Olex-Genetic Algorithm Based Information Retrieval Model from Historical Document Images

N. Vanjulavalli

***Abstract:** Presently, the process of retrieving the historical documents is treated as an important challenge due to the fact that the document possessesindividual structure and level of deprivation. The textual characters present in the printouts come together with the typographical objects. The retrieval and perusal of the visual typographical objects indicates that the content of heritage documents helps to effectively interpret the documents. The extraction of the visual objects finds useful in the interpretation and conveyance of more details regarding diverse practices of demonstration in past documents and the impact in the present status of publication. A pair of essential typographical objects linked to the history of knowledge and information is footnotes and tables. In this paper, the main intention is the detection of the existence of the visual elements from historical printable documents. A new Olex-GA based footnote recognition model (OFR) is developed. The footnote detection model make use of a collection of layout-based models for the extraction of few features concerning to the font and appearance. In addition, Olex-GA algorithm is for the classification purposes. This model is validated using a massive set of 18thcentury printed documents with higher than 32 million images, and the outcome showed their effective outcome of the presented model.*

***Keywords:** Information retrieval; Historical documents; OlexGA; Classification*

I. INTRODUCTION

Historical reports pass on important data about social legacy. Be that as it may, recovering the ideal information from them is a difficult assignment since they are not joined by adequate lists and metadata. Our point in this exploration is to encourage the authentic archive data recovery process by distinguishing typographical items in reports throughout the 18thcentury. The two objects focused in the study are footnote and table.Document image retrieval (DIR) offers the required way of retrieving, indexing and annotating the visual details from the documents [1]. DIR models are mainly based on two dimension of document images namely text by the use of optical character recognition (OCR) tool, or image [2]. By the application of OCR models on the history documents with irregular and poor structure is a difficult and error-prone task. So, the recognition free models hold the priority to deal with the historical documents. An important and differentiable characteristic of the European Enlightenment period is the extension of printed document [3]. The extension brought additional feature of this period to concentration: the recurrent usage of typography objects for conveying necessary data. These objects are always employed for representing the data to minimize the complexity as well as imprecision. The range of typographical practice is taking into accountas the growth of the advanced printing technologies helps the research people from diverse domain of history. The human sciences and the investigation of document images helps to attain an effective grab of “the page image” and its influence on the interpretation of the past scientific knowledge.

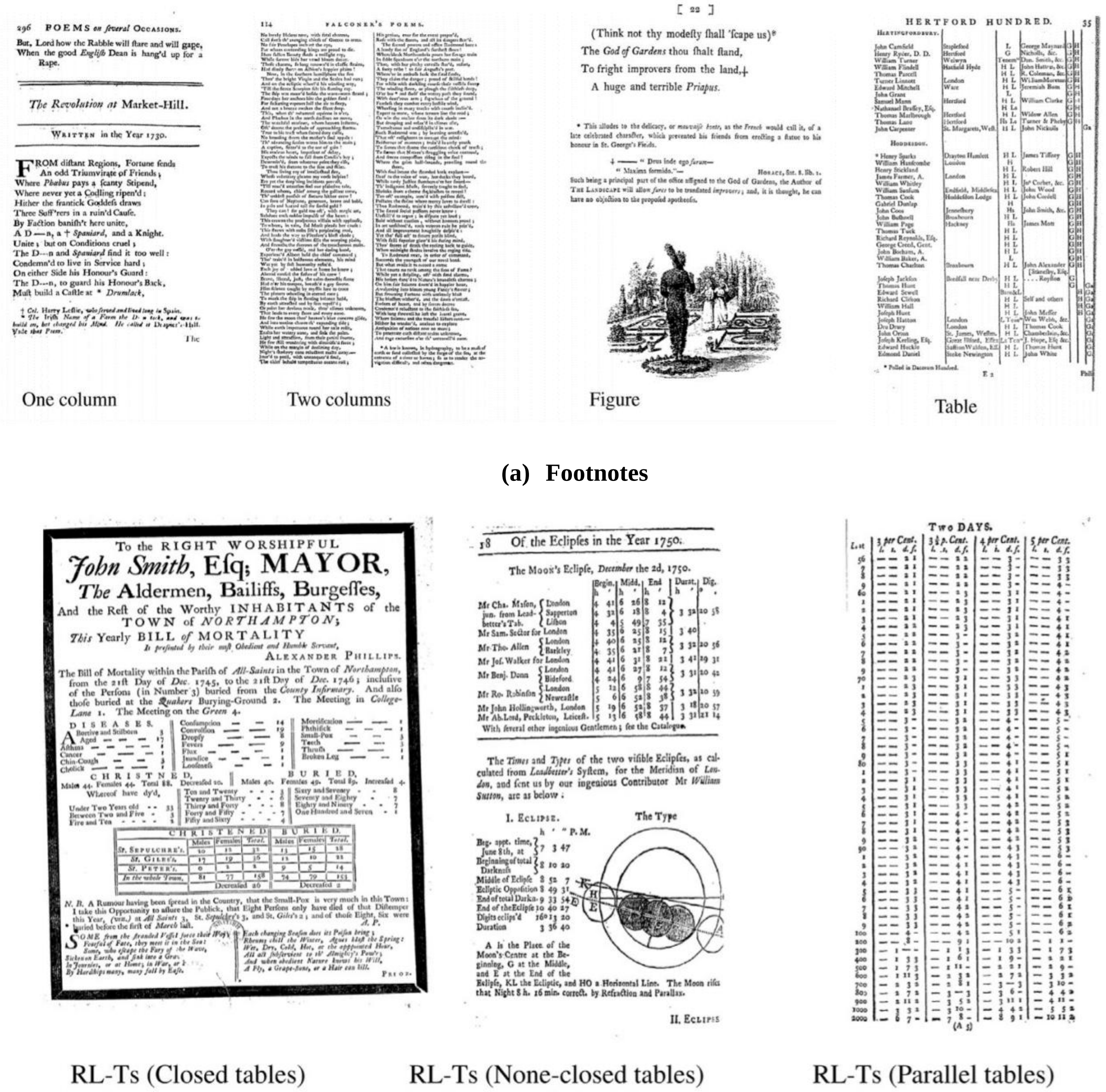
A set of 4 essential visual practices which has been employed in the advanced printed documents are footnote, table, diagram and illustration. Here, the main concentration has been given only to the footnotes. It is a collection of indexical print forms comprising in-text citation, marginal note, appendix and bibliography [4]. Since the name denotes, it exists at the end of the document and offer details related to the author denotations [5]. The existence of footnote as cross-reference is an important practice in the 18th century and it is continuously enhanced with the printable document count. So, the identification of footnote is an important complement in the connection of various scholars and printable documents [4]. Table is an important typographical object among diverse document classes. They have gathered the separate and unorganized details and build a compact geometrical demonstration of identical data [6].

Several works has been developed to analyze the structure of the documents. This work aimed to find various components of the document layout by the use of one or both of the textual as well as pixel-wise details of the image. In addition, the influence of degradation affect the retrieval process of identical shapes of footnote markers (e.g., ‘*’, ‘*)’, ‘1)’). On the other hand, [7] presented a model to detect the footnotes through the investigation of text line. But, it is restricted to easy layouts and assumes that footnotes are present in an individual line. [8] devised a model using deep belief network (DBN) comprising two stages namely unsupervised pre-training and supervised fine-tuning. It defines every image by the combination of two text lines from the top of the image and two from the bottom. A major advantage of this method is the management of sparse label data particularly for heritage documents were inadequate labeled data only exists. [9]

Manuscript published on November 30, 2019.
* Correspondence Author

Dr. N. Vanjulavalli, *Asst. Professor, Dept of Computer Science Annai College of Arts and Science affiliated to Bharathidasan University, India. vanjulavallisn@gmail.com

employs a 1-dimensional convolutional neural network (CNN) for tackling the issue. Here, the representation of document images undergo representation by the combination of the horizontal intensity histogram of the top 2 and bottom 3 text lines that reduces the complexity of the application of CNN. In addition, [10] utilizes a collection of rule-based characteristics attained from the visual observation that the footnote has a small font size. Then, the classification takes place by the use of support vector machine (SVM).



(a) Footnotes

(b) Tables

Fig. 1. Heritable document samples

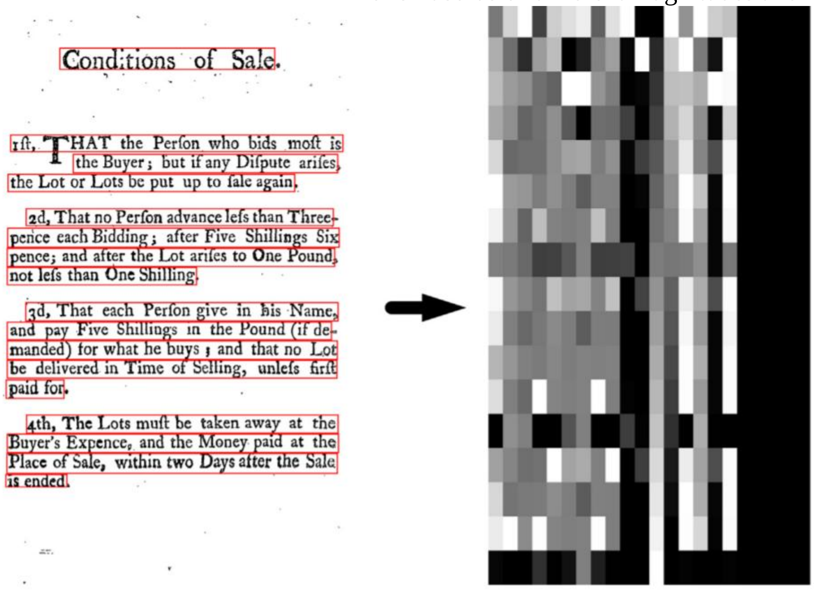
Fig. 1 illustrates few instances of the footnote and table present in the document from the 18th-century. The typographical objects in the Enlightenment period does not involve a direct format, the heterogeneous and degrading nature of digitalization of past document makes the task of DIR very difficult over the advanced versions. The major intention of this study is the design of a new model to retrieve the typographical objects from the document images and categorizes it. This model seems to be reliable to the dynamic kinds of layout type's artifacts of heritage documents. To detect the footnote, a 2 layout-based model is present which extract the collection of rule based characteristic from the images. The filtered features are then undergoing classification based on the existence of the footnotes or tables present in the page. A pair of essential typographical objects linked to the history of knowledge and information is footnotes and tables. In this paper, the main intention is the detection of the existence of the visual elements from historical printable documents. The footnote detection model make use of a collection of layout-based models for the extraction of few features concerning to the font and appearance. In addition, Olex-GA algorithm is for the classification purposes. This model is validated using a massive set of 18th century printed documents with higher than 32 million images, and the outcome showed their effective outcome of the presented model.

II. PROPOSED MODEL

A two layout-based framework to classify the footnote-based document images are presented here.

2.1. The first layout-based model

The organization and characteristics of the text lines on a page holds important details concerning the existence of a footnote on the page. This concept is depends on the horizontal projection of the document images. Next, the identification of the vertical position of every text line takes place through the identification of void space among the histogram peaks utilizing an adaptive determined threshold value. The horizontal position of the text line is determined by the application of vertical projection in every line. The outcome of this procedure composes of a set of 4 values distance of the text-line box from left (x), top (y), width (w) and height (h) of the text-line box. Once the text line is extracted, the subsequent process is the calculation of the collection of measurement. These values represent the features of text line of every line. Fig. 2 provides the sample conversion of the image to a novel type of representation. It is shown that the modifications in the characteristics of the text line is reflection in the table value, for instance, the end of the paragraph text lines appear with small amounts (black pixels). The normalization of the values among the minimal and average values takes place. It is due to the fact that the average over the common maximum functions for avoiding cases such as catchwords, signature marks or volume figures (see Fig. 3) which adequately comes at the bottom of the page. Because of the different number of text lines in every document, the normalization of the feature vectors in every image takes place by the use of a discrete cosine transform (DCT) of the feature matrix. The DCT assists in the representation of the feature matrix as a real value summary of sinusoids of different magnitudes and frequencies.



Conditions of Sale.

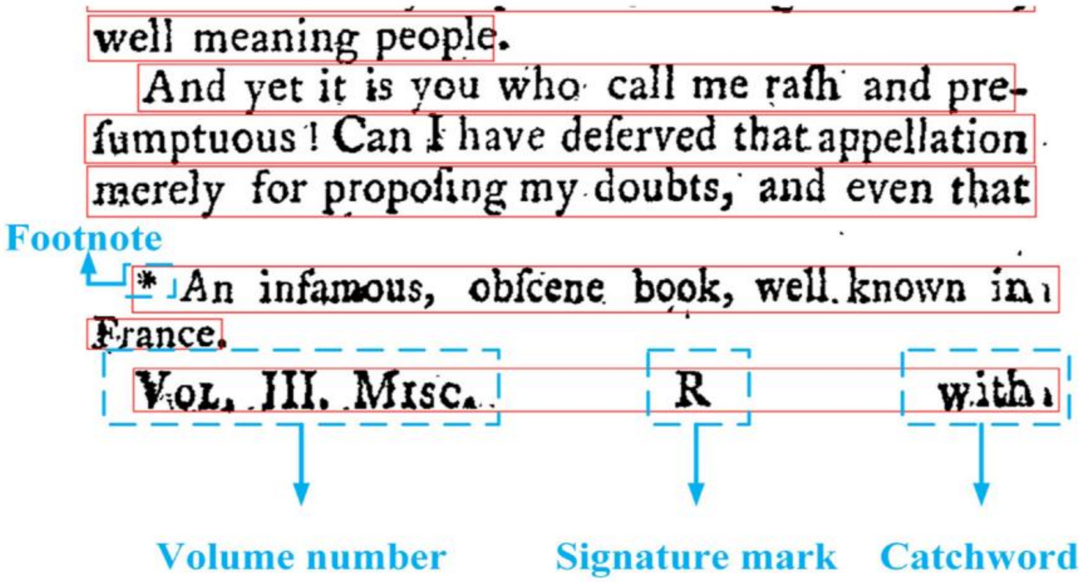
1ſt, THAT the Perſon who bids moſt is the Buyer; but if any Diſpute ariſes, the Lot or Lots be put up to ſale again.

2d, That no Perſon advance leſs than Three-pence each Bidding; after Five Shillings Six pence; and after the Lot ariſes to One Pound, not leſs than One Shilling.

3d, That each Perſon give in his Name, and pay Five Shillings in the Pound (if demanded) for what he buys; and that no Lot be delivered in Time of Selling, unleſs firſt paid for.

4th, The Lots muſt be taken away at the Buyer's Expence, and the Money paid at the Place of Sale, within two Days after the Sale is ended.

Fig. 2. Sample conversion of an image to text



well meaning people.
And yet it is you who call me raſh and pre-
ſumptuous! Can I have deſerved that appellation
merely for propoſing my doubts, and even that

* An infamous, obſcene book, well known in
France.

VOL. III. MISC. R with

Footnote
Volume number
Signature mark
Catchword

Fig. 3. Sample bottom portion of the document image

2.2. The second layout-based approach

It is an improved version of the earlier one. The small font size of the footnote undergoes a comparison with the body text. Here, extra two hypotheses based on the position of the footnotes and the void space among the footnote line and main text are provided. By the use of hypothesis and observance of atleast 100 images from every individual class, the feature vector gets modified and improved.

A rule based model is used in this model. The characteristics are altered by the inclusion of a preprocessing level in prior to the investigation of the layout. A part of the page is cropped in a vertical way from the right for avoiding upcoming interference in any line comprising catchwords in the segmentation process and in the footnote identification process. It is assumed that the usual word length and the right boarder's width once diverse values are tried, the terminal cropping value is assumed to 1/4. In addition, once the text line is segmented, the final text lines are elimination beginning from the vertical middle line of the page. It assists in the elimination of the identification of signatures as footnote. On the integration of every feature, a feature vector of size 72 is developed. And then, OlexGA classifier is presented for the classification of the pages based on the occurrence of the footnote.

2.3. Olex-GA classifier

GA is generally contains a set of 3 fundamental components namely (1) population, i.e., collection of candidate solutions known as individual or chromosome, which gets evolved in the iteration count, (2) fitness function utilized for assigning a value for every individuals of the population; (3) evolution strategy depending upon the operators like crossover, mutation and selection.

Population Encoding

In GA based model, every individual present in the population can be represented as an individual rule or a collection of rules. The first model arranges the individual encoding simpler, however, the fitness of an individual could not be a meaningful indicator of rule quality. At the same time, diverse rules per individual model where an individual indicates the whole classification, and it needs a proper individual coding, however the fitness does not offers a reliable sign. Hence, there is a balance among the simplicity of encoding and efficiency of the fitness function.

Fitness Function

The fitness of a chromosome K, indicating $H_c(Pos, Neg)$, is the value of the F-score comes from employing $H_c(Pos, Neg)$ to the training set TS. This option obviously comes from the formulation of problem $MAX - F$. At present, when $D(K) \subseteq TS$represents the set of every document comprising the positive term in Pos and none of the negative term in Neg, i.e.,

$$D(K) = \cup_{t \in_{Pos}} \Delta(t) \setminus \cup_{t \in_{Neg}} \Delta(t) \qquad (1)$$

opening from the description of $F_{c,\alpha}$, following some algebra, the representation of $F_{c,\alpha}$ is attained:

$$F_{c,\alpha}(K) = \frac{|D(K) \cap TS_c|}{(1-\alpha)|TS_c| + \alpha|D(K)|} \qquad (2)$$

Evolutionary Operators

The selection process is carried out using the roulette-wheel technique and crossover takes place by the uniform crossover scheme. Mutation holds the process of flipping every individual bit offered with the provided possibility. For losing efficient chromosomes, elitism is applied verifying that the optimal individuals comes from the present generation is given to the subsequent one with no alternation by a genetic operator.

GA

Initially, the collection $Pos *$ and $Neg *$of entrant positive and negative terms, will be determined0 from the input vocabulary $V\ (k, f)$. Once population initialization has been done, it starts to evolve by the process of iterating elitism, selection, crossover and mutation, till a specific number n of generations is generated. In every step, a repair operator ρ, intends at the correction of the probable illegal individuals generation through crossover or mutation, is employed.

III. PERFORMANCE VALIDATION

3.1. Dataset used

The presented OFR model is validated utilizing the relative dataset and diverse performance validation setup. The presented model is tested over a total of 32 million historical images. Here, the dataset and the results analysis of presented model allocated to every footnote and table classification process. A major challenge in the process of classifying footnotes in the document is the absence of extensive studies and proper dataset. Thus, a dataset "The Visibility of Knowledge" 1 project is created and is utilized in this study. In addition, a set of digital images from Eighteenth Century Collections Online (ECCO) is applied. It holds a maximum of 32 million document pages from over 155,000 volumes and 8 subjects which undergone distribution over two subsets namely ECCOI and ECCOII. The images undergo arbitrary selection from the ECCOI dataset for keeping the uniform distribution of the ECCO and are partitioned into a set of 5 subsets. Fig. 4 illustrates further details regarding every subset.

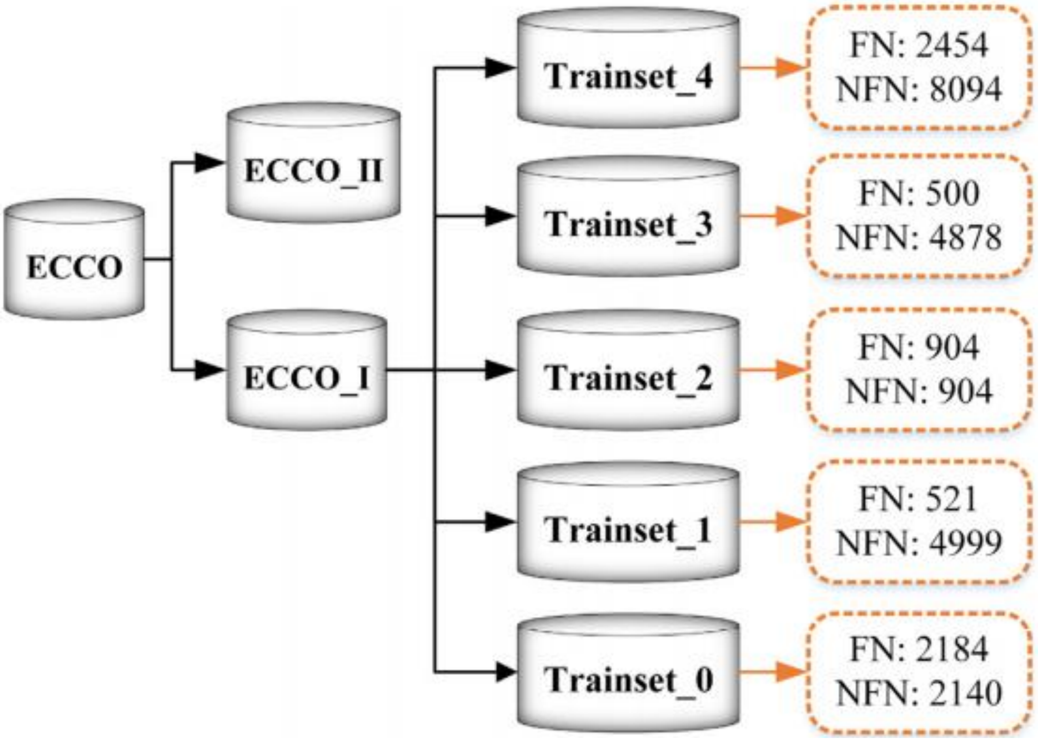


Fig. 4. Details about the labeled footnote dataset

3.2. Results analysis

A set of evaluation metrics used for the evaluation of the presented model is given in Table 1 under the use of 10 fold cross-validation. The existing models are used for making a comparison with the proposed method. Fig. 5 demonstrates the results attained by diverse methods with respect to precision. Looking into the values present in the table, it is shown that the presented OFR model shows better results which can be clarified by the maximum precision value of 88.58. In addition, the SVM model shows manageable results over the presented OFR model by achieving a lower precision value of 85.21. At the same time, the Bbox with Proj based model and space-location based model showed identical and competitive performance of the applied images by attaining the maximum precision values of 87.80 and 87.54 respectively. However, the presented OFR model is superior to other methods with respect to the precision value of 88.58.

Table 1 Classifier results analysis of OFR model

Methods	**Precision**	**Recall**	**F-measure**
Proposed	88.58	85.21	85.87
SVM Classifier	85.21	84.32	84.76
Bbox + Proj based approaches	87.80	76.02	81.45
Space &location based approach	87.54	74.77	80.66

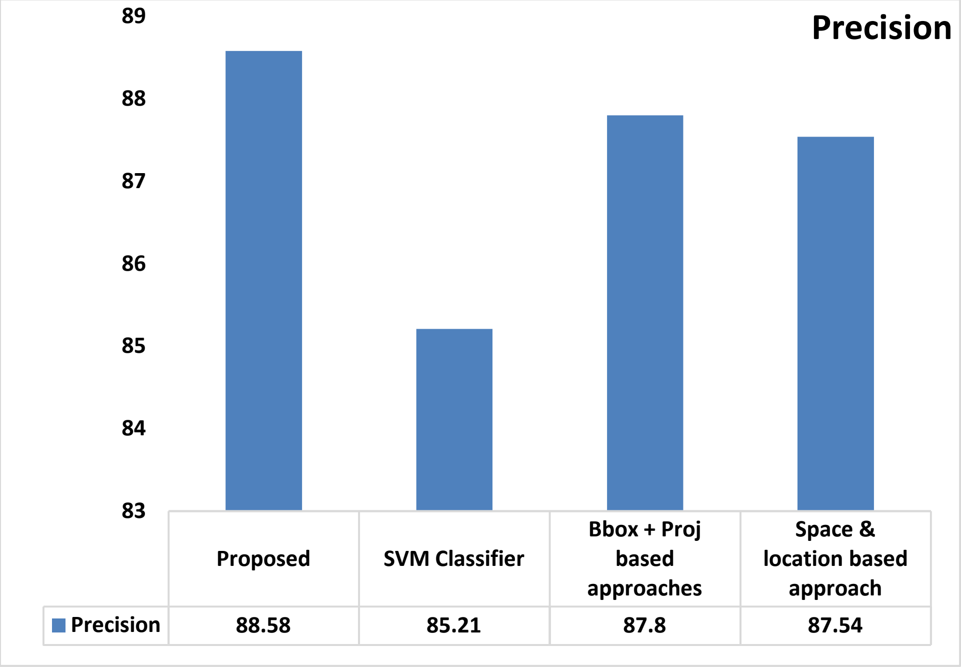


Fig. 5. Classifier results analysis of OFR model based on precision

Fig. 6 demonstrates the results attained by diverse methods with respect to recall. Looking into the values present in the table, it is shown that the presented OFR model shows better results which can be clarified by the maximum recall value of 85.2. In addition, the SVM model shows manageable results over the presented OFR model by achieving a slightly lower recall value of 84.32. At the same time, the Bbox with Proj based model and space-location based model showed ineffective performance of the applied images by attaining the maximum recall values of 76.02 and 74.77 respectively. However, the presented OFR model is superior to other methods with respect to the recall value of 85.21.

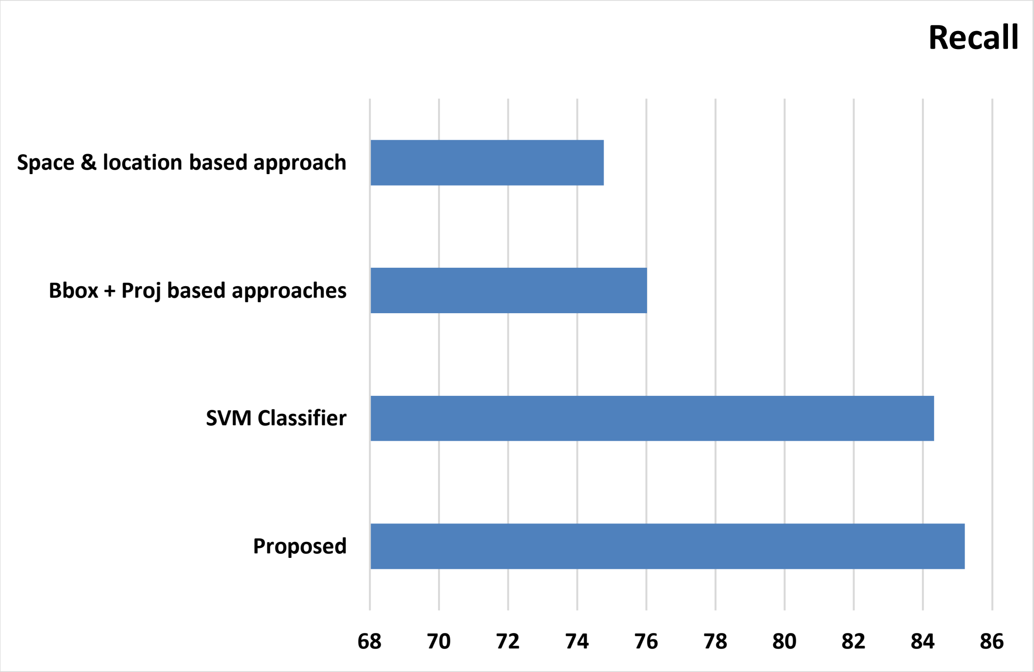


Fig. 6. Classifier results analysis of OFR model based on recall

Fig. 7 demonstrates the results attained by diverse methods with respect to F-measure. Looking into the values present in the table, it is shown that the presented OFR model shows better results which can be clarified by the maximum F-measure value of 85.87. In addition, the SVM model shows manageable results over the presented OFR model by achieving a slightly lower F-measure value of 84.76.

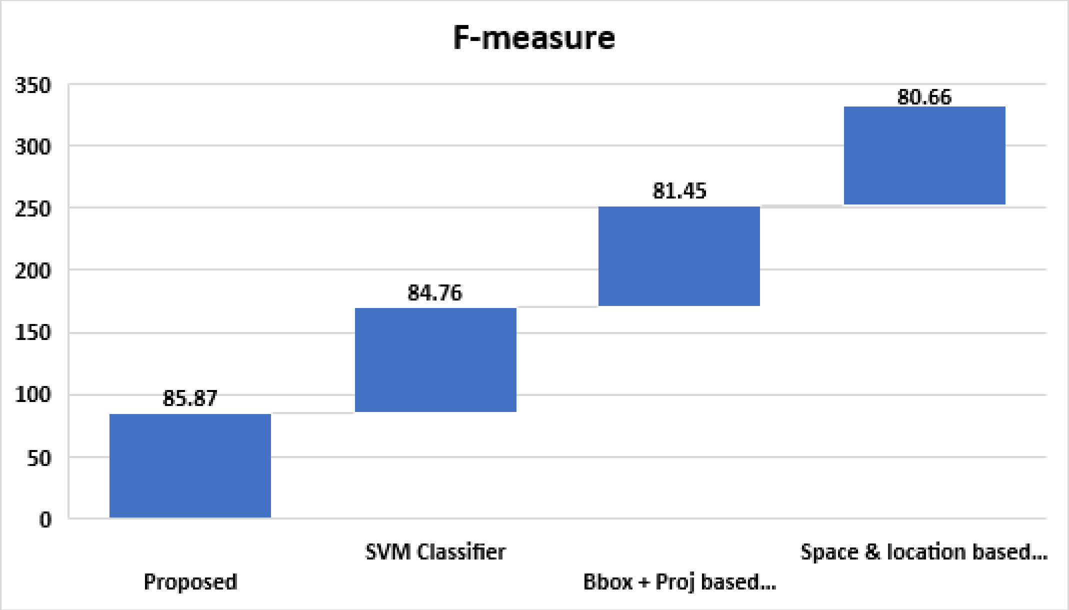


Fig. 7. Classifier results analysis of OFR model based on F-measure

At the same time, the Bbox with Proj based model and space-location based model showed ineffective performance of the applied images by attaining the maximum F-measure values of 81.45 and 80.66 respectively. However, the presented OFR model is superior to other methods with respect to the F-measure value of 85.87.After observing the values present in the table and figures, it is evident that the presented model shows effective results with the maximum precision of 88.58, recall of 85.21and 85.87 of F-measure.

IV. CONCLUSION

Several works has been developed to analyze the structure of the documents. This work aimed to find various components of the document layout. The major intention of this study is the design of a new model to retrieve the typographical objects from the document images and categorizes it. This model seems to be reliable to the dynamic kinds of layout type's artifacts of heritage documents. To detect the footnote, a 2 layout-based model is present which extract the collection of rule based characteristic from the images. The filtered features are then undergoing classification based on the existence of the footnotes or tables present in the page. In addition, Olex-GA algorithm is for the classification purposes. This model is validated using a massive set of 18th century printed documents with higher than 32 million images, and the outcome showed their effective outcome of the presented model. After observing the values present in the table and figures, it is evident that the presented model shows effective results with the maximum precision of 88.58, recall of 85.21and 85.87 of F-measure.

REFERENCES

1. S. Marinai, E. Marino, G. Soda, Exploring digital libraries with document image retrieval, in: International Conference on Theory and Practice of Digital Libraries, Springer, 2007, pp. 368–379.
2. G.G. Chowdhury, Introduction to Modern Information Retrieval, Facet publishing, 2010.
3. C. Wellmon, Organizing Enlightenment: Information Overload and the Invention of the Modern Research University, JHU Press, 2015.
4. B. Pasanek, C. Wellmon, The enlightenment index, Eighteenth Century 56 (2015) 359–382.
5. A. Grafton, The Footnote: A Curious History, Harvard University Press, 1999.
6. T.A. Tran, H.T. Tran, I.S. Na, G.S. Lee, H.J. Yang, S.H. Kim, A mixture model using random rotation bounding box to detect table region in document image, J. Visual Commun. Image Represent. 39 (2016) 196–208.
7. V. Eglin, A. Gagneux, Visual exploration and functional document labeling, in: icdar, IEEE, 2001, p. 0816
8. S. Abuelwafa, M. Mhiri, R. Hedjam, S. Zhalehpour, A. Piper, C. Wellmon, M. Cheriet, Feature learning for footnote-based document image classification, in: International Conference Image Analysis and Recognition, Springer, 2017, pp. 643–650.
9. M. Mhiri, S. Abuelwafa, C. Desrosiers, M. Cheriet, Footnote-based document image classification using 1d convolutional neural networks and histograms., in: Tools and Applications (IPTA), 2017 Seventh International Conference on Image Processing Theory, IEEE, 2017, pp. 1–5.
10. S. Zhalehpour, A. Piper, C. Wellmon, M. Cheriet, Footnote-based document image classification, in: F. Karray, A. Campilho, F. Cheriet (Eds.), Image Analysis and Recognition, Springer International Publishing, Cham, 2017, pp. 634–642.

AUTHORS PROFILE

Dr. N. Vanjulavalli is working as head of the department of Computer Science at Annai College of Arts and Science affiliated to Bharathidasan University, Trichy. She has 16 years of teaching experience and 8 years of research experience. She is specialized in data mining, ontology and e- learning and Neural Networks. She has published5 research papers cited in scopus and science citations. She acted as resource person cum guest speaker for national level seminars at various institutions.