzacions on és probable que es trobi un cert símbol. Aquesta etapa de validació està formulada en termes d'informació espacial i relacional.

A més a més, es proposa un protocol per avaluar el rendiment dels mètodes de localització aproximada de símbols en termes de tases de reconeixement, precisió en la localització i escalabilitat. Es mostra que les mesures que es proposen permeten determinar els punts forts i febles dels mètodes analitzats. Totes les contribucions proposades s'han posat a prova experimentalment amb una col·lecció de planells arquitectònics, amb la corresponent base de dades de referència.

***Paraules clau**: Reconeixement de Patrons, Reconeixement de Gràfics, Descripció de Símbols, Localització Aproximada de Símbols, Recuperació d'Informació Dirigida, Indexació de Patrons Gràfics, Avaluació del Rendiment.*

Abstract

Usually pattern recognition systems consist in two main parts. On the one hand, the data acquisition and, on the other hand, the classification of this data on a certain category. In order to recognize which category a certain query element belongs to, a set of pattern models must be provided beforehand. An off-line learning stage is needed to train the classifier and offer a robust classification of the patterns. Within the pattern recognition field, we are interested in the recognition of graphics and, in particular, on the analysis of documents rich in graphical information. In the particular case of graphical symbol recognition, descriptors are extracted from the symbol to recognize and are subsequently matched with the set symbol models. In this context, one of the main concerns is to see if the proposed systems remain scalable with respect to the data volume so as it can handle growing amounts of symbol models. In order to avoid to work with a database of reference symbols, symbol spotting and on-the-fly symbol recognition methods have been introduced in the past years.

Generally speaking, the symbol spotting problem can be defined as the identification of a set of regions of interest from a document image which are likely to contain an instance of a certain queried symbol without explicitly applying the whole pattern recognition scheme. Our application framework consists on indexing a collection of graphic-rich document images. This collection is queried by example with a single instance of the symbol to look for and, by means of symbol spotting methods we retrieve the regions of interest where the symbol is likely to appear within the documents. This kind of applications are known as focused retrieval methods.

In order that the focused retrieval application can handle large collections of documents there is a need to provide an efficient access to the large volume of information that might be stored. We use indexing strategies in order to efficiently retrieve by similarity the locations where a certain part of the symbol appears. In that scenario, graphical patterns should be used as indices for accessing and navigating the collection of documents. These indexing mechanism allow the user to search for similar elements using graphical information rather than textual queries.

Along this thesis we present a spotting architecture and different methods aiming to build a complete focused retrieval application dealing with a graphic-rich document collections.

Different symbol descriptors encoding geometric and structural information are proposed in this thesis. These descriptors aim to describe parts of the symbols in a very compact and efficient way. Vectorial signatures, attributed strings and off-the-shelf shape descriptors are used to cluster parts of the symbols by similarity.

Several strategies aiming to search for graphical information by similarity are used in this thesis. In order to retrieve locations from the document collection where parts of the symbols appear we use lookup tables and grid files indexed by graphical patterns. A final validation phase is introduced to validate the hypothetic locations where a symbol is likely to be found. This validation stage is formulated in terms of spatial and relational information.

In addition, a protocol to evaluate the performance of symbol spotting systems in terms of recognition abilities, location accuracy and scalability is proposed. We show that the evaluation measures allow to determine the weaknesses and strengths of the methods under analysis. All the proposed contributions have been tested with an experimental scenario consisting of a collection of architectural drawings with its corresponding ground-truth.

Keywords*: Pattern Recognition, Graphics Recognition, Symbol Description, Symbol Spotting, Focused Retrieval, Graphical Pattern Indexation, Performance Evaluation.*