A KEYWORD SPOTTING APPROACH USING AGGLOMERATIVE HIERARCHICAL CLUSTERING METHOD

[1]SUSHMA S N, [2]SHARADA B

[1]Department of Studies in Computer Science, University of Mysore, Karnataka, India
[2]Department of Studies in Computer Science, University of Mysore, Karnataka, India
E-mail: [1]sushisn2007@gmail.com, [2]sharadab21@gmail.com

Abstract -
A load of vital historical document gatherings are lying quiescent in National libraries that needs to conserved in digital form. The problem of information retrieval from historical handwritten scripts is challenging task due to its variableness in writing styles. Retrieval of relevant information from these scanned historical document images would ideally necessitate an efficient Keyword spotting system which can be replaced with recognition systems. We propose a cluster based keyword spotting system in which agglomerative hierarchical clustering techniques are used for matching the handwritten word images. The technique is demonstrated using historical handwritten Devanagari document images. Compared to prior results, the proposed method gives favorable results.

Keywords - Keyword Spotting, Agglomerative Hierarchical Clustering, Single Linkage, Complete Linkage

I. INTRODUCTION

In today's world, in spite of the electronic media, a huge quantity of handwritten information is available in National libraries which are precious and need to conserve. George Washington's papers present in Library of Congress and Isaac Newton's papers present in the University of Cambridge Library [1, 2] are the some examples for valuable handwritten historical documents. There is a rising mandate from researchers of different discipline to access this valuable information. So digitization of historical manuscripts is required which helps in automation of information retrieval. Automatic recognition and transcription of degraded historical manuscripts is an extensively unsolved problem. The objective of our research is to advance a complete approach for automatic retrieval of pool of word images of unconstrained historical documents based on the occurrence of one or several keywords which can be indicated by the user. An Optical Character Recognition (OCR) is suitable for machine printed documents, but it will poorly work with handwritten documents. Information retrieval from historical documents belongs to the field of document image analysis research field. The document image processing that includes automation of information extraction, classification, search, and retrieval of documents is in demand. Complexity lies in handling with historical documents that are noisy and unstructured. There are various challenges in processing handwritten historical variations such as poor quality of images which contains lot of noises and distortions. Handwritten document image retrieval methods can be widely classified into two categories namely: template-based and recognition-based retrieval methods. Template-based retrieval compares image data with group of template images. Whereas, recognition-based retrieval uses OCR correction methods and modified information retrieval methods. Recognition-based methodology can use either holistic or non-holistic approach for word recognition.

Instead of absolute transcription, keyword spotting system can be used, which is a content-based retrieval technique of spotting all similar word image of a given query word image from a document. Keyword spotting is a recognition-free retrieval approach where all instances of query image is spotted in a set of document images using either segmentation based or segmentation free method. In segmentation based approach the document image is segmented into lines and words. Whereas in segmentation free approach, the document image is divided into patches and similar words are spotted without segmentation using sliding window concept. Based on the representation of the query, there exist two foremost approaches of keyword spotting. In Query-by-string (QBS) [3], an arbitrary string is the input which requires large amount of training material and in Query-by-example (QBE) [4], an exemplary image is the input (image matching Problem) which does not require learning. In general, keyword spotting methodology contains image pre-processing, word segmentation and spotting similar words. The mutual matching of representations of unique features between query word image and candidate word images returns ranked list of output word images which are most related to the query. Hence, the selection of appropriate features and using robust word matching techniques are the most key phases of a keyword spotting system.

The last years have perceived increased research work of document image analysis for word spotting of historical information in English, Latin, Arabic and Chinese. Fewer research works have been seen for Indic scripts such as Devanagari, Kannada, Tamil and Malayalam and so on. To the best of our knowledge,

this approach is the first to propose a keyword spotting system based on agglomerative hierarchical clustering for historicalDevanagari documents. The structure of this paper is outlined as follows: Section 2 discusses the overview of the keyword spotting system, experimentation and discussion of results are presented in section 4 and finally, the paper conclusion is presented in section 5.

II. RELATED WORKS

Manmatha et al., [5] presented first application of keyword spotting where different algorithms are proposed for matching word images. Gatos et al., [6] presents a segmentation free approach for word spotting where salient regions (particularly a line) are identified, document image descriptors are allocated to block region and matching is achieved only on the regions of interest. Safwan et al., [7] presents line-based keyword spotting which is based on Hidden Markov Model (HMM). It simulates the keywords in model space as sequence of character models and it uses filler models for background or non-keyword text. The use of filler models improves the retrieval result as non-relevant words are handled appropriately by reducing their cost from the overall cost. Volkmaret al.,[8] presents a keyword spotting method based on BLSTM neural network and then applying CTC token passing algorithm. Serrano et al.,[9] presents a model-based approach where each sequence is first mapped to a semi-continuous HMM (SCHMM) and an improved version of DTW is used for matching. Andreas et al., [10] presents an extension to the existing HMM based spotting method by including character n-gram language models. Frinken et al., [11] proposed to use a bidirectional long short–term memory neural network handwritten recognition system for keyword spotting in order to select new elements. Werner Pantke et al., [12] proposed a segmentation-free word spotting approaches. Canet al., [13] proposed a line-based representation for matching words in historical manuscripts. They represented word images with a set of line segments. Then provided a criterion for word matching based on matching the lines. Wshahet al., [14] proposed a statistical script independent line based word spotting framework for offline handwritten documents based on Hidden Markov Models. Zhanget al., [15] proposed Keyword Spotting in Online Chinese Handwritten Documents with Candidate Scoring Based on Semi-CRF Model. Zhanget al., [16] proposed a Character confidence based method on N-best list for keyword spotting in online Chinese handwritten documents. Rothackeret al., [17] proposed a method of Bag-of-Features HMMs for segmentation-free word spotting in handwritten documents. Wang et al., [18] proposed graph based technique for handwritten word representation. Srihari et al., [19] proposed a method for Spotting Words in Handwritten Arabic Documents where the ranking criterion is a similarity score between each prototype word and the candidate words based on global word shape features. Gaurav et al.,[20] proposed a method for Keyword Spotting Framework using Dynamic Background Model . Laszlo et al., [21] proposed a Pivot-based Search for Word Spotting in Archive Documents. Sargur et al., [2] proposed a method of Spotting Words in Latin, Devanagari and Arabic Scripts where the candidate words that are searched in the documents are retrieved and ranked, where the ranking criterion is a similarity score between the query and the candidate words based on global word shape features.

III. SYSTEM OVERVIEW

In proposed method there are three important stages: Image pre-processing and segmentation, feature extraction and word spotting through agglomerative hierarchical clustering. A given query word image is compared with all the available handwritten word images in the source documents and related words are spotted in the documents. Firstly the historical handwritten documents are scanned and are stored as word image database. A set of operations are achieved like image preprocessing, segmentation of handwritten document images into set of candidate word images, combination of unique features are used for representation of segmented word, use clustering approach to analyze all segmented word images features and to group related word image occurrences into clusters, the clustering algorithm spots the related word images which match the query word image. In the proposed approach, keyword spotting is performed on historical handwritten Devanagari documents. The overview of the proposed system is shown in Figure 1.

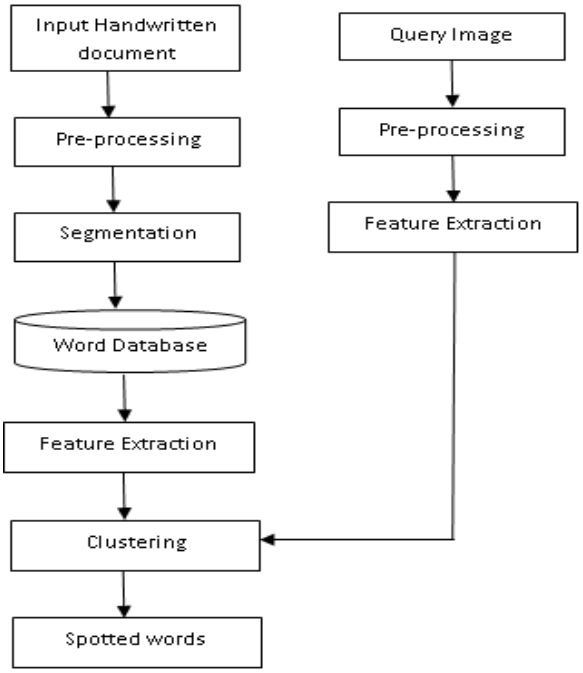


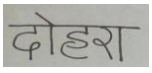
Figure. 1: An overview of Keyword spotting system

3.1 Preprocessing and Segmentation

Historical handwritten documents generally require preprocessing in order to overcome from difficulties such as overlapping of handwriting, slant and touching letters, low quality and low resolution of document images because of occurrence of noise and illumination. Hence earlier to the actual process, the quality of the document image is improved by suitable preprocessing method. Elimination the noise existing in the document image is the primary criteria which is performed using gaussian filter approach. The document image is binarized by threshold value obtained through Otsu's method [22]. Skew angle is identified and Skew angle is corrected by horizontal projection profile approach. The handwritten document image is undergone with thinning process reducing width of line like entity to just single pixel entity and to exterminate all kind of variances arose while writing in the document normalization process is accomplished.For segmentation of document word image, the binary image of document is divided into lines and words. Segmentation is executed using projection profile approach [23] [24]. In line segmentation, the text lines are identified through the valleys of the projection profile which is calculated by a row-wise sum of black pixels. In word segmentation, the line images are segmented into words images using vertical projection profiles approach where the zero valley peaks signifies the spaces between the words.

3.2 Feature Extraction

Feature extraction is the method of dimensionality reduction where the given input data is characterized through group of unique features [25]. This representation is used in comparing the images and in retrieval processes. We considered significant statistical features for the representation of the handwritten word image in the proposed approach. End points, branch points, orientation points, associated components in word images from the thinned segmented word images, Background-to-ink transitions, projection profile, upper projection profile, lower projection profile, distance between upper & lower profile, number of foreground pixels, centre of gravity (C.G.) of the column obtained from the location of fore ground pixels and transition at centre of gravity were grouped as feature set. The starting and termination of a word are represented as end points, connection points of characters are represented as branch points, the loops of word where the point path of orientation deviates are the orientation points and associated components represents number of components present in the word image. Selected examples are presented in Fig. 2.



(a) the original word image



(b) binarized image with interest points

Figure. 2: Illustration of feature extraction with an example.

3.3 Agglomerative hierarchical clustering (AHC)

Clustering helps in designating a set of objects into groups known as clusters. The objective of clustering is to separate a set of 'N' object into 'C' number of clusters where the objects within the cluster should be similar to each other and the objects in various clusters should be dissimilar to each other [26] . Clustering can be used for approximation of the obtainable data, for extracting set of cluster prototypes for the consolidated representation of the data set into homogeneous subsets. There are two types of hierarchical clustering strategies: Agglomerative clustering which is a "bottom up" approach in which each observation starts in its own cluster clusters are unified based on the each observation starts in its own cluster and pairs of clusters are unified as one move up the hierarchy. In Agglomerative clustering based on the distance measurement between clusters, there are three approaches namely single-linkage, complete-linkage and average-linkage. Whereas Divisive clustering is a "top down" approach in which all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In the proposed approach, Agglomerative hierarchical clustering is used for an efficient keyword spotting approach for handwritten Devanagari documents. In a keyword spotting, the major features of segmented word images are characterized in the form of vectors of equal length. The word image database is represented in the form of matrix with 'N' word images with 'M' significant features.

$$F=\begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1m} \\ f_{21} & f_{22} & \cdots & f_{2m} \\ f_{n1} & f_{n2} & \cdots & f_{nm} \end{bmatrix} \quad (1)$$

Where F denotes the feature matrix and the distance matrix is represented using D.

$$D=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix} \quad (2)$$

Where distance between the sample i and sample j are represented by dij. Performance analysis is conducted using single linkage and complete linkage method.

Single linkage: $d(u, v) = \min(dist(u[i], v[j]))$ (3)

Nearest Points are calculated for all points i in cluster u and j in cluster v.

Complete linkage: $d(u, v) = \max(dist(u[i], v[j]))$ (4)

Farthest Points are calculated for all points i in cluster u and j in cluster v. And for calculating the distances, Euclidean distance measure was used.

$$\text{Euclidean distance} = \left|\left|u - v\right|\right|2 = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \quad (5)$$

Dendrogram is used for the representation of result. The distances between the samples are represented

using tree representation called dendrogram. Cluster fragmented representation is acquired by breaking of dendrogram by a definite point. Using Single linkage method and complete linkage method and average linkage method relation between the clusters are specified and appropriate metrics for distances are chosen for best keyword matching results. In general Keyword spotting system, the given query word image is compared with set of all candidate word images stored in the database. In the proposed keyword spotting system using clustering methodology, the segmented word images are grouped into word images clusters. A representative word image is selected from all clusters. A given query word image is then compared with all representative images. The similarity score for which the threshold of similarity lies between 0 to 1 is calculated between the pair of words. Based on the similarity score, particular word cluster is identified and similar word images are spotted in documents.

IV. EXPERIMENTS AND RESULTS

As per our awareness, this is the first work of using Agglomerative hierarchical clustering for keyword spotting of historical handwritten Devanagari documents. Currently as there are no benchmarking dataset available for handwritten Devanagari document images, the proposed method is experimented with three datasets of Devanagari documents. Extensive experimentation was performed using the proposed approach on all three datasets. Scanned handwritten Devanagari documents of the Indian philosopher, Shankaracharya from the Oriental Research Institute @ Mysuru was used as datasets for experimentation of the proposed keyword spotting system. For experimentation from all the handwritten word classes in the segmented word database, we select a subset of subset of 10 classes from UOM Dataset 1, 10 classes from UOM Dataset 2 and 10 classes from UOM Dataset 3. The number of instances per class differs from 3 to 20 and each sample from all classes was used as class representative which is compared with the given query word. For evaluating the performance of different keyword spotting methods, Precision and Recall are computed. The ratio of precision and recall are defined as:

$$\text{Precision} = \frac{|\text{ret} \cap \text{rel}|}{|\text{ret}|} \tag{6}$$

$$\text{Recall} = \frac{|\text{ret} \cap \text{rel}|}{|\text{rel}|} \tag{7}$$

Where ‘rel’ indicates group of relevant words with regard to the query word and ‘ret’ indicates group of spotted words from the word database. Mean average precision (mAP) is measured based on accuracy of word extraction. For a given query word image, score of average precision is calculated using:

$$\text{Average precision} = \frac{\sum_{k=1}^{\text{ret}} (P(k) \times r(k))}{\text{rel}} \tag{8}$$

Where P(k) denotes precision as ‘k’ in the word list and rel(q) denotes the indicator function equaling 1 if the word rank k is a similar word, zero otherwise. The results of proposed method is tabulated in Table 1 and graphically shown in Figure 4.

Class No.	Dataset 1		Dataset 2		Dataset 3	
	SL	CL	SL	CL	SL	CL
1	83.74	73.25	83.95	74.12	84.29	75.43
2	83.12	74.56	84.00	74.95	85.91	75.12
3	84.84	74.83	83.85	74.01	84.34	74.83
4	83.50	73.10	82.89	74.85	84.63	75.82
5	84.39	74.77	85.00	74.88	85.92	74.08
6	83.99	73.00	84.11	75.92	84.71	75.73
7	84.28	74.89	84.94	74.21	84.80	74.60
8	84.84	74.71	84.75	74.93	85.39	74.81
9	83.17	74.62	83.80	73.42	84.20	74.29
10	84.08	74.53	85.17	73.11	83.84	75.98

Table 1: Mean average precision of the proposed clustering based keyword spotting approach.
Where SL denotes Single Linkage, CL denotes Complete Linkage.

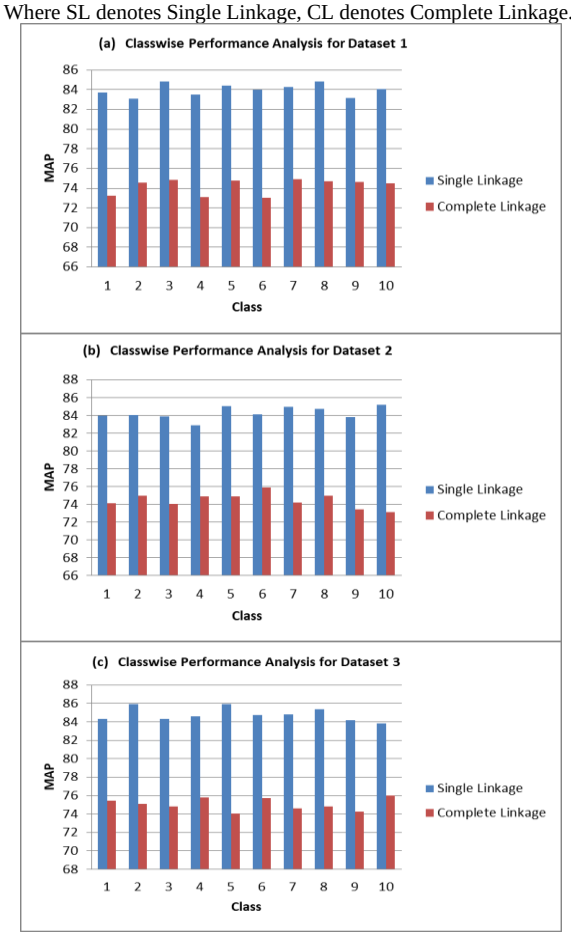


Figure 3: Class wise labeling results of proposed clustering based keyword spotting approach in terms of Mean average precision for (a) Dataset 1 (b) Dataset 2 (c) Dataset 3.

V. CONCLUSION AND FUTURE WORK

In this paper, a keyword spotting of historical handwritten Devanagari documents using agglomerative hierarchical clustering techniques is illustrated. There are only limited works reported using cluster based keywords spotting for handwritten documents. However, the proposed work is the first effort for spotting keywords from handwritten Devanagari document images using agglomerative hierarchical clustering approach. In this paper, the authors have also presented the comparison of performance of a keyword spotting using both single linkage and complete linkage method. Wide experimentation is conducted using the handwritten Devanagari documents which shows that keyword spotting by agglomerative hierarchical clustering using single linkage techniques significantly outperforms than keyword spotting by agglomerative hierarchical clustering using complete linkage techniques. In conclusion, considering the huge complexity in processing handwritten Devanagari script, the impact of the proposed approach may be considered momentous with acceptable keyword spotting performances.

In the future research, the proposed methodology can be also used for the keyword spotting in numerous Indic languages as well as Arabic language. Most of the keyword spotting systems uses segmentation approach. Hence, segmentation free keyword spotting approach for handwritten documents can be focused in future.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We acknowledge Oriental Research Institute, Mysore, Karnataka, India for providing scanned handwritten historical Devanagari manuscripts.

REFERENCES

[1] Rath, T.M., Manmatha, R., Word spotting for historical documents. International Journal on Document Analysis and Recognition 9, 139–152, 2007.. https://doi.org/10.1007/s10032-006-0027-8.

[2] Srihari, S., Srinivasan, H., Babu, P., Bhole, C., Spotting words in handwritten Arabic documents. Document Recognition and Retrieval XIII 6067, 2006. 606702-606702–12. https://doi.org/10.1117/12.643107.

[3] Almazan, J., Gordo, A., Fornes, A., Valveny, E., Word spotting and recognition with embedded attributes, in: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp.2552–2566,2014.https://doi.org/10.1109/TPAMI.2014.233981

[4] Rothacker, L., Rusinol, M., Fink, G.A., 2013. Bag-of-features HMMs for segmentation-free word spotting in handwritten documents, in: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR.pp.1305–1309. https://doi.org/10.1109/ICDAR.2013.264

[5] Manmatha, R. and W. B. Croft. "Word spotting: Indexing handwritten archives." Intelligent Multimedia Information Retrieval Collection pp.43-64, 1997.

[6] Gatos, Basilios and Ioannis Pratikakis, Segmentation-free word spotting in historical printed documents, Document Analysis and Recognition, 10th International Conference on. IEEE, 2009.

[7] Wshah Safwan, Gaurav Kumar, and Venu Govindaraju, Script independent word spotting in offline handwritten documents based on hidden markov models, Frontiers in Handwriting Recognition (ICFHR), International Conference on. IEEE, 2012.

[8] Frinken Volkmar, ,A novel word spotting method based on recurrent neural networks, Pattern Analysis and Machine Intelligence, IEEE Transactions on Vol.34(2), pp.211-224, 2012.

[9] Rodríguez-Serrano, José A., and Florent Perronnin, A model-based sequence similarity with application to handwritten word spotting, Pattern Analysis and Machine Intelligence, IEEE Transactions on Vol. 34(11), pp.2108-2120, 2012.

[10] Fischer Andreas, Improving hmm-based keyword spotting with character language models, Document Analysis and Recognition(ICDAR),International Conference on. IEEE, 2013.

[11] Volkmar Frinken , Andreas Fischer, Markus Baumgartner , Horst Bunke, Keyword spotting for self-training of BLSTM NN based handwriting recognition systems, Pattern Recognition, Vol.47, pp.1073–1082, 2014.

[12] Werner Pantke, Martin Dennhardt, Daniel Fecker, Volker M¨argner, Tim Fingscheidt, An Historical Handwritten Arabic Dataset for Segmentation-Free Word, International Conference on Frontiers in Handwriting Recognition, 2014.

[13] Ethem F. Can , Pınar Duygulu, A line-based representation for matching words in historical manuscripts, Pattern Recognition Letters Vol.32, pp. 1126–1138, 2011.

[14] Safwan Wshah n, Gaurav Kumar ,Venu Govindaraju ,Statistical script independent word spotting in offline handwritten documents, Pattern Recognition Vol.47, pp.1039–1050, 2014.

[15] Heng Zhang, Xiang-Dong Zhou, Cheng-Lin Liu, Keyword Spotting in Online Chinese Handwritten Documents with Candidate Scoring Based on Semi-CRF Model, International Conference on Document Analysis and Recognition, 2012.

[16] Heng Zhang, Da-Han Wang, Cheng-Lin Liu, Character confidence based on N-best list for keyword spotting in online Chinese handwritten documents, Pattern Recognition Vol.47, pp.1880–1890, 2014.

[17] Rothacker, L., Rusinol, M., Fink, G.A, Bag-of-features HMMs for segmentation-free word spotting in handwritten documents, in: Proceedings of the International Conference on Document Analysis and Recognition, ICDAR. pp. 1305–1309, 2013.

[18] Peng Wang, Veronique Eglin, Christine Largeron, Josep Llados, Alicia Fornes, A Novel Learning-free Word Spotting Approach Based On Graph representation, IAPR International Workshop on Document Analysis Systems, 2014.

[19] Sargur Srihari, Harish Srinivasan ,Pavithra Babu and Chetan Bhole, Spotting Words in Handwritten Arabic Documents, Center of Excellence for Document Analysis and Recognition (CEDAR).

[20] Gaurav Kumar, Zhixin Shi, Srirangaraj Setlur, Venu Govindaraju, Sitaram Ramachandrula, Keyword Spotting Framework using Dynamic Background Model, International Conference on Frontiers in Handwriting Recognition, 2012.

[21] Laszlo Czuni, Peter Jozsef Kiss, Agnes Lipovits, Monika Gal, Pivot-based Search for Word Spotting in Archive Documents, Proceedings of the International Conference of Machine Vision and Machine Learning, pp. 133, 2014.

[22] Papandreou, B. Gatos, G. Louloudis, and N. Stamatopoulos, “Document image skew estimation contest,” in ICDAR, pp. 1444–1448, 2013.

[23] V. Lavrenko, T. Rath, R. Manmatha, Holistic word Recognition for Handwritten Historical Documents,” in Proc. Doc. Image Anal. Libr. DIAL’04, pp.278-287, 2004.

[24] S. Abirami and D. Manjula, "Profile Based Information Retrieval from Printed Document Images," in International Conference Computer Graphics, Imaging and Visualization. Vol.3, pp.268-272, 2013.
[25] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” in International journal of computer vision, vol. 60, pp. 91–110, 2004.
[26] F. Murtagh, A survey of recent advances in hierarchical clustering algorithms which use cluster centers, Computer Journal, Vol. 26, no. 4, pp. 354-359, 1984.

★★★