Spotting a Target Advertisement in a Source Document Image

[1] P. Nagabhushan, [2] Sahana D.Gowda

ABSTRACT

Many times one would search for a particular content in an entire document image, such as if a company has released an advertisement, the concerned manager would be interested in locating the advertisement in a newspaper. This problem is termed as *advertiser's problem*. The target text component is a-priori known in terms of its structure and content, however the complete details of the layout where the target image appears in a source document image will not be known. Given the fact that font size and font style employed by a source document is generally known before hand, the problem is to locate the target image in the source document image. The major objective of this paper is to accomplish this task of spotting the target text component with the additional and the major constraint that for such a purpose any language processor such as OCR neither should be employed nor would be available. For this purpose the components in a document image are characterized in terms of energy values called entropy [4, 9]. The matching of the entropy is proposed for locating the target component in a source document image.

Keywords : Spotting, Target advertisement, Source document image, Conventional Entropy Quantifier (CEQ), Spatial Entropy Quantifier (SEQ), Frequency Histogram, Spatial Histogram, Regression Line, Distance Measure.

[1&2] Department of Studies in Computer Science, University of Mysore, Mysore-570006, India. Email : pnagabhushan @hotmail.com, sahanagowda@rediffmail.com

1. Introduction

Searching for the location of an advertisement amidst vast contents in a newspaper, to check for the correctness of the advertisement by the concerned manager of the company which has released the advertisement is a problem belonging to the class of target image search in a given document image. This specific problem may be described as **advertiser's problem**. There are many problems which are similar to this, such as given an article, an author/a reader would be interested in first locating its position in the source document. Parents who have advertised inviting for marriage proposals for their son/ daughter in matrimonial columns, would search to confirm that their advertisement has appeared without any errors. In all such cases, it is not possible to locate the target by reading all the contents of the source document. It would be more comfortable if a mechanism, which is quick and does not involve the process of *reading* could be employed for the purpose. For the purpose of automation the target and the source documents have to be scanned and therefore we restate the problem as the problem of locating a target image in a source document image or alternatively the source could already be available in the electronic media. Some of the facts, which could favor the development of a procedure for the purpose, are –(i) The structure and contents of the target image are a-priori known. (ii) The publisher's choice of font type for a given source document is generally known before hand. However finding the location and the layout of the target in the source are the major factors involved in spotting process as layout of the target text component will not be known since the publisher would have fit the target text component according to the availability of space in the source document. The condition imposed in this research is that for such a purpose any language processing can not be employed, because *OCRing* the entire source document to locate a *small* target is not advisable, *OCRing* is not possible if the source/target documents are multilingual [1] and there are still some languages for which OCRs are not available [1].

Towards achieving the aforesaid objective, some of our earlier papers [4,5,8,9,10] are the motivators. In [10] we have proposed methods for quantifying the text components at various hierarchical levels – paragraph (coarse), line (medium) and word (fine) level, in terms of their energy values called entropy. Two types of entropy parameters are defined-Conventional Entropy Quantifier(CEQ)[5,9,11] and Spatial Entropy Quantifier(SEQ)[4]. We have used the entropy values to establish the equivalence between the two text document images. As a further improvement [4, 9] we have proposed histogramming the entropy values for a component to reduce the computational complexities.

In this research paper it is proposed to extend the above ideas of entropy computations to search, locate and establish the equivalence of the target image in the source document image. For the purpose it is proposed to segment the source document into blocks. The blocks further could be made up of paragraphs. A paragraph consists of sub paragraphs and so on. Similarly a target image if has more than one block, then blocks would in generally be spatially adjacent. The target blocks decomposed into levels up to paragraphs/subparagraphs are quantified in terms of entropy values and matching the entropies is proposed to locate the target in the source document image.

Some works for spotting a word as a target [1,2,12,15] are reported. Some of these are OCR based and some of them are structure based. But research works beyond the word level are not reported, particularly without employing OCR. However in our paper a block or a set of blocks comprising any advertisement / article is the target for spotting.

The rest of the paper is organized as follows. The complexities of an advertiser's problem are brought out in section-2. The features employed-geometrical features and entropy features for establishing the equivalence between two document images without *OCRing* are briefed in section-3 for the purpose of providing a complete reading and to make this paper a self contained one, although these are dealt in detail in our earlier papers [4,8,9,10,11]. The proposed algorithmic model is presented in section-4. In section –5 experimental details are reported. The paper concludes in section-6.

2. Complexities In The Problem

An advertisement can be as small as a paragraph containing only text. At higher level it is necessary to interpret an advertisement as a block containing many sub blocks as shown in figure-6. Every block/sub block may be made up of a paragraph or sub paragraphs. In our paper we propose to decompose an advertisement up to paragraph/sub paragraph level. For the reason which becomes clear subsequently, it is not required to go up to further finer levels such as lines and words. Generally the publisher's choice of font style, font size and other dimensions could be known. Yet, if the target is composed of adjacently placed blocks more than one in number, the layout, which could be finalized by the publisher for inserting an advertisement in the source, depends upon the availability of space. The different ways in which a case of three blocks could be placed in the source are shown in figure-1. Further, as seen in figure-6, within a block it is possible that the sub blocks and paragraphs could be shuffled without damaging the content /meaning of the advertisement. Different possibilities of such cases are schematically exemplified in figure-2.

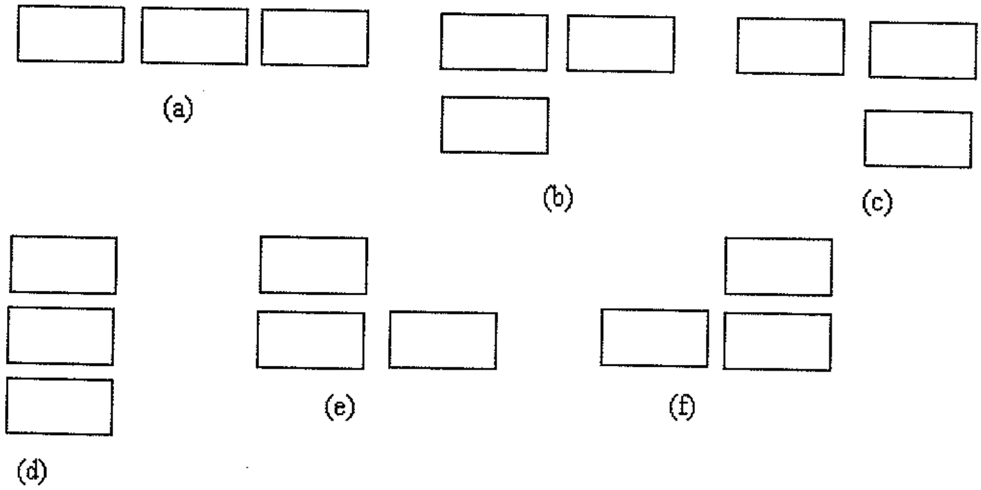


Figure 1: Different Combinations of Spatial Arrangement of Blocks in an Advertisement.

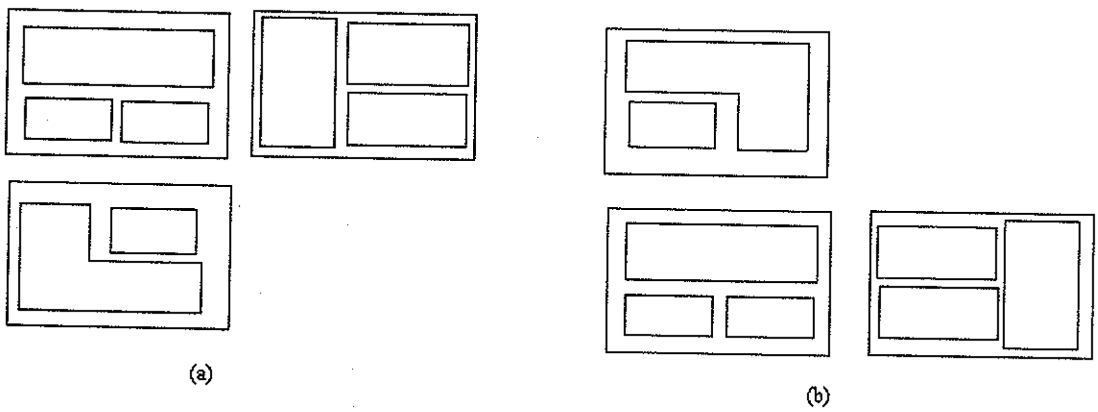


Figure 2: Different Possible Placements With In A Block

3. The Notion Of Entropy For OCR-less Operations On Document Images

A document image is the visual interpretation of newspaper, magazine or technical paper. Visually a typical document image is interplay of 0-1 and 1-0 transitions due to foreground and background. These transitions cause energy packets, which are quantified in terms of entropy [4,5,9,11]. A document image consists of different text components in a hierarchical sequence, namely paragraphs, lines and words. If the components at paragraph level are found dissimilar then there is no need for analyzing at line or word level in few cases where the structure of the paragraph components is kept constant, such as in the case of searching for a specific advertisement in a newspaper. If the components at paragraph level are found to appear similar then for further verification the components at line and word level would be analyzed. The details have appeared in our earlier papers [11]. To compare and establish the equivalence between two components at the same hierarchical level some of the basic steps required are.

1. Tight Segmentation [8]-to extract the components.
2. Extraction of geometrical features [9] to characterize the structure of the components.
3. Extraction of entropy values of the components to characterize the content of the components through.
   a. Conventional Entropy Quantifier (CEQ)[5,9,10,11].
   b. Spatial Entropy Quantifier (SEQ)[4,11].
4. Assimilation of entropy values of CEQ into frequency histograms and transformation of histograms into regression lines for comparison [4,10].
5. Assimilation of entropy values of SEQ into spatial histograms and transformation of histograms into regression lines for comparison [4].

The gist of each of the methods is summarized below, for the sake of completion and to build up to new algorithm presented subsequently.

3.1. Tight Segmentation [8]

Tight segmentation [7] is a top down approach, which segments the components at different hierarchical levels such as paragraph level, line level and word level and encloses them in a tight enclosure. Each component is identified based on the blank space existing between the components at same level. The threshold is automated by calculating the width of the blank space existing between the adjacent components.

In this paper we have considered two levels of extraction namely paragraph and sub paragraph level. Segmentation is applied on a document image with Manhattan layouts.

3.2. Position and Geometrical feature extraction [9]

Once the document image is subjected to tight segmentation, a component in the document is enclosed in the tight structure. From the X-Y plot of the contour enclosing the tight boundary, the position features- index point (x, y) of every component, its distance from the reference origin ($d=\sqrt{[x^2+y^2]}$) and its inclination ($\theta = \tan^{-1}(y/x)$) are extracted for coarse level components.

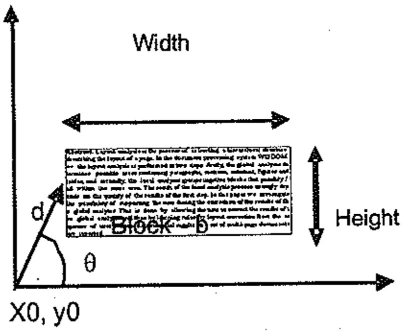


Figure 3: Feature Representation and Transition Points of a Component

In the sequel the GEOMETRICAL features of the component are extracted. These features describe the geometrical structure of every component. The suggested geometrical features are (i) Height (H), (ii) Width (W), (iii) Area (A=H * W), (iv) Density (D)of the high pixels in the component and (v) Sparsity (S) of the high pixels in the component is the density of null pixels in the component.

3.3. Conventional Entropy Quantifier (CEQ)[5,9,11]

Entropy is used to measure the probability(**p**), where **p** represents number of times 0-1 transitions occurs in each row and **1-p** represents the non-probable occurrence of transitions to measure entropy. Entropy is formulated as

$$E(t) = p\log(1/p) + (1-p)\log(1/(1-p)) \quad (1).$$

Considering the transitions only along the rows in an image matrix would yield in the same entropy value though the rows are shuffled in the image matrix. Also the entropy values remain same for rows with same number of transitions in them although they could occur in different positions in that row. Hence this method has been modified to improve the analysis of the component structure to take care of the short comings mentioned above.

For a component at any hierarchical level, entropy value is calculated for +ve and -ve transitions (0-1 and 1-0 transitions) in both horizontal (row) and vertical (column) directions. We have termed them based on the transitions taking place in row and column respectively as $E^+_h(r)$, $E^-_h(r)$ and $E^+_v(c)$, $E^-_v(c)$. Total entropy of each row and column is the summation of $E^+(r)$ and $E^-(r)$ and $E^+(c)$ and $E^-(c)$ as in equation- 2.

$$E(R)= E^+(r) + E^-(r) \ldots (2)$$

Total entropy of every component is given by

$$E(T)=E(R)+E(C)\ldots (3)$$

To overcome the limitation of CEQ, **spatial-entropy quantifier** has been introduced which calculates the entropy at every point where transition takes place.

3.4. Spatial Entropy Quantifiers(SEQ)[4]

The formulation of SEQ essentially is based on the equation-1 of CEQ, which is morphed to take an improved form as explained below.

The variables in CEQ expression are heuristically replaced to formulate an expression for SEQ. The position of transition is represented by ***pos***, which specifies the column number in row transition and row number in column transition. The ***pos*** parameter heuristically replaces ***p*** in CEQ and ***pos*** is always a number greater than zero, since ***pos*** represents the row number or column number of the image matrix. As the position number increases the entropy value also increases, this may be true in all rows and columns and there by yields same entropy value if the transition takes place at same point in different rows. To overcome this we have divided the position parameter ***pos*** by the corresponding row or column number where transition takes place but this induces the decrease in entropy value when ***pos*** value increases. The final entropy at every point is a negative value as the result of dividing the ***pos*** parameter with the row or column number yields in a value less than one and the logarithm of a number less than one is a negative number. Finally, the whole value is scaled by dividing the row/column number by the width or length of the component. This heuristically accounts for relative energy contributed by each row or column to the structure of the component. One may see equation-4 and 5, which represent the restructured equation-1 as explained in this paragraph.

Consider a component of size m x n, obtained from any hierarchical level

Let **R** represent the set of sequence of 'm' rows and **C** represents the set of sequence of 'n' columns.

R= {$\mathbf{r}_\alpha$ / every r_α is a horizontal row of consecutively placed **n** pixels one after the other} and $1 \leq \alpha \leq m$.

C= {$\mathbf{c}_\beta$ / every c_β is a vertical column of consecutively placed **m** pixels one below the other} and $1 \leq \beta \leq n$.

$$E(r_\alpha)=(r_\alpha \div m)(((pos \div n)\log(n \div pos))+((m-(pos \div n))\log(n \div ((m \times n)-pos))))\ldots(4)$$

If the transition occurs in a column c_β then it contributes to column entropy and it is given by

$$E(c_\beta)=(c_\beta \div n)(((pos \div m)\log(m \div pos))+((n-(pos \div m))\log(m \div ((m \times n)-pos))))\ldots(5)$$

If the transition occurs in a row r_α then it contributes to row Entropy and it is formulated as:

The total entropy of each component is the summation of all the entropy values at every point of transition both for +ve and –ve transitions as shown in equation-6, 7 and 8. Then the total horizontal entropy $E_h(R)$ is given by,

$$E_h(R)=\sum_n E(r_\alpha)\ \ldots(6)$$

Similarly for vertical entropy $E_v(C)$ is given by,

$$E_v(C)=\sum_m E(c_\beta)\ldots(7)$$

Total entropy of the component is given by,

$$E(T)=E_h(R)+E_v(C)\ \ldots(8)$$

It should be emphasized here that the entropies (CEQ, SEQ) introduced eliminate the process of *reading*.

the range as shown in fig-4.1.We have considered a 10 bin frequency histogram for the purpose of illustration.

$$H=\{E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8, E_9, E_{10}\}$$

Where E_i is the frequency count of the bin for a particular entropy range.

Subsequently cumulative frequency distribution is computed which is illustrated in fig-4.2.

$$CH=\{ce_1, ce_2, ce_3, ce_4, ce_5, ce_6, ce_7, ce_8, ce_9, ce_{10}\}$$

Where $ce_j = \Sigma E_k$ for k= 0 to j;

The cumulative histogram is normalized as shown in fig-4.3. A 1st order polynomial is fitted across these normalized cumulative histo points through a polyfit function [7] to

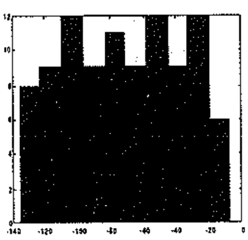


Figure 4.1: Histogram

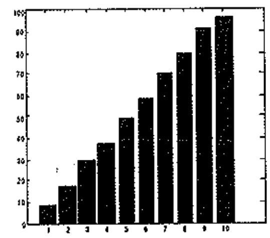


Figure 4.2 : Cumulative Histogram

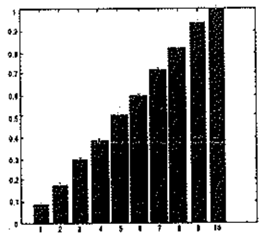


Figure 4.3: Normalized Cumulative Histogram

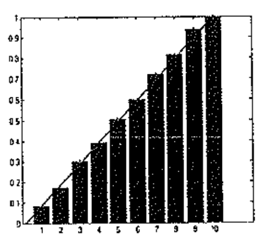


Figure 4.4: Fitting a Regression Line

3.5. Distance between regression lines obtained through transformation of Frequency histograms

The whole procedure is explained in detail in [3,6]. The following is the gist of the procedure as applied to entropy data.

Consider the entropy values of a set say $E^{-}(R)$, Let the total number of such transitions be N. Construct the histogram for the data values.

The x-axis represents the range of entropy values and y-axis represents the count of number of entropies with in

obtain a regression line through a ployval function[7] as illustrated in fig-4.4. The two regression lines of the corresponding components may intersect, may not intersect or they may overlap.

If the two regression lines overlap then the components is content wise identical, this also implies that the components are identical in terms of structure and position.

If the two regression lines do not intersect then the components are not equivalent in terms of content or they may not be similar in terms of structure also. But if the two regression lines do not intersect and if they are parallel to each other then the two components are structure wise similar but they are present at different positions.

3.6 Distance between regression lines obtained through transformation of spatial histograms[4]

Consider a row with 10 columns, let the number of transitions be 4 at positions 2, 4, 7 and 9. One such experimental observation is presented below.

X = [2, 4, 7, 9]

Y = [-124.32, -146.12, -178.98, -198.56];

Y' = [1. 0.85080, 0.6946, 0.626108];

Here X represents the position of transitions and Y represents the entropy at every position in X. Y' is the normalized entropy values.

The nature of entropy values extracted is in a sorted sequence, which can be observed in Y, the cause for this is explained in the earlier section. The advantage of this is that the histogram is normalized directly without constructing the cumulative histogram since the Y set is already in the sorted sequence. This is one of the major advantages of the entropy values extracted through spatial-entropy quantifier.

Y' is the normalized values of Y. Using the functions ployfit [7] and ployval [7] the first order polynomial is fitted through these points to obtain a regression line with y ranging from 0 to 1 as shown in figure 5.1, 5.2 and 5.3. For each row and column two regression lines are obtained since there are two sets of entropy values caused due to transitions 0-1 and 1-0. The number of histograms depends on number of rows and columns where transitions have occurred.

The spatial-entropy quantifier considers the position of occurrence of transition as one of the main feature, based on the position of the component the entropy also changes For each component; four regression lines are obtained corresponding to four entropy terms. The distance between the corresponding regression lines for any two comparable components is computed based on the behavior of the regression lines. The two regression lines of the corresponding components may intersect, may not intersect or they may overlap.

If the two regression lines overlap, then the components are content wise identical.

If the two regression lines do not intersect then the components are not equivalent in terms of content.

4. Proposed Algorithmic Model

From section-3 it is evident that looking at a document in terms of its energy composition measured in terms of entropy values eliminates the need to read the content of a document recognizing characters, words, lines and so on.

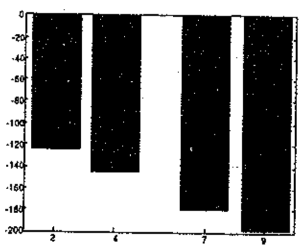


Figure 5.1: Spatial Hi stogram of a Single Row

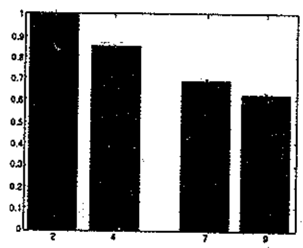


Figure 5.2: Normalized Spatial Histogram

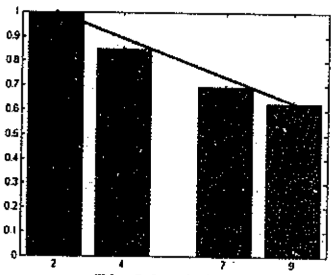


Fig 5.3: Fitting a Regression Line

In the Advertiser's problem, the inputs are

(i) Source document image containing many advertisements packed

(ii) The specific target (advertisement) image made up of one block or more than one block.

The source document image is segmented into blocks. The geometrical features of the blocks in the source document image are compared with the geometrical features of the target advertisement. If none of them match, every block of source document and target text advertisement are segmented into paragraphs. If the number of paragraphs within each block match then they are compared with geometrical features. If they are found comparable and show one to one correspondence then the CEQ values are extracted and represented in the form of frequency histogram and regression line based distance measure is applied.

(i) If the regression line of CEQ values overlaps, for further confirmation the SEQ values are extracted and assimilated through spatial histogram and the distance is computed through regression line based distance measure. If the regression lines overlap then they are identical paragraphs, if not they are further segmented into subparagraphs within each paragraph, if any, and the process is repeated to check for the equivalence.

(ii) If the regression lines are found dissimilar, then they are further segmented into sub paragraphs if any and then again CEQ values are extracted and process is repeated for finding the equivalence, if found similar then SEQ values are extracted and distance is computed constructing the spatial histograms using regression line based distance measure.

The algorithmic procedure is explained in detail step by step.

Procedure: Advertiser's problem

Input: Source Document Image, Target Advertisement.

Output: Position of the target text component in the source document image.

1. Extract the block components and the sequence of occurrence of both the target text component and reference document image.

2. Extract the geometrical span features of every block component extracted.

3. Compare every geometrical span feature by applying the absolute distance between the corresponding block components and the target text component.

a. If sequence of occurrence and geometrical features match, the target text component and the reference document image could have the same layout structure. Then they are subjected to further paragraph level component extraction.

b. If paragraphs matches in sequence then the paragraphs structure are **identical then the corresponding components are subjected to content analysis. go to step-5.**

c. If the paragraphs match in shuffled positions then the **first level paragraph components are subjected to content analysis correspondingly. go to step-5.**

4. If the paragraph components do not match, then they are further subjected to segmentation to extract sub paragraphs.

a. If the subparagraphs match in sequence then the corresponding subparagraph components are subjected to content analysis. go to step 5.

b. If the subparagraph match in shuffled positions then the layout of each paragraph is different and further the subparagraphs are subjected to content analysis. go to step-5.

5. Extract the entropy values of each component through CEQ.
6. Represent the entropy values into frequency histogram bin and transform the histograms into regression lines.
7. Compare the two regression lines of two corresponding components under test.
   a. If the lines overlap then the components are exactly identical in terms of structure and content.
   b. If they do not overlap, then the components are dissimilar.
8. If the components content appear to be identical the corresponding components are further subjected to SEQ values extraction.
9. Represent the entropy values of SEQ into spatial histogram [4] bin and transform the histograms into regression lines.
10. Compare the two regression lines of two corresponding components under comparison.
   a. If the lines overlap then they are identical.
   b. If they do not they are dissimilar components.

5. Experimental Analysis

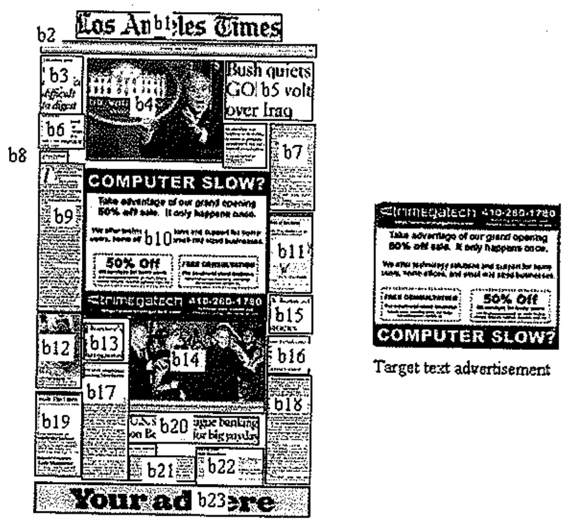
Experiments have been conducted on target advertisements with different structures and placed in different layouts in the source document. One of the experiments is illustrated in detail for better understanding of the proposed model and the results of other two are presented briefly.

Figure-6, shows a typical source document image and the target image for the purpose of spotting

Source document image

Target text advertisement

Figure 6 : (a) Source Document Image
(b) Target Advertisement

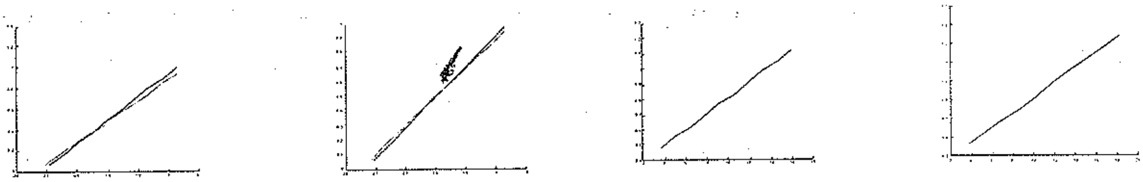


Source document image

Target text advertisement

Figure 7: Blocks Extracted In Source Document Image

In figure-7, blocks are extracted from the source document image through tight segmentation. The geometrical features of each block in source document image are compared with the target advertisement block. Block (b10) matches with the target advertisement. Further block(b10) is compared in terms of content through entropy feature CEQ as shown in figure-8.



8.1:Distance between $E^{+}(C)$

8.2: Distance between $E^{-}(C)$

8.3: Distance between $E^{+}(R)$

8.4: Distance between $E^{-}(R)$

Figure 8 : The Distance Between The Regression Lines For Entropy Terms

Since they do not match, further they are segmented into paragraphs as shown in figure-9.

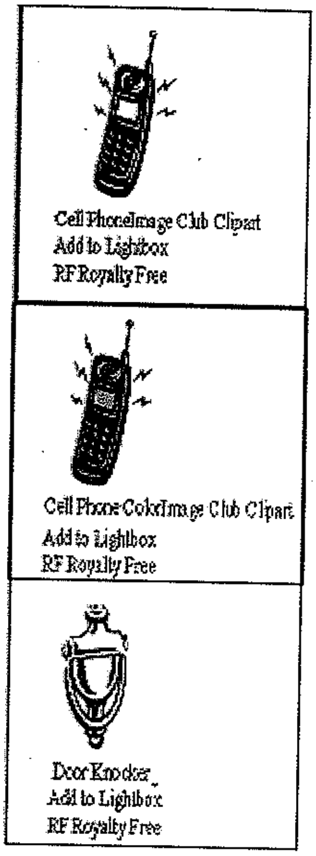
Figure 9 : The Paragraphs Extracted From The Block (B10) Of Source Document Image And Target Advertisement

In figure-9 block (b10) of source document image and target advertisement are segmented into paragraphs and they are compared with the geometrical features of paragraphs but they match with paragraphs at different positions, so the layout of the two blocks do not match. Paragraphs that have correspondingly matched are further subjected to extraction of entropy values through CEQ and distance is computed through regression line distance measure. As shown in figure-8 the regression lines of the corresponding paragraphs overlap and hence they prove that the components are similar. For confirmation further SEQ entropy values are extracted for each component both in source as well as for target advertisement and spatial histograms are constructed. Since spatial histogram is for every row and column of the component, it is more space consuming, hence we have not shown it, but the details of spatial histogram, with an example has been illustrated in our earlier paper [4].

The results of some more experiments are briefly presented below.

In figure-11, the target text advertisement consists of 3 blocks and they are present one below the other, but due to publisher's convenience, they are placed one beside the other. They do not match with each other at block level. when the block is further segmented into paragraphs, but the paragraphs match with the paragraphs in source document image but their spatial adjacency among them is different. therefore, the blocks have different layout format.

Source document image



Target text advertizement

Figure 10: The Target Text Advertisement Consists Of 3 Paragraphs and Their Spatial Adjacency Is Different In The Source Document Image

In the next experiment for the images shown in figure-11, the target advertisement is an article in the source document image. The layout of the target and the block containing the article in the source document image is different. At first level of analysis, none of the blocks match with the target advertisement, they are further subjected to segmentation, paragraphs are extracted within each block in source document image and the paragraphs are extracted in the target advertisement also. Further, the geometrical features of the paragraphs in the target advertisement are compared with the corresponding paragraphs in the source document image. In this experiment the paragraphs do not match in the sequence of extraction, they match with the paragraphs in shuffled positions and also the layout of the block in source document image is different when compared on the X-Y plot.


Filmdreh auf dem Rathausplatz

Ein Hauch von Hollywood-Flair vor dem Rathaus

Marianne Hettinger dreht Szenen für ihren Kino-Film

Source Document Image


Ein Hauch von Hollywood-Flair vor dem Rathaus

Marianne Hettinger dreht Szenen für ihren Kino-Film

Target Advertisement

Figure 11: Spatial Arrangement Is Different In Target Advertisement And Source Document Image

6. Conclusion

Searching for a particular target in a vast document image is a tedious job when one tries to find out manually or when one tries to automate with a language processor (OCR). To over come this, a new method has been proposed in the framework of Advertiser's problem, where the job is to search, locate and establish the equivalence of a target Advertisement in a source document image. This works in the way similar to human performance when he tries to spot a target in a newspaper which he cannot read. Although the method is basically designed to search for the text target components, it can be extended to the targets containing pictures or images also.

References

[1] A.Balasubramanian, Million Meshesha, and C.V. Jawahar, *"Retrieval from Document Image Collections"*, H. Bunke and A.L. Spitz (Eds.): DAS (2006), LNCS 3872.

[2] Li.Zhang, Chew Lim Tan, *"Word image coding technique and its applications in information retrieval from image documents"*, IWDA, Kolkata(2005).

[3] Nagabhushan.P, R. Pradeep Kumar, *"Histogram PCA*, ISNN (2) (2007).

[4] P.Nagabhushan, Sahana.D.Gowda, Chew Lim Tan, *"OCR-Less Procedure for Equivalence between two Document Images through Spatio-Entropy Quantifier"*, international journal of computing and mathematical applications, Vol 2. No. 12, January –December 2008.

[5] P.Nagabhushan, Sahana D.Gowda, *"Establishing Equivalence between Two Layout-Variant Document Images without Reading"*, International conference on Modeling and Simulation, Kolkata, India 2007.

[6] Pradeep Kumar.R, Nagabhushan.P, *"An Approach Based on Regression Line Features for Low Complexity Content Based Image Retrieval"*, ICCTA, 2007.

[7] Rafael C. Gonzalez, Richard E.Woods and Steven L.Eddins, *"Digital Image Processing using MATLAB"*, Pearson Education, 2006.

[8] Sahana.D.Gowda and Nagabhushan.P, *"Page Segmentation and Rule Based Identification Method for Technical Journal Pages"*, International conference on Cognition and recognition, India, 2005.

[9] Sahana D. Gowda and Nagabhushan.P, *"Equivalence between Two Document Images In Terms Of Geometry and Entropy"*, ACVIT, Aurangabad, INDIA, 2007.

[10] Sahana.D.Gowda, P.Nagabhushan, *"Equivalence between two document images through iterative histo analysis of the conventional entropy values"*, accepted for international journal of system, cybernetics and informatics, pentagram, Hyderabad.

[11] Sahana.D.Gowda, P.Nagabhushan, *"Entropy Quantifiers useful for establishing equivalence between text document images"*, International conference on computational intelligence and multimedia applications, Sivakasi, INDIA2007.

[12] Santanu Chaudhury,Geetika Sethi ,Anand Vyas and Gaurav Harit, *"Devising Interactive Access Techniques for Indian Language Document Images"*, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)

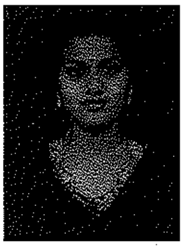
[13] Shin.C and D.Doermann, *"Classification of document page images"*, in workshop on document layout interpretation and its applications (DLIA 99) at the ICDAR'99, September, 1999

[14] Véronique EGLIN, Stéphane BRES *LIRIS – RFV, INSA de Lyon,*Document page similarity based on layout visual saliency:Application to query by example and document classification, Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)

[15] Weihua Huang, Chew Lim Tan, Sam Yuan Sung and Yi Xu, *"Vertical Bar Detection for Gauging Text Similarity of Document Images"*, proceedings of 6th International conference on document analysis and recognition, (2001).

Author's Biography

Prof.P.Nagabhushan(BE-1980, M.Tech-1983, Ph.D-1988-89) has been serving as a Professor in the Department of Studies in Computer Science since 1994. He is an active researcher in the area of Cognition and Recognition. He has extensively worked on feature extraction, reduction, classification, cluster analysis, data and knowledge mining, digital image analysis, recognition and so on. Till now 17 candidates have completed Ph.D successfully with him and presently 6 more are working. He has more than 100 publications and proceedings and totally has about 450 papers to his credit. He is a fellow of ***Institution of Engineers*** and an active member in many professional bodies.



Prof. Sahana D.Gowda (BE-1999, ME-2004, Ph.D-2009) has been serving as a Professor in the Department of Computer science, B.N.M. Institute of technology, Banaglore since 2009. She is actively involved in research activities. Her area of interest is Image processing, Pattern recognition and Computer Architecture. She has 6 conference papers and 6 journals to her credit.