A Content Spotting System For Line Drawing Graphic Document Images

Muhammad Muzzamil Luqman*†, Thierry Brouard*, Jean-Yves Ramel* and Josep Lladós†

*Laboratoire d'Informatique, Université François Rabelais de Tours, 37200 France
†Computer Vision Center, Universitat Autónoma de Barcelona, 08193 Spain
Email: {brouard, ramel}@univ-tours.fr, {mluqman, josep}@cvc.uab.es

***Abstract*—We present a content spotting system for line drawing graphic document images. The proposed system is sufficiently domain independent and takes the keyword based information retrieval for graphic documents, one step forward, to Query By Example (QBE) and focused retrieval. During offline learning mode: we vectorize the documents in the repository, represent them by attributed relational graphs, extract regions of interest (ROIs) from them, convert each ROI to a fuzzy structural signature, cluster similar signatures to form ROI classes and build an index for the repository. During online querying mode: a Bayesian network classifier recognizes the ROIs in the query image and the corresponding documents are fetched by looking up in the repository index. Experimental results are presented for synthetic images of architectural and electronic documents.**

***Keywords*-content spotting; graphic document retrieval; query by example; fuzzy structural signature**

I. Introduction and related works

The graphic document research community has seen a gradual shift of attention over the last few years, from the hard problems of symbol recognition, segmentation and localization to the relatively softer problem of symbol spotting. An important reason behind this is the growing size of document repositories and the increasing demand from users to have an efficient browsing mechanism for graphic content. The format of these documents mainly restricts to use keyword based searching and indexing mechanisms. Thus a very interesting topic of research is to investigate into mechanisms of indexing the graphic content of these documents; in order to offer to the users, the advantages of Query By Example (QBE) and focused retrieval. The research surveys by Chhabra [1], Llados et al. [2], Cordella & Vento [3] and Tombre et al. [4] provide a detailed and state of the art historical review of work done in the field of symbol recognition over last two decades. The graphic documents are generally represented by symbolic representations based structural methods of pattern recognition. Graph in one form or another has remained a popular choice for most of the methods of symbol recognition and segmentation, because of its natural adaptation to the content of these documents, but has an associated drawback of computational inefficiency. On the other hand, the new developments in statistical pattern recognition offer highly efficient mathematical tools for learning and classification.

Fonseca et al. [5] have presented a detailed review of content based retrieval of technical drawings. Some of the notable recent works for symbol spotting include : a region string based method in [6], a method based on graph representations and vectorial signature [7], a chain point dendrogram based approach by [8] and a shape context descriptor based approach in [9]. The PhD dissertations of Rusinol [10] and Nguyen [11], in recent past, are good contributions to the literature on symbol spotting.

We are more interested in investigating into graph based representations for symbol spotting and have selected a method from Qureshi et al. [12] for our work. This system is based on a graph based structural approach. First, it vectorizes the image into a set of quadrilateral primitives, extracts topological and geometric features and represents the image content by an attributed relational graph (ARG). The nodes of the graph are the quadrilateral primitives and arcs are the relationships between these primitives. Nodes of graph have relative lengths and arcs have relative angle and relation type as their attributes. In the second step, the system looks for potential ROIs corresponding to symbols. It detects parts of the ARG that may correspond to symbols i.e. symbol seeds. Scores corresponding to probabilities of being part of a symbol are computed for all edges and nodes of the ARG. They are based on features such as lengths of segments, perpendicular and parallel angular relations, degrees of nodes etc. The symbol seeds are detected during a score propagation process. This process seeks and analyzes the different shortest paths and loops between nodes in the ARG. To obtain the symbols seeds, the scores of all the nodes belonging to a detected path are homogenized i.e. propagation of the maximum score to all the nodes in the path until convergence. And finally they employ a greedy algorithm for sub-graph matching. The system achieves good localization and spotting results. The results of this system have also been evaluated by Delalandre et al. [13], where the authors have concluded that this method offers high confidence detection results without any multiple detections but lacks in precision of localization results. We argue that a content spotting and document retrieval system should offer a high recall rate and low precision automatically becomes tolerable. The underlying sub-graph matching algorithm restricts this method to scale to huge document repositories.

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.835

IEEE computer society

Our main motivation behind this work is to propose that the two decades of research in the symbol recognition and segmentation for graphic documents could be adapted to be processed by statistical pattern recognition tools; for addressing the problem of symbol spotting and graphic document browsing and retrieval. We detail a system for content spotting in line drawing graphic document images. The proposed system exploits the representation power of structural approaches of pattern recognition and the computational efficiency of clustering and classification tools from statistical pattern recognition. This work is a continuation of our past works on symbol recognition [14] [15]. In [15] we have proposed a method to embed an ARG into a feature vector; which we term as fuzzy structural signature and which has been shown capable of encapsulating the structure of image content represented by the ARG. The signature was previously proposed to describe graph based representations of presegmented symbols by feature vectors (or signatures). Such a signature comes with a very important advantage of enabling graph based representations to access the whole range of feature vector based statistical classifiers and machine learning tools. The use of fuzzy intervals while designing the structure signature incorporates uncertainty management into the system and helps to achieve graceful degradation in noisy environments. In this work, we take forward the work of Qureshi et al. [12] and use our fuzzy structural signature from [15] along with efficient clustering and classification tools to develop a complete system for document retrieval and content spotting, which offers (unsupervised) automatic learning and indexation abilities, the ease of QBE and granuality of focused retrieval.

II. The proposed system

A. Offline learning mode

The system starts with a repository of line drawing document images, automatically learns the symbol classes in the documents and builds an index for the repository. The fact that the learning process is fully automatic and does not require any manual labeling or keyword assignment by human operators, is a very important contribution of our work. In first step, the system uses the method of Qureshi et al. [12] to mark probable ROIs in the documents. A ROIs could ideally contain only one symbol but may also correspond to a part of symbol, a complete symbol with a part of another symbol (context noise) or even more than one symbols : the system has been designed to incorporate this behavior. Once the ROIs have been extracted from the documents, we use our previously proposed method of fuzzy structural signatures [15] to describe each ROI by a signature.

After describing all the ROIs by fuzzy structural signatures, we deploy an agglomerative (or hierarchical) clustering method from Statistics Toolbox of Matlab [16]. The clustering process starts with computing a pairwise city

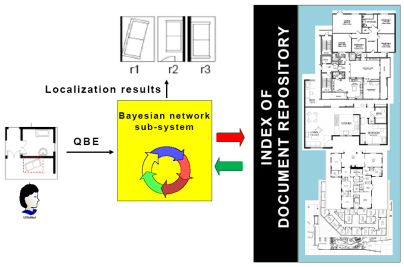

Figure 1. The proposed system.

distance metric for the features in fuzzy structural signature and builds a linkage tree using the single link algorithm. We use a method from [17] for getting an optimal cutoff for clusters. This method is based on an econometric approach to verify that variables in multiple regression are linearly independent. We have selected the agglomerative clustering approach, in order to keep our system independent of requiring any specific details about the type or number of symbol models in underlying documents; the system has been designed to work well for all type of line drawing document images. We cluster the fuzzy signatures of ROIs from documents and assign class labels to cluster members. Meanwhile an index for the document repository is also built. This index contains the details about ROIs, their locations in documents and the labels of the clusters to which they are assigned. We maintain this index in a database management system (DBMS) and have used MySQL [18] for our implementation.

As the last step for offline learning mode, we learn a Bayesian network from the fuzzy structural signatures of ROIs from documents. First the features in the signature are quantized by using a histogram based binning technique which uses Akaike's information criterion (AIC) as cost function [19]. Then we learn the structure of the Bayesian network using the quite recently proposed genetic algorithms by Delaplace et al. [20]. And finally the parameters of the network, which are the conditional probability distributions associated to its nodes, are learned by using Maximum Likelihood Estimates (MLE) from [21]. This Bayesian network will serve as a window to the document index for retrieving relevant documents from the repository, during the online querying phase.

B. Online querying mode

The system is based on Query By Example (QBE) information retrieval mechanism. The user selects a complete document image or defines region in the image : there isn't any restriction on the number of symbols in the query image. The query may contain one or more complete or partial symbols. Once the query is posed to the system, it first marks the ROIs in the query image, then it computes fuzzy structural signatures for each of these ROIs and poses a set

of classification queries to the Bayesian network. Given the prior probabilities and the evidence, the Bayesian network performs probabilistic inference by deploying Bayes rule (Eq.1) to compute posterior probabilities for all possible cluster labels. We use these posterior probabilities as a measure of similarity for deciding about the probable membership of the query signature to one or more clusters i.e. the confidence or rank of retrieving the documents corresponding to the ROIs in that cluster.

$$\Pr(c_i|e) = \frac{\Pr(e, c_i)}{\Pr(e)} = \frac{\Pr(e|c_i) \times \Pr(c_i)}{\Pr(e)} \tag{1}$$

where,

$$e = f1, f2, f3, ..., f16 \tag{2}$$

$$\Pr(e) = \Pr(e, c_i) = \sum_{i=1}^{k} \Pr(e|c_i) \times \Pr(c_i) \tag{3}$$

After receiving the ranks of the cluster labels for all the ROIs in the query image, the system fetches the associated documents for corresponding clusters from the document repository using the index (built during offline learning). The posterior probabilities or ranks are used by our system for defining the threshold for making a decision for the inclusion of a retrieved document in the final results. Lower threshold on rank means the system is returning more documents and could lead to a high recall performance. Whereas high threshold on rank leads to very precise results but the recall rate drops. Generally a threshold of 0.5 gives acceptable results, however the system could retrieve all the results and could sort them to show more relevant results on top. An important contribution of this work is the ability of the system to highlight the search results in the documents i.e. focused retrieval. Instead of returning atomic results and leaving it on the user to spot the required information himself in the retrieved documents, focused retrieval enables our system to return precise and accurate regions in the results. This allows the user to browse in the document repository using visual queries (as in Figure 1).

III. Experimentation

We have evaluated the performance measures for our system on two document repositories. The line drawing graphic document images and the queries are generated synthetically along with the ground truth using [22]. These datasets are available online at [23] and the details of the datasets are give in Table I.

Experimentation for both of the datasets was done by using synthetically generated queries consisting of cropped regions of documents with three different levels of context noise. Since we have focused on the document retrieval aspect of the problem, we have used the standard performance measures of precision and recall for evaluating the performance of our system. However, a detailed discussion on performance evaluation of spotting systems is recently compiled by [24], where the authors have detailed the list of performance measures that are interesting for symbol spotting systems. During learning phase our system detected a total of 10285 ROIs in electronic diagrams and 4586 ROIs in floorplans, which approximately corresponds to 108% of the symbols in each of the datasets. Figure 2 presents the results of our experimentation and compares them with the results of [12]. The precision-recall curve of the proposed system shows the ability of the system to achieve a constant high recall rate with an acceptable precision rate. An overall high recall rate clearly shows the ability of our system to retrieve all the relevant documents. The reason for low precision rates, for some of the queries, is the inability of the localization method to find ROIs in query images with high context noise. The system has successfully succeeded in improving the performance of the underlying localization method by allowing (unsupervised) automatic learning and real time querying facilities.

Table I
Dataset details.

Electronic diagrams				
	Document		Symbol	
Settings	Backgrounds	8	Models	21
Dataset	images	800	Symbols	9600
Floorplans				
	Document		Symbol	
Settings	Backgrounds	2	Models	16
Dataset	images	200	Symbols	4216

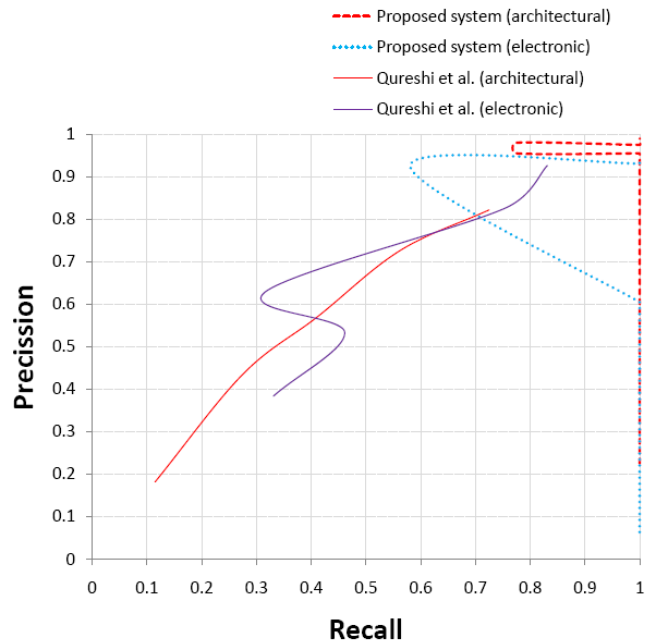

Figure 2. Experimental results for posterior probability (rank) ≥ 0.5.

IV. Conclusion and future work

We have presented our proposed system for content spotting and document retrieval in line drawing graphic document image repositories. The system hybrids the representational strength of structural methods of pattern recognition and the computational efficiency of statistical tools of machine learning. The important contributions of our work include the incorporation of the learning abilities into structural representations without requiring any labeled learning set : which, in fact, is a very important issue in graphic document research. The system offers the advantages of Query By Example (QBE) and focused retrieval. Moreover, the system does not impose any restriction on the number of symbols in query image and the user has the liberty to select any region from any document image for browsing the document repository - visual queries are more natural way of searching in graphic document repositories. Our main motivation behind this work is not only to study hybrid approaches of structural and statistical methods for graphics recognition but also to take existing structural methods forward to the statistical domain so that the last two decades of research on structural approaches of graphics recognition does not go in vain and these methods may also take advantage of the state of the art machine learning tools and techniques. The result are very encouraging and in future we plan to take this work forward by incorporating fuzzy graph mechanism for spotting ROIs, so that the system becomes independent of the external methods for localization of ROIs.

Acknowledgment

This research was supported in part by PhD grant PD-2007-1/Overseas/FR/HEC/222 from Higher Education Commission of Pakistan.

References

[1] A. K. Chhabra, "Graphic symbol recognition: An overview," *LNCS*, vol. 1389, pp. 68–79, 1998.

[2] J. Llados, E. Valveny, G. Sanchez, and E. Marti, "Symbol recognition: Current advances and perspectives," *LNCS*, vol. 2390, pp. 104–128, 2002.

[3] L. P. Cordella and M. Vento, "Symbol recognition in documents: a collection of techniques?," *IJDAR*, vol. 3, no. 2, pp. 73–88, 2000.

[4] K. Tombre, S. Tabbone, and P. Dosch, "Musings on symbol recognition," in *GREC*, vol. 3926 of *LNCS*, pp. 23–34, Springer, 2006.

[5] M. J. Fonseca, A. Ferreira, and J. A. Jorge, "Content-based retrieval of technical drawings," *IJCAT*, vol. 23, no. 2-4, pp. 86–100, 2005.

[6] M. Rusiñol, J. Lladós, and G. Sánchez, "Symbol spotting in vectorized technical drawings through a lookup table of region strings," *PAA*, 2009.

[7] M. Rusinol and J. Llados, "Symbol spotting in technical drawings using vectorial signatures," in *GREC*, vol. 3926 of *LNCS*, pp. 35–46, Springer, 2006.

[8] S. Tabbone and D. Zuwala, "An indexing method for graphical documents," in *ICDAR*, pp. 789–793, 2007.

[9] T.-O. Nguyen, S. Tabbone, and O. R. Terrades, "Symbol descriptor based on shape context and vector model of information retrieval," in *DAS*, pp. 191–197, 2008.

[10] M. Rusinol, *Geometric and Structural-based Symbol Spotting. Application to Focused Retrieval in Graphic Document Collections*. PhD thesis, Universitat Autonoma de Barcelona, 2009.

[11] T. O. Nguyen, *Localisation de symboles dans les documents graphiques*. PhD thesis, Université Nancy 2, 2009.

[12] R. J. Qureshi, J.-Y. Ramel, D. Barret, and H. Cardot, "Spotting symbols in line drawing images using graph representations," in *GREC*, pp. 91–103, 2007.

[13] M. Delalandre, J.-Y. Ramel, E. Valveny, and M. M. Luqman, "A performance characterization algorithm for symbol localization," in *GREC*, vol. 8, pp. 3–11, 2009.

[14] M. M. Luqman, T. Brouard, and J.-Y. Ramel, "Graphic symbol recognition using graph based signature and bayesian network classifier," in *ICDAR*, vol. 10, pp. 1325–1329, IEEE Computer Society, 2009.

[15] M. M. Luqman, M. Delalandre, T. Brouard, J.-Y. Ramel, and J. Llados, "Employing fuzzy intervals and loop-based methodology for designing structural signature: an application to symbol recognition," in *GREC*, vol. 8, pp. 22–31, 2009.

[16] http://www.mathworks.com/ (Accessed on 25.01.2010).

[17] Y. Okada, T. Sahara, S. Ohgiya, and T. Nagashima, "Detection of cluster boundary in microarray data by reference to mips functional catalogue database,," in *GIW2005*, 2005.

[18] http://www.mysql.com (Accessed on 25.01.2010).

[19] O. Colot, C. Olivier, P. C., and E. M. A., "Information criteria and abrupt changes in probability laws," in *Signal Processing VII : Theories and Applications*, pp. 1855–1858, 1994.

[20] A. Delaplace, T. Brouard, and H. Cardot, "Two evolutionary methods for learning bayesian network structures," in *CIS*, vol. 4456 of *LNCS*, pp. 288–297, Springer, 2006.

[21] P. Leray and O. François, "BNT structure learning package : Documentation and experiments," tech. rep., Laboratoire PSI - INSA Rouen- FRE CNRS 2645, Nov. 15 2004.

[22] M. Delalandre, T. Pridmore, E. Valveny, H. Locteau, and E. Trupin, "Building synthetic graphical documents for performance evaluation," in *GREC*, vol. 5046 of *LNCS*, pp. 288–298, Springer, 2007.

[23] http://mathieu.delalandre.free.fr/projects/sesyd/index.html.

[24] M. Rusiñol and J. Lladós, "A performance evaluation protocol for symbol spotting systems in terms of recognition and location indices," *IJDAR*, vol. 12, no. 2, pp. 83–96, 2009.