An intelligent and unified text and non-text object extraction from PDF using Support Vector Machine

Dr.V.Anandkumar,
Professor, Department of Information Technology, Sri Krishna College of Engineering and Technology, Coimbatore. anandkumar@skcet.ac.in
Mr.A.Vijay,
Assistant Professor, Department of Business Administration and Information Systems, Arbaminch University, Sawla campus, Ethiopia. vijay.kpu@gmail.com
Ms.G.Divya,
Assistant Professor, Department of Computer Science and Engineering, Saveetha School of Engineering, Chennai. mailtodivya16@gmail.com
Mr.V.Arulkumar,
Assistant Professor, Department of Computer Science, Sri Krishna College of Engineering and Technology, Coimbatore. arulkumarv@skcet.ac.in

Article Info
Volume 83
Page Number: 3538 - 3546
Publication Issue:
July-August 2020

Article History
Article Received: *25 April 2020*
Revised: *29 May 2020*
Accepted: *20 June 2020*
Publication: *10 August 2020*

Abstract:
Today's e-book plays an important role in all fields to learn new things through personal computers, laptops or mobile phones. There are several formats for an eBook. The most used format is PDF because it preserves the original format of the document. Segmentation is used to reuse content, but in the existing system documents are only segmented as textual content. It does not take into account non-textual elements, such as graphics, tables and images. In this survey, the design analysis is performed by extracting text objects and non-text objects from the PDF document and segmenting the objects separately using the Support Vector Machine (SVM) classifiers. Finally, we get the output as text objects and non-text objects separately. This method uses a bottom-up approach to extract lines of text and a top-down approach to split the diagram tree generated by Kruskal's algorithm into sub diagrams that use the Euclidean distance between adjacent vertices. Text and non-text objects are classified using SVM techniques. With each section using the SVM technique for each segmented and non-textual text, different dimensional characteristics are extracted for labeling purposes. Different eBook PDF documents are tested, and some sample input and output PDF documents are shown in the experimental results.

Keywords*: E-book, PDF object, Support vector machine, Graph based Image Segmentation.*

I.INTRODUCTION

An e-book is an electronic model where we get a traditional print book from either a personal computer or by using an e-book reader. The e-book is available in various formats like MOBI, AZW, AZW1, AZW4, EPUB, and PDF. The most widely used e-book format is PDF because while transferring PDF documents it maintains the original formatting and security; no one can change the content of the document. The PDF document may contain text objects and image objects. Text objects contain only the text data. Image objects include graphs, tables, lists, and images. Document segmentation plays an important role in e-book which is used to reuse the content of the document. It is a method of sub dividing the document regions as text regions and image regions and it leads to layout analysis.

The document can be divided into text segmentation and image segmentation. In existing work, you can only segment the text content of a PDF document. However, it is more important to segment the image with text segmentation. Text segmentation is a precursor to text retrieval, auto synthesis, information retrieval, language modeling, and natural language processing. In written texts, text segmentation is the process of identifying boundaries between words, phrases or other important units of language such as sentences and arguments. The term, separated from such processing, is useful for helping people read text and is mainly used to help computers perform certain man-made processes as basic units. Line extraction is a preprocessing step for handwriting recognition and document structure extraction, and image segmentation is a mid-level processing technique. The main reason for the segmentation process is to get more information about the area of interest from an image.

Most PDF documents contain both text objects and images. This examination takes into account both text and image objects for segmentation in PDF documents. The proposed research takes into account the segmentation of text and image components in the PDF e-book format. This overcomes the limitation of segmentation in tables, images, graphics, etc. in the PDF document. In this system both text layer and image layer are taken into consideration for segmentation. Each layer segments its data independently. Finally the results of both text and image layers are merged together for final segmentation. Text segmentation and image segmentation are used for reusable purpose

II. LITERATURE REVIEW

Neha Gupta et al introduced a text extraction concept which is based on Image Segmentation. The text involved in these images includes critical and useful information. Text extraction in images has been used in a large variety of applications such as vehicle license plate detection, document retrieving, mobile robot navigation, and object identification. In this system, we retrieve text information from complex input images by using Discrete Wavelet Transform (DWT). But a preprocessing step is required for color to extract text edges in the color image. The edge map is formed using resultant edges. Morphological operations are applied to improve the performance on the processed edge map and then thresholding is applied in the image.

Chandranath Adak et al introduced a new method for Unsupervised Text Extraction from G-Maps. Text extraction is a method of extracting passage of text from a non-text background. Due to an unsupervised approach, no prior knowledge or training is required on the textual and non-textual parts. The fuzzy C mean clustering technique or the Prewitt method are used for image segmentation and edge detection. The limitation of this system is that it is not fully automatic due to the threshold and the selection of a better result depends on the human eye.
Q. Yuan et al introduced a new text extraction technique which is based on Edge Information. The designed scheme presents a well-designed approach that uses area statistics to take out textual blocks from grey scale record pictures. The main objective of this scheme is to find out textual regions on heavy noise- infected newspaper photos and split them from graphical regions. The algorithm traces the function points in unique entities and then groups the ones with facet points of textual areas. From this method we can obtain accurate web page decomposition with green computation and reduced reminiscence size by copying with line segments.

Thai V. Hoang et al introduced a new text extraction method which is based on Sparse Representation. Input document image includes both text and graphics which is processed to produce two output images, one returns with text and the other returns with graphics. Graphical file pictures containing textual content and graphic additives are taken into consideration as two-dimensional indicators. The proposed set of rules fully depends upon a sparse representation framework with the content as it should be chosen discriminative over complete dictionaries. Every one offers sparse illustration above one type of signal and non-sparse illustration above the other. Separation of text and image additives is obtained via promoting sparse graphic of input images in these dictionaries. The proposed approach overcomes the problem of handling among text and images.

S.Ranjini et al implemented a new method of extracting and recognizing text taken from an English digital cartoon image using the median filter. In this work, blob extraction functions are used and Japanese text is extracted vertically from Manga Comic Image. At the same time, Optical Character Recognition (OCR) removes several text restrictions at the same time and converts Japanese manga language to some additional languages in the traditional way, for the satisfaction of learning manga on the internet.

III. EXISTING METHODOLY

In the existing work the text documentation is to group text into visually homogeneous blocks. From PDF document we separate the text components from the image components such as images, tables and graphs. Here, line segmentation is considered over a horizontal reading order. This method involves three modules which are text information retrieval, the merging of words into text lines and the grouping of text lines into text blocks.

A. Text Information Retrieval

A PDWordFinder extracts words from a PDF file, and enumerates the words on a single page or on all pages in a document. The visual attributes such as font family, font size, color and bounding box are retrieved. The bounding boxes can be formed based on the font size of each word in a paragraph or line. In this module text information is retrieved.

B. The Merging of Words into Text Lines

In this module the extracted text is grouped into line segment. Initially the words or quads are sorted in the order of top down or bottom up which is based on center position of bonding boxes. Then the words are merged horizontally even if the vertical distance between two words is lesser than threshold. For line segment, font size and vertical centers of bounding box are taken as attributes and then they are computed by weighted averaging. From this logical text line fragmentation can be achieved.

C. The Grouping of Text Lines into Text Blocks

This module is to merge text segments into homogeneous text blocks. The problem of stemming in Bloches can be overcome by decoupling line space and font size and carefully detecting block boundaries during region growing.

The existing module doesn't consider the text objects such as titles, tables, lists and maps. Graphic recognition and integration with text segmentation are not considered in segmentation. In many cases, graphic components such as lines and color background are used to separate text. The detection of graphic components and their integration with text segmentation will greatly improve performance. Lists and tables are not considered in this text segmentation. Text belonging to map regions often has various orientations and excess character space. These are the most challenging cases for text segmentation.

IV. PROPOSED METHODOLY

In the proposed work the segmentation of both text and image components in e-book PDF format is considered. This overcomes the restriction of segmentation in tables, images and graphs in PDF document. In this system both text layer and image layer are taken into consideration for segmentation. Each layer segments its data independently. Finally the results of both text and image layers are merged together for final segmentation. Text segmentation and image segmentation are used for reusable purpose. This work is composed of text segmentation and image segmentation. In text segmentation text content in the PDF documents of e-book is segmented and in image segmentation image objects are segmented. Hence considering text and image objects in PDF document, the accuracy, precision, recall and F-measure for segmented documents will be increased.

A. Graph based Text Segmentation

In PDF document the words and quads are accessed through Word Finder and visual attributes are retrieved. Then PDWordGetNthQuad is used to get the bounding boxes of quads. The bounding boxes of each word or quad may vary from one to another word and it also varies from one line to another. Additionally the vertical center lines are computed from the bounding boxes. Further the words or quads are merged into text lines by selecting up a quad that has no longer been assigned a line identity to begin a new line segment. Then the line is extended by adding qualified quads on both left and right to the line. When no qualified quad can be added to the line, a new line is started until all quads are assigned a new line identity. This merging criterion is similar to Bloechle's. If horizontal distance between two words is smaller than threshold value those words are merged horizontally and we cannot consider the vertical distance between the words. Here we use font size, vertical center and width of the quad which are assigned as attributes to form line segment. After getting the text line segment we build homogeneous text blocks which avoid the pitfall by decoupling line space and font size. Relative difference between two line spaces is defined as $\Delta\,(line_{i,m}, line_{i,n})$ which is distance between vertical center lines and find the block boundary is found by comparing relative line space difference with a threshold value. For example if line i is block boundary it must satisfy the condition $\Delta\,(linespace_{i,m}, linespace_{i,n})>$ threshold. Similarly relative difference between font size is also calculated to find block boundary by using condition as $\Delta\,(fontsize_i, fontsize_n)>$ threshold where font size is the average of font sizes within the line *i*. The block boundary can be measured using line space and font size and also the type of block boundary. This can be explained as algorithm in the following section.

Text Segmentation Algorithm

```
Input: PDF document
Output: Text content in text pad
Step 1 : Access words and quads in PDF document
Step 2 : Check the document in horizontal reading
         direction
Step 3 : Calculate geometric center point and form
         block
Step4 : For boundary detection assign
         $linesegment_{boundary} = -1$
         $boundary_{id} = 0$
Step5: Define boundary for each line and increase
         the $boundary_{id}$ by 1.
         $boundary_{id} = boundary_{id} + 1$
Step 6: Merge the lines using queue
         lines.$boundary_{id} = boundary_{id}$
Step 7: Using Kruskal define the edge weight. Sort the edge weight in descending order and calculate mean and variance value
         $Mean = \frac{1}{vertices-1}\sum_{n=1}^{vertices-1} w(edges_n);$
         $Variance=\frac{1}{vertices-1}\sum_{n=1}^{vertices-1}[w(edges_n) - Mean]^2$
Step 8: Set the threshold value   $\Theta$=n*variance
Step 9: Remove the edges w($edges_c$)-Mean>$\Theta$
Step 10: Calculate angle distribution for segmentation
         Angle distribution=$\left[\tan^{-1}\frac{\Delta x_{i,j}}{\Delta y_{i,j}}\right]_{180°}$ //$\Delta y_{i,j} = |vertex_j(y) - vertex_i(y)|$, $\Delta x_{i,j} = |vertex_j(x) - vertex_i(x)|$
Step 11: Calculate line spacing and word spacing
         word spacing>interline spacing
         Merge according to width of block.
```

B. Graph based Image Segmentation

In image segmentation process a digital image is partitioned into a number of segments. The image may contain tables, lists, and graphs. By this segmentation process those contents are partitioned separately and saved in required location. It is a hybrid method. In this system both text and image layers are taken into consideration for segmentation. Each layer segments its data independently. Finally the results of both text and image layers are merged together for final segmentation.

For every layout analysis the image objects are not considered in the segmentation. This is the main goal of this research work considering both textual and image objects. It is considered that image objects are spatially far away from text blocks. Then cluster properties of Delaunay tessellation neighborhood system are used to reject non-textual objects. For layout segmentation only the clusters in the text region are considered. Hence it plays an important role in reflow the able reconstruction of PDF document structure. There are two systems available to segment identification for PDF document pages. One is from the PDF path which is directly used to extract geometric features of bounding boxes and to group the elements into desired physical segments by image streams.

Thus the bounding box ensures to include the elements for graphics but the smallest bounding box that encloses white background which is invisible to users. Such issues will return inaccuracy for graphic segmentation. Additionally, when the path and image elements for making a holistic graphic composite are vast in numbers, the computational speed will be reduced for the grouping process. One more option is to utilize the well- researched image based segmentation methods. In this research work image objects are processed as a separate layer using traditional image analysis method. From the visual perspective component analysis is obtained. Local text features describe the spatial closeness of graphic objects. Merging process is required to detect graphic composite holistically. Thresholds are set for connected component grouping based on inter text line spacing. As for graphics embedded or surrounded by text elements added, post processing of integration is handled.

Graph based Image Segmentation Algorithm:

```
Input: PDF document
Output: Image objects
Step 1: Components are analyzed from visual
        perspective
Step 2: Geometric features get from component
        analysis
Step 3: Define interline spacing and Set threshold
        Interline spacing < threshold
        Merge component objects
```

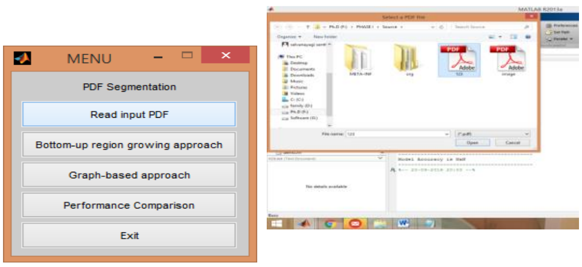
C. SVM Classification

In this module the objects are classified. SVM classifier uses test data and train data to classify the data. The output of layout analysis is bounding boxes of text line composite objects for text layer and graphic composite objects for image layer. Then from the analysis result of text or image layer a feature vector is extracted for each composite object. The source of feature extraction comes from different layers for classification which are indicated by character features of graphic components with zero. It is the main difference between the text and image features. For both textual content and image content segments, all the segmented sub images are saved, and image features describing texture spectrum are extracted. In this SVM is used for classification of text and image objects. Two-class SVM classifier is used for classification where mathematical expression for both isolated and embedded in PDF documents are detected with an accuracy of 90%. In this work, a larger sort of class labels is considered. The previous analysis within the document is taken into consideration for segment extraction where document is segmented into physical class labels such as footer text, body text, page number text, graphic text, and header. Multi-class SVM classifier is used in labeling task to discover the dissimilarity capacity of the presented features.

V. EXPERIMENTAL RESULTS

In our research work we used 50 e-book PDF documents to evaluate the performance of graph based approach in terms of accuracy, precision, recall, time measure and F-measure.In bottom up growing region approach it segments the text content in PDF document while in the graph based approach the image objects are segmented by using a hybrid method.

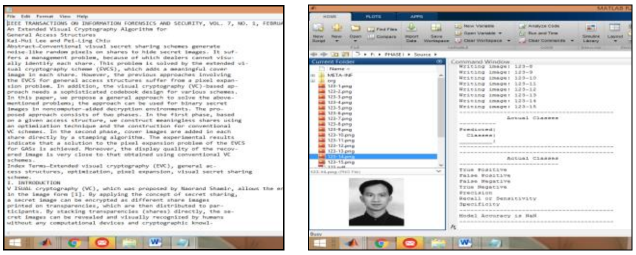
A. Sample input and output for graph based approach



```
Model Accuracy is NaN
```

Fig 1. Input PDF document

In the above Fig 1, PDF document is taken as input to evaluate text and image segmentation.



```
IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 7, NO. 1, FEBRUA
An Extended Visual Cryptography Algorithm for
General Access Structures
Kai-Hui Lee and Pei-Ling Chiu
Abstract—Conventional visual secret sharing schemes generate
noise-like random pixels on shares to hide secret images. It suf-
fers a management problem, because of which dealers cannot visu-
ally identify each share. This problem is solved by the extended vi-
sual cryptography scheme (EVCS), which adds a meaningful cover
image in each share. However, the previous approaches involving
the EVCS for general access structures suffer from a pixel expan-
sion problem. In addition, the visual cryptography (VC)-based ap-
proach needs a sophisticated codebook design for various schemes.
In this paper, we propose a general approach to solve the above-
mentioned problems; the approach can be used for binary secret
images in noncomputer-aided decryption environments. The pro-
posed approach consists of two phases. In the first phase, based
on a given access structure, we construct meaningless shares using
an optimization technique and the construction for conventional
VC schemes. In the second phase, cover images are added in each
share directly by a stamping algorithm. The experimental results
indicate that a solution to the pixel expansion problem of the EVCS
for GASs is achieved. Moreover, the display quality of the recov-
ered image is very close to that obtained using conventional VC
schemes.
Index Terms—Extended visual cryptography (EVC), general ac-
cess structures, optimization, pixel expansion, visual secret sharing
scheme.
I. INTRODUCTION
V ISUAL cryptography (VC), which was proposed by Naorand Shamir, allows the en
in the image form [1]. By applying the concept of secret sharing,
a secret image can be encrypted as different share images
printed on transparencies, which are then distributed to par-
ticipants. By stacking transparencies (shares) directly, the se-
cret images can be revealed and visually recognized by humans
without any computational devices and cryptographic knowl-
```

```
Writing image: 123-8
Writing image: 123-9
Writing image: 123-10
Writing image: 123-11
Writing image: 123-12
Writing image: 123-13
Writing image: 123-14
Writing image: 123-15
-------------------------------------------
                Actual Classes
----------
Predicted|
  Classes|
_________|
-------------------------------------------
-------------------------------------------
                Actual Classes
---------------------------
True Positive
False Positive
False Negative
True Negative
Precision
Recall or Sensitivity
Specificity
-------------------------------------------
Model Accuracy is NaN
-------------------------------------------
```

Fig 2. a) Text Segmentation b) Image Segmentation

Thus the image objects in the input PDF document such as graph, images, pictures and tables are segmented by using graph based approach which is shown in Fig 2(b). The text contents are segmented by using Kruskal's algorithm based approach as shown in Fig 2(a).The above fig 2, shows the first page text segmentation. Like this the text is segmented for whole document.

For performance evaluation, the proposed graph based text segmentation is compared with OCD document processing; XY cut segmentation and bottom up growing approach. Then for image segmentation the proposed graph based approach is compared with segmentation using connected components, Eigen vector and bottom up nearest neighbor application.

A. Accuracy

Social Accuracy is defined as the proportion of true positives and true negatives among the total number of results obtained. Accuracy is evaluated as,

$$\text{Accuracy} = \frac{(Truepositive + Truenegative)}{(Truepositive + Truenegative + Falsepositive + Falsenegetive)}$$

Fig.3 shows graph based segmentation for text which shows higher accuracy than the existing approaches.

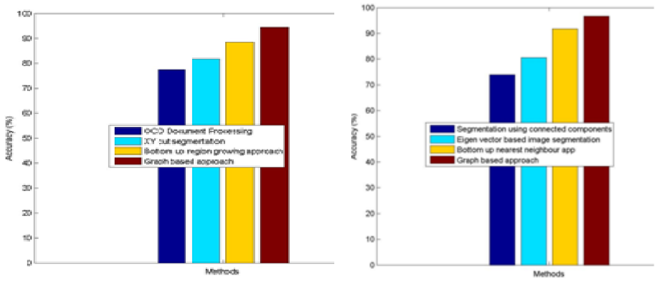


Fig 3. Accuracy on Graph based approach and existing for text segmentation and image segmentation

B. Precision

Precision value is evaluated according to the relevant information at true positive prediction, false positive.

$$Precision = \frac{Truepositive}{(Truepositive + Falsepositive)}$$

Fig.4 shows graph based segmentation which shows higher precision than the existing approaches.

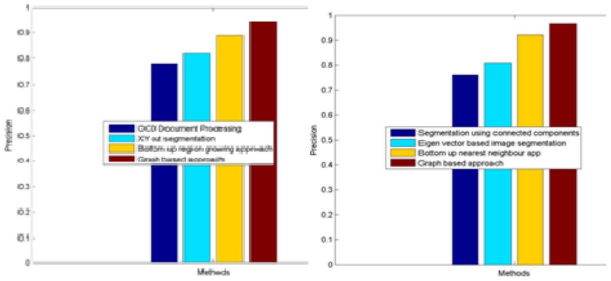


Fig 4. Precision on Graph based approach and existing for text segmentation and image segmentation

C.Recall

The Recall value is evaluated according to the retrieval of information at true positive prediction, false negative.

$$Recall = \frac{Truepositive}{(Truepositive + Falsenegative)}$$

Fig 5 shows graph based segmentation which shows higher recall than the existing approaches.

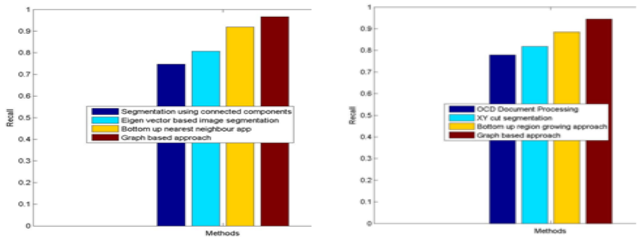


Fig 5. Recall as on Graph based approach and existing for text segmentation and image segmentation

D. F-measure

F-measure is calculated from the precision and recall value. It is calculated as:

$$f - measure = 2 \times \left(\frac{precision \times recall}{precision + recall}\right)$$

Fig 6 shows graph based segmentation which shows higher f-measure than the existing approaches.

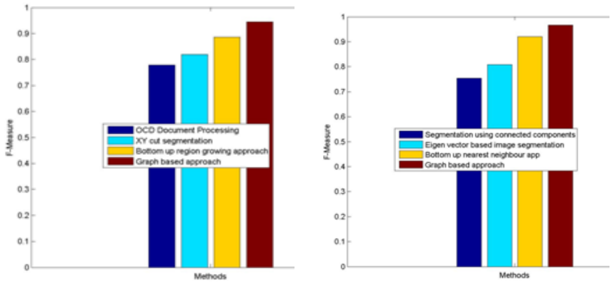


Fig 6 Comparison of F-measure between Graph based and Bottom up region growing approach

The values of precision, recall, accuracy, time measure and f-measure are tabulated in the following table 1

TABLE I: COMPARISON TABLE

	Accuracy	**Precision**	**Recall**	**F-Measure**
Text segmentation				
OCD Document Processing	77.5	0.8	0.8	0.8
XY cut segmentation	81.8	0.8	0.8	0.8
Bottom up region growing approach	88.5	0.9	0.9	0.9
Graph based approach	94.5	0.9	0.9	0.9
Image Segmentation				
Segmentation using connected components	73.9	0.8	0.7	0.8
Eigen vector based image segmentation	80.6	0.8	0.8	0.8
Bottom up nearest neighbor app	91.8	0.9	0.9	0.9

Graph based approach	96.6	1.0	1.0	1.0

From the experimental results it is proved that the proposed graph based approach more effectively segment the e-book PDF format than the existing segmentation approaches.

VI. CONCLUSION

In this work, e-book PDF format is segmented considering both text objects and image objects. This process involves text layer and image layers processed separately and finally the objects are classified by SVM classifier. Then experimental results are conducted in various e-book PDF documents to prove that the proposed graph-based approach is better than the existing bottom up region approach in terms of accuracy, precision, recall, f-measure, time measure.

REFERENCES

[1] Gupta, N., &Banga, V. K. (2012, April). Image Segmentation for Text Extraction. In 2nd International Conference on Electrical, Electronics and Civil Engineering (ICEECE’2012) (pp. 182-185).

[2] Adak, C. (2013, August). Unsupervised text extraction from G-maps. InHuman Computer Interactions (ICHCI), 2013 International Conference on (pp. 1-4). IEEE.

[3] Gautam, A. (2013). Segmentation of Text From Image Document.International Journal of Computer Science and Information Technologies,4(3), 538-540.

[4] Hassan, T. (2010). User-guided information extraction from print-oriented documents.

[5] O'Gorman, L. (1993). The document spectrum for page layout analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 15(11), 1162-1173.

[6] Yuan, Q., & Tan, C. L. (2001). Text extraction from gray scale document images using edge information. In Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on (pp. 302-306). IEEE.

[7] Lienhart, R., & Wernicke, A. (2002). Localizing and segmenting text in images and videos. IEEE Transactions on circuits and systems for video technology, 12(4), 256-268.

[8] Ranjini, S., &Sundaresan, M. (2013). Extraction and Recognition of Text From Digital English Comic Image Using Median Filter. International Journal on Computer Science and Engineering, 5(4), 238.

[9] Tehsin, S., Masood, A., &Kausar, S. (2014). Survey of Region-Based Text Extraction Techniques for Efficient Indexing of Image/Video Retrieval.International Journal of Image, Graphics and Signal Processing, 6(12), 53.

[10] Arulkumar V. "An Intelligent Technique for Uniquely Recognising Face and Finger Image Using Learning Vector Quantisation (LVQ)-based Template Key Generation." International Journal of Biomedical Engineering and Technology 26, no. 3/4 (February 2, 2018): 237-49. doi:10.1504/IJBET.2018.089951

[11] Hoang, T. V., &Tabbone, S. (2010, June). Text extraction from graphical document images using sparse representation. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (pp. 143-150). ACM.

[12] V Arulkumar, Charlyn Puspha Latha, Daniel Jr Dasig, "Concept of Implementing Big Data In Smart City: Applications, Services, Data Security In Accordance With Internet of Things and AI" International Journal of Recent Technology and Engineering 8, no. 3 (September 2019): 237-49. 2277-3878

[13] Arulkumar, C. V., and P. Vivekanandan. "Multi-feature based automatic face identification on kernel eigen spaces (KES) under unstable lighting conditions." Advanced Computing and Communication Systems, 2015 International Conference on. IEEE, 2015

[14] Kumari, S., & Vijay, R. (2012). Effect of symlet filter order on denoising of still images. Advanced Computing, 3(1), 137.

[15] Wu, L., Shivakumara, P., Lu, T., & Tan, C. L. (2015). A New Technique for Multi-Oriented Scene Text Line Detection and Tracking in Video. IEEE Transactions on Multimedia, 17(8), 1137-1152.

[16] Mehta, A., Parihar, A. S., & Mehta, N. (2015, September). Supervised classification of dermoscopic images using optimized fuzzy clustering based Multi-Layer Feed-forward Neural Network. In Computer, Communication and Control (IC4), 2015 International Conference on (pp. 1-6). IEEE.

[17] Green, R., & Oliver, C. (2013, November). Layout analysis of book pages. In2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013) (pp. 118-FIN123). IEEE.