Volume 4, No. 4, April 2013
Journal of Global Research in Computer Science
ISSN-2229-371X
REVIEW ARTICLE
Available Online at www.jgrcs.info

COGNITIVE WAY OF CLASSIFYING DOCUMENTS: A PRACTITIONER APPROACH

Muheet Ahmed Butt*, Majid Zaman**
*Scientist, Directorate of IT & SS, University of Kashmir, Srinagar, J&K, India
**Scientist, PG Department of Computer Science, University of Kashmir, J&K, India
ermuheet@gmail.com, zamanmajid@rediffmail.com

Abstract: Documents are the sources of data which result in information and knowledge. Processing documents to extract their contents in an automated fashion is an essential task in all types of organizations for varied applications. The classification of documents being processed is required for their efficient recognition as it reduces the number of searches and also reduces the chances of error at different stages during the process. Therefore, in this proposed research a robust classification mechanism for document images based on the layout structure of its different elements which results in cognition based recognition is presented. The document image is considered to contain text only or text as well as tables and images. The classification is based on a scheme of preserving the structure of the layout of a document image. The algorithms are based on the spatial relationships existing among the visual components present in the document.

INTRODUCTION

Document processing in any organization whether having its operations manual or computerized, forms an essential activity in its functioning's. Within document processing, the key activity prior to all other activities is the recognition of documents and hence their categorization [1] [7] [10]. Human vision system achieves instantaneous recognition of documents without contextual analysis. This is primarily accomplished by recognizing, in a glance, the appearance of a document, the key distinguishing features within the document image and their spatial relationship with each other, rather than actually analyzing the content. Further, from one domain to another and one observer to another, the areas of interest and the priority of their importance within a document image may vary [11][12][13]. As for example, when an assistant within an office set-up of an organization tries to achieve classification of incoming documents, the prominent feature of interest might be the logo of the document concerned, so as to decide whether the document is from within or outside the organization and the addressee of the document. The administrator of the department while viewing the same document has contents as the more important area of interest

The conventionally applied, structured approach to document segmentation, which tries to achieve segmentation of the document in a rigid vertical and horizontal or top to bottom fashion uses the system resources extensively. A generalized approach to segmentation, without considering the specific domain and the particular context in which classification is being attempted, is bound to be more time consuming and give extraneous and irrelevant information and also lead to a higher rate of errors. [5]. Since the context in which a document is viewed and the areas of interest are inevitably linked, a more intelligent approach would be to achieve segmentation assigning priority to the areas of interest based on the context in which the document is being considered in the particular domain. This is bound to be faster and give more accurate results with the percentage of errors considerably minimized.

THE PROPOSITION

Consistent with the concept of how a human observer's gaze shifts from one area of interest to another, as per priority of importance/interest, our proposition here is to adopt a similar sequence for segmentation of regions of interest. In a domain specific study, with a pre-decided context of classification, the regions of interest and their priorities can be decided in advance and the outcome of segmentation obtained, not as a physical structuring of the document image but a typical layout of key segments whose identification, in order of priority, may lead to classification in a few or even one or two passes.

In the particular study under consideration, the incoming documents to a particular teaching department in a university are being considered. In this context, documents are first classified as being from within the organization or from an outside agency. Hence, from the particular classifiers viewpoint, two major areas of interest would be a logo/emblem (if present) on the document and the organization name. For both these areas of interest, the typical spatial positions are known, as both of these are assured to be at the top of the document image. Further sub-classification can be achieved by obtaining the particular spatial arrangements of the features of interest, like addressee, sender, subject line, heading/title, main text body, copy to block, etc.

A human observer accomplishes instant recognition of documents by taking into consideration appearance features like the size of a document, texture/color of the paper in addition to key segments like emblem, title, addressee, etc. and their relative spatial arrangements with respect to one another [5][6][7]. When considered prior to segmentation these considerations may lead to immediate differentiation between major document categories as for instance between an invitation card, an official document, a newsletter/magazine/journal and postal letters.

FRAMEWORK

The Sequence of Steps:

The sequence of steps as depicted in Fig1 is proposed here. Any document under processing is subjected to pre-processing, followed by the extraction of appearance based features first to achieve a document image categorization technique, aiming at a higher rate of success and a reduction in error rate, by utilizing the domain and context information [8][9][10]. In this case, these features are the dimensions of the document. Immediately after this step a preliminary broad classification based on these features is attempted. Refer to Fig 2, the domain specific schema, whereby it is apparent that the physical dimensions of a document can immediately help us differentiate between an official document (e.g. letter/order/notification, all of which are printed on standard sized pages), Invitations (which have different dimension information), magazines/ journals/newsletters (which have a separate set of dimension) and letters (again the envelope dimensions are different from the previous categories). Followed by this preliminary classification, the segmentation algorithm appropriate to any one of the above categories is applied, extracting the key features applicable to that particular class. In the next stage, the key features of the incoming document are compared to the key features of the previously stored class samples to arrive at a decision about the class with which the incoming document has highest resemblance. The next phase assigns class label to the document from amongst the various labels available in within the knowledge base. This assignment is a direct result of the results of the previous step. Hence, the output so obtained is a document with a class label assigned to it.

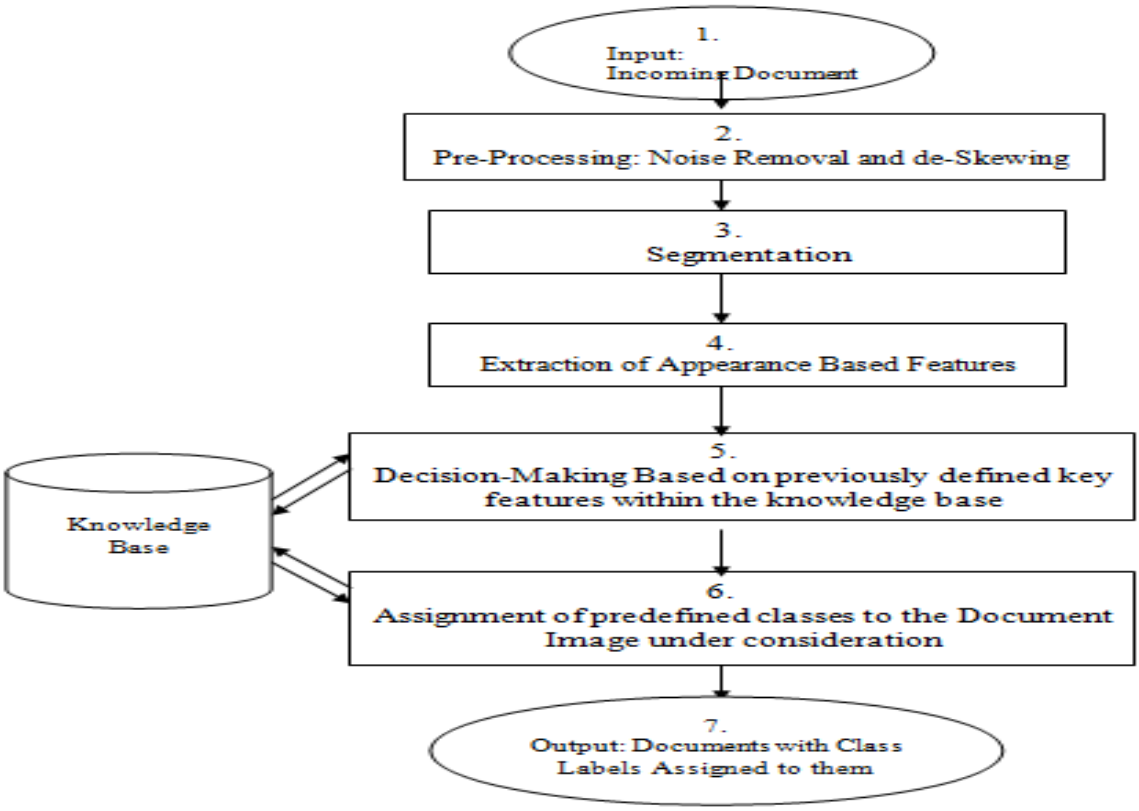

Figure.1 A hierarchical framework for cognition based document categorization

The Domain Specific Classification Schema

In the present study, the following basic classes of the documents have been identified and hence it is assumed that an incoming document has to be assigned to one of these classes, depending upon the value of a measure of nearness with a class. This value is obtained by a comparison of the extracted discriminating features of the document under study, with the features of the class prototypes stored within the knowledge base. The sample is assigned to the class with which the highest value of the measure of nearness is obtained.

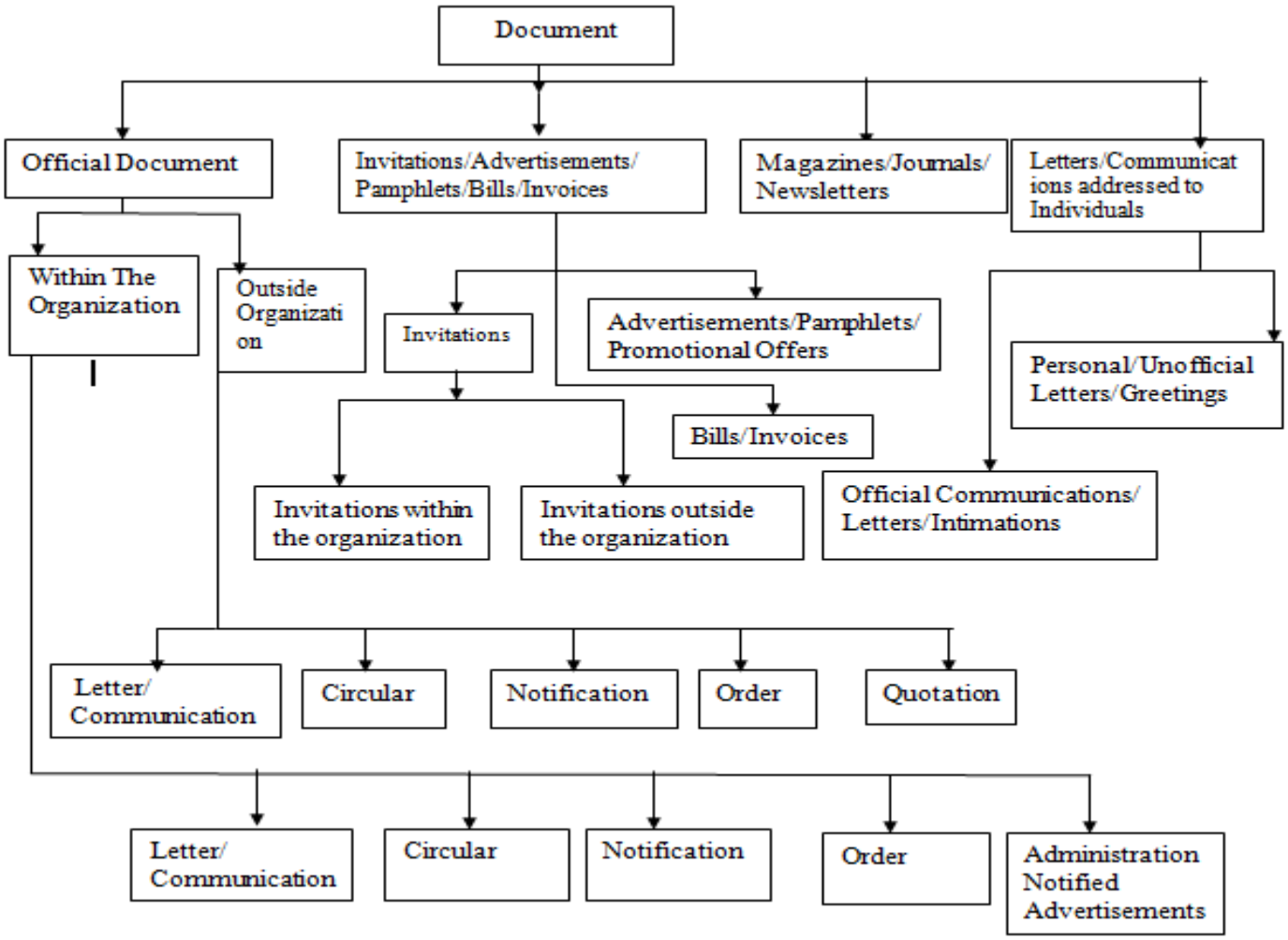

Figure: 2 Domain Specific document classification, for the present study

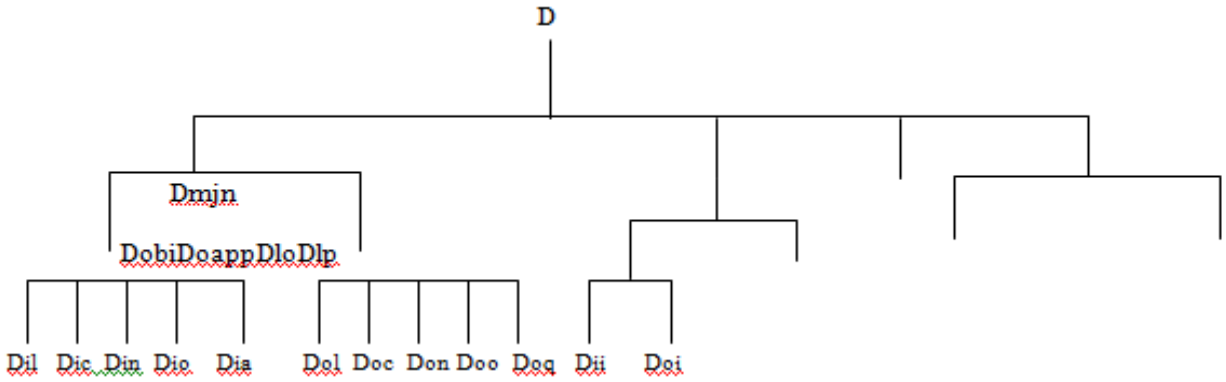

Figure 3. A dendogram representation of the Domain Specific Classes and the class label descriptions

Assume that within a particular domain, the total number of document types is 'n', then
The number of classes = n,
The class prototypes are represented as:
D1, D2, D3 ….Dn.
Each class prototype is a set of similar documents.
For instance, in the present study, the major classes are:

D= {D1, D2, D3, D4}

With D1= {Doi, Doo}
D2= {Din,Dbi,Dapp}
D3= {Dmnj}
D4={Dlo,Dlp}

The overall classification scheme would be represented as:

{((Dil,Dic,Din,Dio,Dia),(Dol,Doc,Don,Doo,Doq)), ((Diin,Doin),Dbinv,Dapp), Dmjn, (Dlo,Dlp) }

The first subscript '***i*** 'indicates a document from ***inside*** the organization

The first subscript '***o***' indicates a document from ***outside*** the organization

The second subscript ***'l, c, n, o'*** could indicate a ***letter, circular, notification, order***, each of these could belong to one of the above super sets.

The second subscript '***in***' indicates an ***invitation***,

The second subscript '***bi***' indicates ***bill or invoice***

The second subscript '***app***' indicates ***ad, pamphlet or promotional offer***

The subscript '***mnj 'indicates magazine,newsletter, journal***
The first subscript 'l' indicates a ***letter***, the second subscript '***o***' or '***p***' indicates ***official*** or ***personal.***

CONCLUSION

The processing of documents for the purpose of discovering knowledge from them in an automated fashion is a challenging task and hence an open issue for the cognitive research community. As discovery of knowledge from documents can be achieved efficiently once the class of a document is known in advance. Moreover, as classification is a step prior to knowledge extraction, hence achieving classification on the basis of global appearance based features, and specifically by preserving the layout structure of document images is naturally a logical approach.

REFERENCES

[1]. Michelangelo Diligenti, Paolo Frasconi, Marco Gori, "Hidden Tree Markov Models for Document Image Classification", IEEE Transactions on pattern analysis and machine intelligence, vol 25, no 4, pages 519-523, 2003. Available at

[2]. Antonacopoulos, A., Gatos, B., Bridson, D.: ICDAR 2005 page segmentation competition.In: Proc. ICDAR, Seoul, Korea (2005) 75–80

[3]. H. S. Baird, M. A. Moll. J. Nonnemaker, M.R.C., Delorenzo, D.L.: Versatile document image content extraction. In: Proc., SPIE/IS&T Document Recognition & Retrieval XII Conf., San Jose, CA (2006)

[4]. E. Appiani, F. Cesarini, A.M. Colla, M. Diligenti, M. Gori, S. Marinai, G. Soda, "Automatic document classification and indexing in high-volume applications", International Journal on Document Analysis and Recognition, vol 4, no 2, pages 69-83, 2001.

[5]. Kevyn Collins-Thompson and RadoslavNickolov, "A Clustering-Based Algorithm for Automatic Document Separation", Proceedings of the SIGIR 2002 Workshop on Information Retrieval and OCR, Available at

[6]. Thomas Breuel, "High Performance Document Layout Analysis", 2003 Symposium on Document Image Understanding Technology Greenbelt Marriott, Greenbelt Maryland,

[7]. Thomas M. Breuel, "An Algorithm for Finding Maximal Whitespace Rectangles at Arbitrary Orientations for Document Layout Analysis", Proceedings of the ICDAR 2003, vol 1, pages 66-70, 2003.

[8]. Thomas M. Breuel, "Two Geometric Algorithms for Layout Analysis", Proceedings of Document Analysis Systems V, 5th International Workshop, DAS 2002, vol 2423, pages 188-199.

[9]. P. Duygulu, V. Atalay, "A Hierarchical Representation of Form Documents for Identication and Retrieval", International Journal on Document Analysis and Recognition, vol5, no 1, pages 17-27, 2002.

[10]. Song Maoa, AzrielRosenfelda, Tapas Kanungob, "Document Structure Analysis Algorithms: A Literature Survey", 2003,

[11]. S Diana, E Trupin, F Jousel, J Lecoutier and J Labiche, From Acquisition tpModelisation of a Form Base to Retreive Information.

[12]. George R. Thoma, Glenn Ford, "Automated data entry system: performance issues", Proceedings of SPIE vol. 4670, 2002.

[13]. Zhixin Shi, VenuGovindaraju, "Skew Detection for Complex Document Images Using Fuzzy Runlength", Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol 2, pages 715-720, 2003.

