Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 9 (18): 3317-3325, 2009
ISSN 1812-5654
© 2009 Asian Network for Scientific Information

Publisher Identifier Scheme for Printed Documents using Neural Networks

W.A.J. Rasheed and H.A. Ali
Department of Software Engineering, Faculty of Computer Science and Information Technology,
Isra Private University, Amman, 11622, Jordan

Abstract: This study investigates means used to extract embedded specifications of printing layout in a document when handled as an image rather than to recognize its characters and word constituents. These specifications are manifested by the most significant attributes frequently found in page printouts like advertisements, conference proceedings and magazines. The commonly used tools for printing document are word processors, provided by different software packages along with operating systems for PCS, like MS Windows and Mackintosh. Most of the supported packages in addition to Win Word specify the significant attributes of document formats to fall into font and paragraph design. Font design includes type, size and style characteristics. Moreover, line spacing and inter-character gaps provide another important attributes of paragraph specifications. These attributes were extracted, analyzed and exploited in this study with the aim of constructing the proposed publisher identification system. A three stage software scheme is proposed that consists of paragraph layout and font layout detection algorithms based on statistical measures followed by feed forward neural network identifier. This technique is implemented as a tool for the overall analysis and investigations. Several experiments have been conducted to validate the procedure of the designed system. The system achieved 95% successful publisher identification. This identification inaccuracy can be attributed to the poor quality of printing in addition to the effect of noise. Hence, it can be considered as acceptable performance measure for detection and identification purposes once bad quality printed samples are excluded.

Key words: Text analysis, text identification, character recognition, image processing, neural networks

INTRODUCTION

Publication facilities have been developed in the recent decades with high quality attributes. This can be well recognized by looking at the books, newspapers, magazines and other means of information repositories. Earlier trends had focused on the information itself rather than the demonstration capabilities and step-by-step great amount of attention have been oriented towards the printing facilities of coloring, font design and many other items (Sharma *et al*., 1998).

Shape analysis and characterization of literature has been a field of intense activity in computer vision, image processing and pattern recognition (Theodoridis and Koutroumbas, 2003). Shape of an object represents an important part of the Information which can be based on a description of the boundaries or the whole region of the object and implementing OCR techniques. It can be classified as structural, in which the object is broken into parts and properties of those parts are encoded. In addition, shape representations can be monolithic or hierarchical. Monolithic encode only properties at a single scale, while hierarchical consist of describing the shape at multiple discrete scales. Different methods for shape similarity and recognition have been developed. For example, Sinha and Giardina (1990) used morphological function to decompose a given shape into primitive parts. The properties and relationships among these primitives were used to describe different objects. Another technique, the potential-based approach developed by Chuang (1996), identifies the best match from a selected group of shape templates by measuring the repulsive force and torque when the template and sample are put to interact through a potential field. Kupeev and Wolfson (1996) used a weighted graph to represents the structure and the quantitative elements of a contour for estimation of shape similarity. Tsang and Tsang (2006) presented a neighborhood vector representation as shape parameter for binary images. This method is based on the pixel neighborhood relation. The results show that the shape parameter can be used as a general method of feature extraction for problems in image processing and pattern recognition.

Variety of classification techniques uses page segmentation for pattern recognition. Many of the existing page segmentation algorithms have region

Corresponding Author: W.A.J. Rasheed, Department of Software Engineering,
Faculty of Computer Science and Information Technology, Isra Private University,
Amman, 11622, Jordan

classification modules embedded in the systems. Various ways to segment and classify regions have been proposed ranging from purely top-down approaches that recursively split a page into smaller components, to purely bottom-up approaches that attempt to cluster individually connected components into larger and larger entities. Top-down approaches usually segment the page into homogenous, extracting features and then classify them into appropriate categories. A number of different features have been proposed for region classification. Some of these include the area of the connected component of a block, number of black pixels in the region, the mean horizontal black run length, component width height ratio, component density, mean length of black intervals to the mean length of white intervals, number of black intervals over a certain length, feature-based interaction map (Chetverikov *et al*., 1996), texture discrimination masks (Pappas *et al*., 2001), periodicity measures, the measures of visual attention, i.e., legibility, complexity, attractiveness, etc (Ryu *et al*., 2000). Bottom-up approaches usually classify each image pixel first and then group the pixels into regions, such as multi-scale texture segmentation (Wu *et al*., 1999), texture discrimination masks (Jain and Zhong, 1995), mask based local textural characteristics extraction (Williams and Alder, 1996). Some approaches are based on geometric relations that group the pixels according to various patterns and then use simple statistical features to classify them (Mitchell and Yan, 2001).

Moreover, Non-text regions were examined using neural network based region identification algorithm as key component of a document recognition system implementing segmentation. They were classified into text, graphic, photo and other region types (Andersen and Zhang, 2003), Low quality images of newspaper documents obtained from microfilmed archives and the results compare favorably with other results reported elsewhere.

The invading era of computer systems onto this area of application had played great role in the occurring achievements. With computer implementation, two media emerged in information handling; either a print out on paper sheet or a file saved on storage medium (commonly termed as hard or soft copy modes). Computers with their broad analogous systems of digital instruments are nowadays regarded as the main data developing tools. In accordance to the nature of these systems, information generation can be utilized in two different processes; writing (i.e., typing) and printing (which in turn distinguishes between the soft and hard copy modes). Along with these digital systems, wide range of word processors have been developed by Besser (1996). The ever growing memory sizes support heavily these word processors to demonstrate the documents with net and plain character generators. The outcome of the devoted care on these missions exceeds the informative boundaries and stimulates the artistic attentions in writing to appear as drawing and painting resultants (DeSoi *et al*., 1990). Any information repository these days, despite of its sort whether a magazine newspaper, book, or the like can be simply specified by the artistic appearance of the cover page or the text features themselves. Correspondingly, for the sake of standardization, calls for papers or articles by various institutes or publishers insist on following up publisher regulations or template in order to be considered for acceptance. Hence, these regulations could indirectly be considered as identifiers for the publisher or document generation source.

Based on the last deduction, the present study has adopted an automatic detection system designed to extract the writing specifications of any printed document. The increasing number of information producers with their distinct specifications of writing made the identification activity to invoke a data base utilization in order to integrate the process of identification. By this database the extracted specification can be associated with that of the producer of the information and supports a decision-making activity.

MATERIALS AND METHODS

Broadly speaking, document format is characterized by paragraph layout and font attributes design. The former involves line spacing, number of text columns and gap dimension properties, while the latter denotes font type, size and style selections (Wieser and Pinz, 1993). Based on these two main categories of document description, the proposed detection system is designed to involve two successive procedures followed by an identification process. Therefore, implementation of the proposed system consists of three phases; paragraph layout detection phase, font attribute detection phase and identification phase as shown diagrammatically in Fig. 1 and briefly outlined below. The first two phases determine the attributes which are handed over next process in order to identify the most probable source of the document.

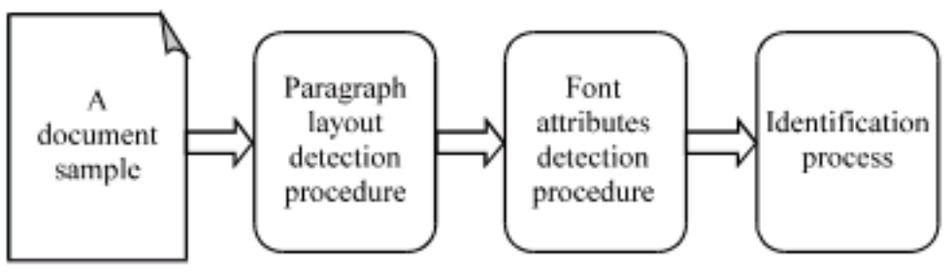


Fig. 1: The proposed identification system

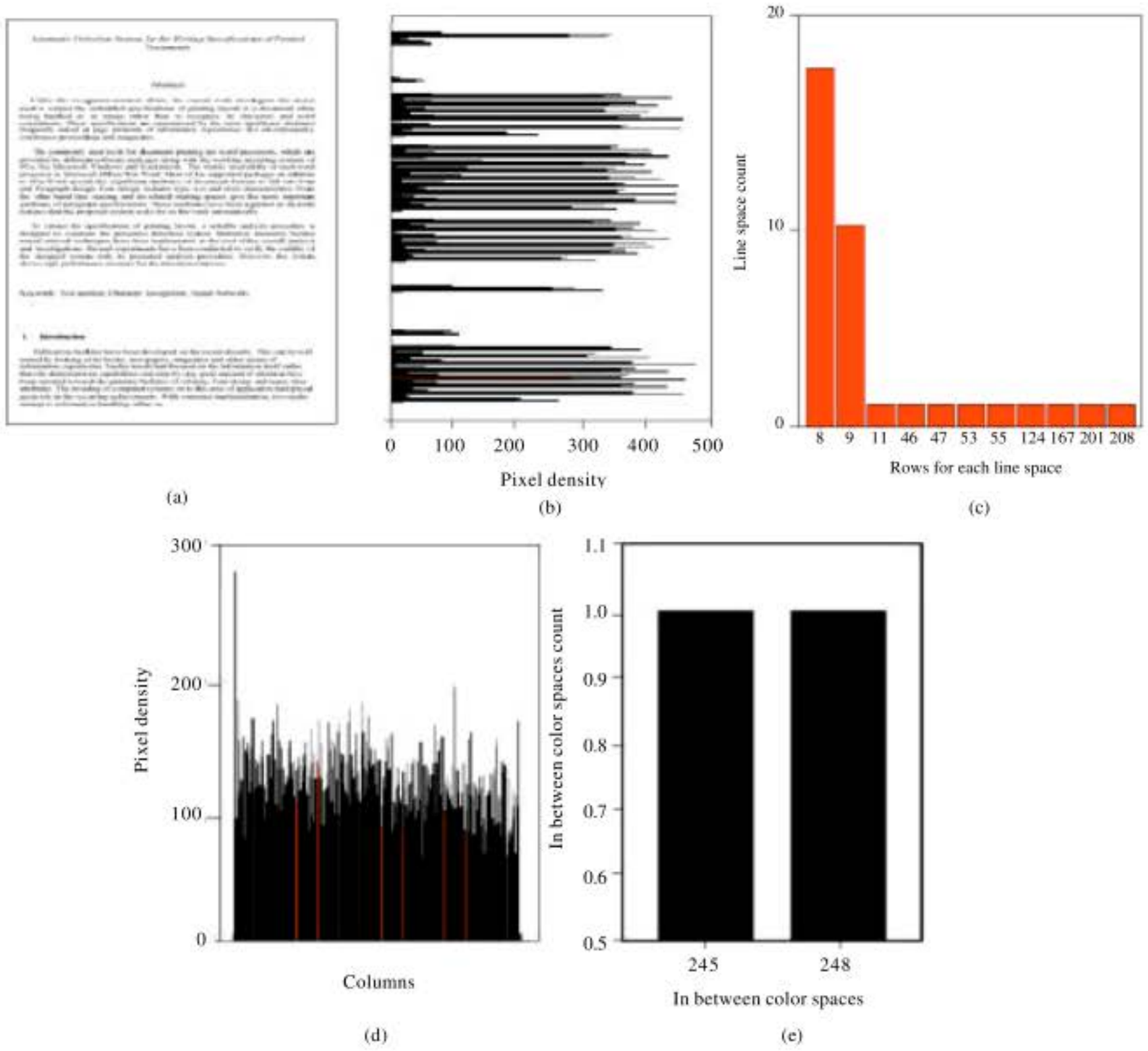


Fig. 2: Document example 1, (a) input document, (b) row wise pixel density, (c) row wise line spacing frequency distribution, (d) column wise pixel density and (e) column wise line spacing frequency distribution

This process is carried out, utilizing a neural network association scheme. It may be added that the design and implementation of the proposed scheme was carried out for about ten months starting in April 2008 at Isra Private University, Jordan.

Phase one: Paragraph layout detection: Principally, the current procedure depends thoroughly on pixel density that is acquired in two modes of row wise and column wise calculations. These density distributions are traced so as to gather the underlined spacing characteristics with the aid of the computed frequency distribution of line spacing (Kaiman, 1968). Examples showing the essential data of pixel density acquisition and computed line spacing frequency distribution for different text contents along the two modes are shown in Fig. 2a-e and 3a-e. However, paragraph setup characteristics are determined according to the following notes:

- From row frequency distribution, line spacing measure tagged to maximum occurrences represents document line spacing attribute 1
- Header and bottom spaces denote attribute 2 and attribute 3. These attributes are computed by tracing and counting the empty rows initiated with the first (upper) row and with the last (lower) row, respectively
- With the exception of considering the header and bottom spaces, in-between paragraph measure is that whose related accumulated number should satisfy Eq. 1

$$\text{In-between paragraph measure} = \text{Amount of paragraphs} - 1 \quad (1)$$

This measure defines attribute 4, which is determined throughout two stages. In the first, paragraph number is computed using the adjacent non empty lines (separated

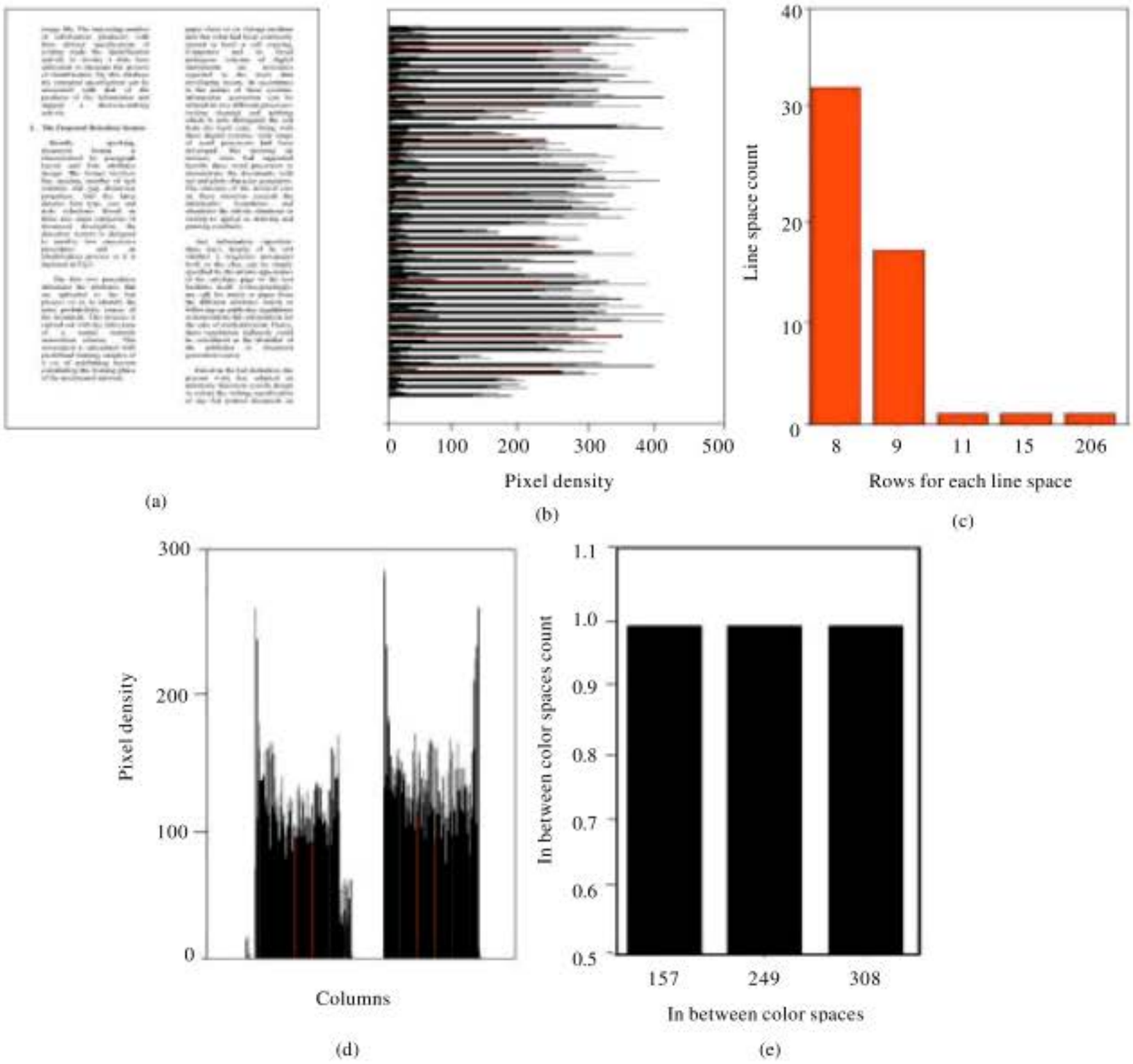


Fig. 3: Document example 2, (a) Input document, (b) Row wise pixel density, (c) Row wise line spacing frequency distribution, (d) Column wise pixel density, (e) Column wise line spacing frequency distribution

by attribute 1) counting that is carried out using pixel density characteristics. Whereas, in the second stage, any other number of line spacing other than attribute 1 and satisfies Eq. 1 will be considered as the target attribute (attribute 4).

In particular circumstances, a verification process resumes execution in order to justify the results gained by this stage. This process is stimulated when in-between paragraph space does not differ from the standard line spacing, attribute 1. Here, a tracing mechanism is started to analyze pixel density diagram extracting attribute 4.

- From column pixel density characteristics both left and right margins are calculated so as to define attribute 5 and 6, respectively. Obviously the algorithm used is similar to that used for header and bottom spaces with the exception of using the column mode density instead of the row mode acquisition
- Number of written column represents attribute 7. The procedure used here is also similar to the one used in Paragraph Amount evaluation but using the column mode instead
- The column in-between measure, considered as attribute 8, is computed in similar way as that used in addressing the process of evaluating the In-between line spacing of attribute 4, but with difference of using the column mode instead of the row mode of data investigation
- Finally, written column width is determined from column pixel density characteristics and represents the last paragraph layout, attribute 9. It might be noted that the column width is considered by ignoring the left and right margins

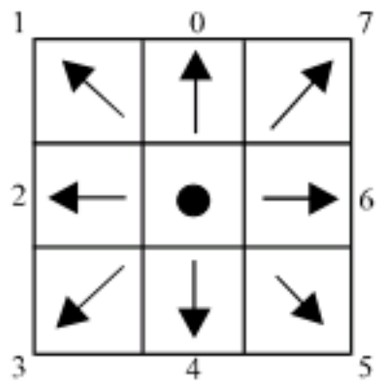
Phase two: Font attributes detection: Due to the great modalities of font and its associated characteristics, the function of the second procedure is divided in two sub steps of processing. The first deals with general detection of the font and can be termed as character recognition stage, whereas the second step investigates the concise descriptions of the underlined character. The procedure in whole adopts a small set of alphabetical character patterns to sketch the overall activities. This set involves the first characters from A to F. The investigations fixes the capital mode of these characters in order to minimize the variations as much as possible and to compress the required database needed for the decision making phase. Moreover, traditional steps of commonly used recognition system are used to design the different activities of the first step of the current procedure (Oh *et al*., 2001). These activities denote line segmentation and character segmentation as first trends. Freeman chain technique is adopted for this purpose (Tsang *et al*., 1999). The most significant feature of base line position is contributed in this stage. The obtained chain hence is subjected to feature extraction operations oriented towards the following features:

- Horizontal propagation of the character
- Vertical propagation of the character
- The relative propagation with the base line
- Whole existence and the count of the existing ones

Based on the above features, the first stage candidates is the character with its global definition. When the gathered features do not convince any available character subset, another character is segmented and processed until a match status is achieved.

The match will in turn commit the character represented by its chain to the second stage of this procedure.

The second stage is designed to investigate the frequency analysis and the statistical measures of the provided character from previous stage. The statistical analysis and statistical classification adopted in the implementation of this stage follows methods outline by Michie *et al*. (1994) and Härdle and Simar (2007). All the characters mentioned earlier are traced along their chains to build the classification repository. This repository is constructed with different tables, involving the eight Freeman chain frequencies. Figure 4a-c show the basic principle of Freeman chain construction and how they are applied to represent the character A as an example. Then Table 1 shows such characteristics of font description summary constructed for character A for a specific font type, style and range of font sizes. For the example of character A written in Times New Roman font with different sizes, the absolute direction frequency count summary and the normalized frequency count with reference to direction 7 is shown in Table 1.



(a)

A

(b)

(c) **{4545432226666543666110101010 0}**

Fig. 4: Freeman chain extraction for letter A as an example

Table 1: Direction frequency count for character A in Times New Roman font with different sizes

	Character size (font size)															
Directions	8	9	10	11	12	14	16	18	20	22	24	26	28	36	48	72
The absolute frequency count of the directions																
0	5	6	5	1	4	6	6	8	8	8	11	11	16	22	31	45
1	4	4	5	0	5	5	6	6	8	8	9	11	12	15	22	33
2	1	0	0	0	3	6	6	9	8	8	11	10	7	13	10	16
3	4	5	5	0	2	2	2	2	4	4	4	4	12	11	22	33
4	5	5	4	1	4	5	6	8	6	6	9	11	16	23	30	43
5	1	1	2	0	3	4	4	4	6	6	7	7	2	8	4	6
6	7	7	7	0	7	9	10	13	14	14	17	18	27	26	47	72
7	1	1	1	0	0	0	0	0	0	0	0	0	2	5	3	4
The normalized frequency (direction 7 is the reference)																
0	4	5	4	1	4	6	6	8	8	8	11	11	14	17	28	41
1	3	3	4	0	5	5	6	6	8	8	9	11	10	10	19	29
2	0	0	0	0	3	6	6	9	8	8	11	10	5	8	7	12
3	3	4	4	0	2	2	2	2	4	4	4	4	10	6	14	20
4	4	4	4	1	4	5	6	8	6	6	9	11	14	18	27	39
5	0	0	1	0	3	4	4	4	6	6	7	7	0	3	1	2
6	6	6	6	0	7	9	10	13	14	14	17	18	25	21	44	68
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

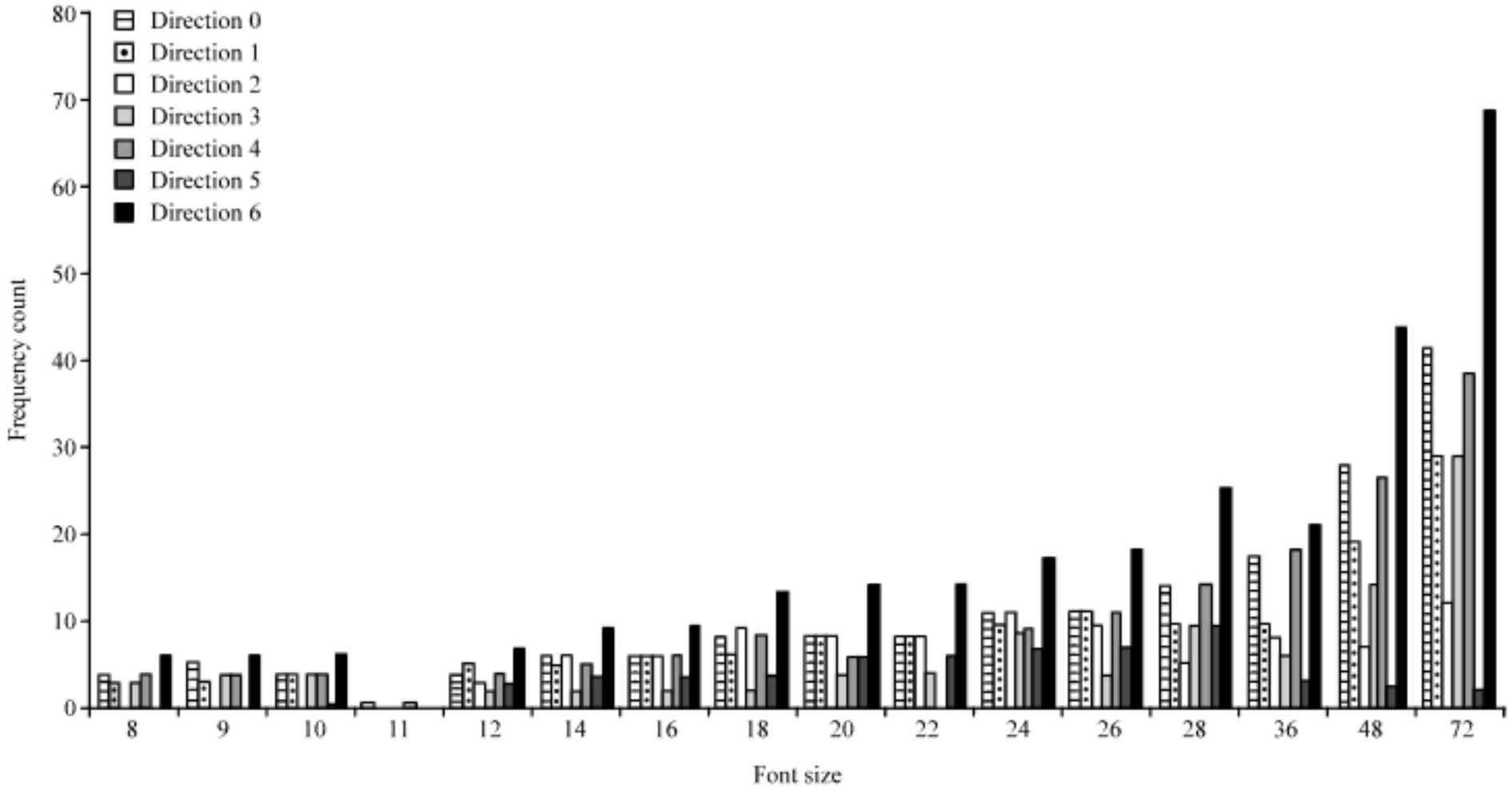


Fig. 5: Frequency count of direction in Times New Roman font for character A with different font sizes

Finally, the direction frequency counts normalized to direction 7 is plotted as a function of the font size for letter A, written in Times New Roman type in Fig. 5. Information of any character provided from the first stage will be compared with the available behaviors of the repository to decide three parameters of its main features of font, i.e., size, font style and font type, taken into consideration that the bold attribute is ignored.

Phase three: The identification process: So far details of the foregoing two procedures are summarized by 12 different features (attributes). Based on this amount of features, a neural network design is used as an identification tool (Crawford-Hines and Anderson, 1997).

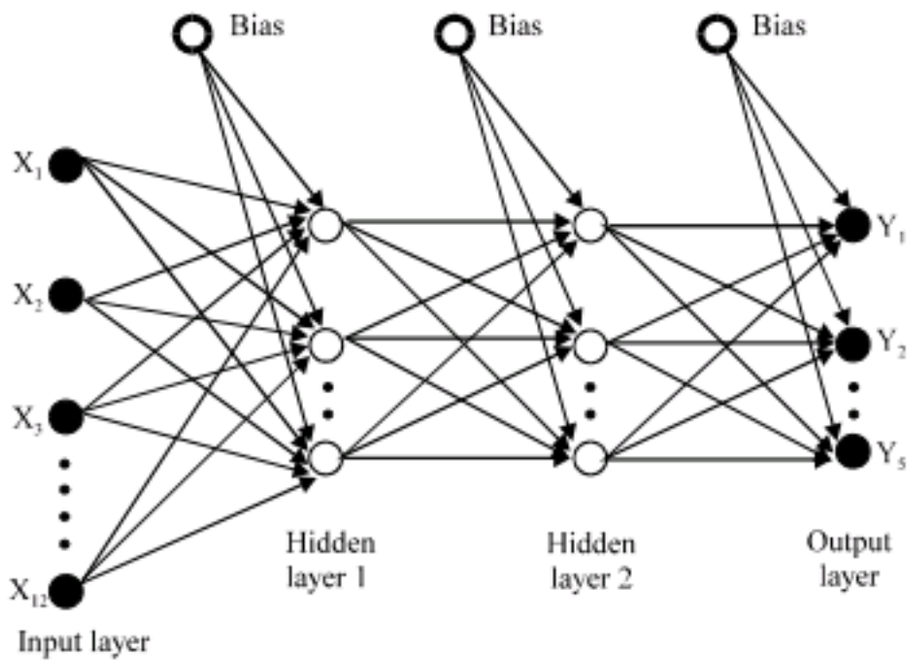


Fig. 6: The adopted identification neural network

The neural network structure: The overall processing of the proposed identification system is summarized by the final stage of neural network utilization structure. The main function of the subjected structure relies on the association capability of the neural network and its generalization power (Feng *et al*., 2006; Inohira *et al*., 2008; Rasheed and Ali, 2009). It is worth mentioning that classifiers other than neural networks for pattern analysis and identification add extra programming burdens on the system, while neural network classifiers reduce data dimensionality to a great extent. This is due to their reliance on features instead of abstract data configuration. Feed forward network can successfully achieve this task (Hagan *et al*., 2002). A net with twelve input layers, two hidden layers of five neurons each and five neurons output layer is used, as shown in Fig. 6. All neurons of the hidden and output layers get input signal from bias neurons. Obviously, the number of the neurons at the output layer specifies the number of the selected publishers identifying the features at the input layer of the scheme. Hence, 32 binary codes denote the whole domain of the identification process under consideration. The five outputs of the network are chosen based on the selected experiment sources of the documents.

On the other hand, input measurements for the features collected have been normalized to be used as proper data for the training purposes. Then a training table of twelve input values and five output values sums up all the needed data of the training. Genetic algorithm

search technique has been adapted for the evaluation of the weights for neural network. Based on the collected data, training has been smoothly tolerated the weights to a satisfactory error level. This scheme on the testing phase has showed suitable outcome of identification that is found easy to be recovered with its associated error value.

Different experiments were conducted with samples of publishers' documents for network training purposes. For each experiment, samples of provenances were taken depending on the allowable coding space that covers the selected provenances. Five samples for each publisher were chosen as data for each publisher in order to include as most expected variations as possible in document contents. Therefore, 160 overall samples are involved as training patterns in this prototype design. Genetic algorithm search technique was adopted for the training of the neural network of the identification process.

RESULTS

Although, diagrams, photos, footers and headers are important features of any published document, this research was limited to the most important characterizing properties of publisher identification which is writing and page layout. Therefore, at this stage, diagrams, photos, footers and headers have not been considered. As the designed neural network in the identification process phase have five output, i.e., allowing for 32 different outputs, different samples of only 29 publishers were selected to be experimented with while the remaining three possible outputs (29, 30 and 31) were reserved for unrecognized printing materials or unknown publisher. In order to have a better representation of any publisher printout layout, four pages of published material are selected as full training sample spans for each publisher. This strategy is followed as a measure to avoid unexpected occurrence of non-standard contents, such as the inclusion of introductory paragraph that disturbs the regular print layout.

The main methodology in computing the parameters is based on analysis of the four pages samples in order to extract the outcome of Fig. 2b. Then Fig. 2b in turn is transformed into Fig. 2c, which demonstrates the bases of measurement. These measurements produce the first four parameters (attributes 1 to 4) of the paragraph layout detection phase. Along the same phase, Fig. 2d is transformed into Fig. 2e, determining the next parameters (attributes 5 to 9). Furthermore, Fig. 3 shows results for different publisher printout sample having two columns text instead of one.

It is worth mentioning that within the second phase, computation passes through two stages, namely character recognition and font feature detection processes. Traditional methodologies of recognition are conducted to identify any selected character (Oh *et al*., 2001). This selection depends on whether its related Dbase is already available or not. In case it is not available, the algorithm resumes processing to select another character of the text. Once a character is defined, the second stage determines the remaining three parameters (i.e., font size, style and type). This process is achieved with the aid of the familiar Freemann description model (Tsang *et al*., 1999).

It is known that fonts can be supplemented to any word processor as additional library. A fast detection of the availability for the defined character can be detected by applying the normalized frequency count of Table 1-b to its content. Once a character is found, concise description can be further extracted through matching its frequency count with that of Table 1. For simplicity and less storage requirement, this prototype study is based on five characters only, therefore five tables of the form shown in Table 1 are constructed and embedded in the system.

For testing purposes, different publishers' printout documents were fed to the system. The correct detection of publishers' identification achievement was about 95%. Failure of identification can be attributed to the illed processing in phase two. Incorrect recognition of noising text stands for the unsuccessful identifications. Poor quality of printed documents that do not supply adequate data statistics may also result in similar consequences.

DISCUSSION

Huge differences in publishing styles and artistic products nowadays are attributed to the wide variety of word processing tools and printing hardware provided by computer technology advancement. For various reasons, such varieties necessitate the attempts to recognized and distinguish publishing resources and their classifications.

Although, not many published study is available on the approach of publisher identification, some relevant techniques might be quoted. For example, Chetverikov *et al*. (1996) uses texture features reported text identification accuracy of 97-98% and non-text identification accuracy of 84-89% while achieving a total accuracy of 96%. Ryu *et al*. (2000) used multi-scale analysis and top-down approach for segmenting and a periodicity for classification reported an accuracy of 97.1% for image identification. Fast scan method and classification for pattern extraction of document by

Mitchell and Yan (2001) resulted into a total accuracy of 98%. Pappas *et al*. (2001) uses a simple mask that makes use of the different correlation properties to classify the region reported a total accuracy of 98.3%. For identification of newspaper documents using neural networks, (Andersen and Zhang, 2003) reported an identification accuracies in the range from 87.5-98% depending on text and non-text contents in the tested materials. Neighborhood vector representation as shape parameter for binary images adopted by Tsang and Tsang (2006) used neural network classifier claimed to be satisfactory, but it was for hand written characters.

Therefore, the results reported in this article, which produced an overall accuracy of about 95% can be seen as approximately equal to most of the above mentioned results, slightly worse than those reported in (Ryu *et al*., 2000) but better than some results of (Chetverikov *et al*., 1996). The identification achieved in this paper would be considered encouraging due to the fact that further efforts to include diagrams, photos, footers and headers as extra attributes at detection and identification stages would certainly enhance the recognition results.

Furthermore, the study reported in this research involves wide range of computation techniques that includes image processing, segmentation, statistics and neural networks capabilities to be utilized for identification. Within the authors' framework, many tasks relied primitively on two objectives, i.e., scanning feature detection and character recognition. Such technique with its analysis procedure can be regarded as a key to investigate the era of style mining of websites. These websites despite their varieties and modalities are classified corresponding to limited publishers. The main criteria appearing to describe the publisher in this application denote the coloring and animation effects.

CONCLUSIONS

Having achieved about 95% correct recognition of publisher printouts identification, the proposed prototype scheme appears to be promising and could be improved with the addition of more printing features and/or network elaboration. This system uses only twelve features of the published material and yet may encourage the implementation of the detection system to be adopted with automatic schemes that could be used effectively for document identification in different investigation departments.

Failure of the publisher detection is captured and can be attributed to lack of enough information or some fuzziness in the fed documents the system. This matter is studied and is encompassed with the generalization capability of the used neural network. Furthermore, other document features such as footers, headers, photos, diagrams and colors may be added to obtain more attributes in order to improve system performance.

REFERENCES

Andersen, T. and W. Zhang, 2003. Features for neural net based region identification of newspaper documents. Proc. Int. Conf. Document Anal. Recognit., 1: 403-407.

Besser, H., 1996. Designing a digital documents curriculum. Proceedings of the 29th Hawaii International Conference on System Sciences, Wailea, HI, USA., Jan. 3-6, IEEE Computer Society Press, Los Alamitos, CA, pp: 153-158.

Chetverikov, D., J. Liang, J. Komuves and R.M. Haralick, 1996. Zone classification using texture features. Proc. Int. Conf. Pattern Recognit., 3: 676-680.

Chuang, J.H., 1996. A potential-based approach for shape matching and recognition. Pattern Recognit., 29: 463-470.

Crawford-Hines, S. and C.W. Anderson, 1997. Neural nets in boundary tracing tasks. Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, Sept. 24-26, Amelia Island, FL, USA., pp: 207-215.

DeSoi, J., M. Lease, W. Lively, S. Sheppard and T. Slade, 1990. A graphical environment for user-interface design and development. Software Eng. J., 5: 289-299.

Feng, N., F. Wang and Y. Qiu, 2006. Novel approach for promoting the generalization ability of neural networks. Int. J. Signal Proc. Waset, 2: 131-135.

Hagan, M.T., H.B. Demuth and M. Beale, 2002. Neural Network Design. China Machine Press, CITIC Publishing House, Beijing.

Härdle, W. and L. Simar, 2007. Applied Multivariate Statistical Analysis, Leopold. 2nd Edn., Springer-Verlag, Berlin, Heidelberg, ISBN: 978-3-540-72243-4.

Inohira, E., T. Uoi and H. Yokoi, 2008. Generalization capability of neural networks for generation of coordinated motion of hybrid prosthesis with a healthy arm. Int. J. Innovative Comput. Inform. Control, 4: 471-484.

Jain, A.K. and Y. Zhong, 1995. Page segmentation using texture discrimination masks. Proc. Int. Conf. Image Process., 3: 308-311.

Kaiman, A., 1968. Computer-aided publications editor. IEEE Trans. Eng. Writ. Speech, 11: 65-75.

Kupeev, K.Y. and H.J. Wolfson, 1996. A new method of estimating shape similarity. Pattern Recog. Lett., 17: 873-887.

Michie, D., D.J. Spiegelhalter and C.C. Taylor, 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood Series in Artificial Intelligence, Prentice Hall, ISBN-13: 978-0131063600.

Mitchell, P.E. and H. Yan, 2001. Newspaper document analysis featuring connected line segmentation. Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing-Volume 11, Sydney, Australia, 2001, Australian Computer Society Inc., Darlinghurst, Australia, pp: 77-81.

Oh, I.S., J.S. Lee and C.Y. Suen, 2001. A class-modularity for character recognition. Proceedings of the 6th International Conference on Document Analysis and Recognition, Sept. 10-13, Seattle, WA, USA., pp: 64-68.

Pappas, T.N., S.H. Tseng and D.A. Kosiba, 2001. A robust and efficient algorithm for bi-level document block classification. Int. Conf. Image Process., 1: 1122-1125.

Rasheed, W.A.J. and H.A. Ali, 2009. Generalization aspect of neural networks on upgrading assimilation structure into accommodating scheme. J. Comput. Sci., 5: 177-183.

Ryu, D., S. Kang and S. Lee, 2000. Parameter-independent geometric document layout analysis. Proc. Int. Conf. Pattern Recognit., 4: 397-400.

Sharma, G., M.J. Vrhel and H. Joel Trussell, 1998. Color Imaging for Multimedia. Proc. IEEE, 86: 1088-1108.

Sinha, D. and C.R. Giardina, 1990. Discrete black and white object recognition via morphological functions. IEEE Trans. Pattern Anal. Mach. Intell., 12: 275-293.

Theodoridis, S. and K. Koutroumbas, 2003. Pattern Recognition. 2nd Edn., Acad. Press, San Diego, CA, USA, ISBN: 0126858756.

Tsang, I.J., I.R. Tsang and D. Van Dyck, 1999. Image coding using Neighborhood Relations. Pattern Recog. Lett., 20: 1279-1286.

Tsang, I.R. and I.I. Tsang, 2006. Neighborhood vector as shape parameter for pattern recognition. Proceedings of the International Joint Conference on Neural Networks, Jul. 16-21, Vancouver, BC, Canada, pp: 3204-3209.

Wieser, J. and A. Pinz, 1993. Layout and analysis: Finding text, titles and photos in digital images of newspaper pages. Proceedings of the 2nd International Conference on Document Analysis, Oct. 20-22, Recognition, UK., pp: 774-777.

Williams, P.S. and M.D. Alder, 1996. Generic texture analysis applied to newspaper segmentation. IEEE Proc. Int. Conf. Neural Networks, 3: 1664-1669.

Wu, V., R. Manmatha and E.M. Riseman, 1999. Text finder: An automatic system to detect and recognize text in images. IEEE Trans. Pattern Anal. Mach. Intell., 21: 1224-1229.