Find a Text in Image File Using Correlation Method

[1]Ziad M. Abood, [2]Intisar Abd Yousif, [3]Ahmed Kawther Hussein
Al-Mustnsriayh University- College of Education

Abstract: *Image correlation is representative of a wide variety of window-based image processing tasks. We can be search for text/ word in documents with extension as DOC, PDF, TXT ... etc., and count the number of it in document or in the page. the difficulty in this process when we are keeping this documents in the form of images and any extension such as(jpg, bmp, ... etc.), thus lead to difficulty in the search for a word in those images, as some peoples depend on convert the document to image file in order to keep document from manipulation or making copies of them.*

This study introduces method by using V.B. and depends on correlation method in search for "word" or "symbol" in image files.

Index Terms- *image correlation, image processing, V.B, search.*

I. Introduction

Image correlation is a widely used procedure in many areas of image and picture processing. This process, also known as template matching, is used to locate an object in a picture [1, 2] or, in image registration, to match pieces of two pictures to one another [3]. It is used in some forms of edge detection to find the step edge between two areas, or to find lines, spots, or curves [2]. In digital photogrammetric, image correlation is used to find the corresponding points of two images of a stereo model [2]. Because of image correlation requires comparing portions of two images in a large number of relative positions, it is an extremely time consuming process. In this application, image sizes are typically at 794*1132 (i.e. A4) or any size at less to find:

- Number of text (word) in an image.
- Any equation or symbol in an image.
- Any part of shape, curve, picture in an image.

II. Image Digital Definition and Correlation

An image is represented by a two-dimensional array where each element ("*pixel*") has an unsigned integer value representing the "gray level" of the pixel. Image correlation involves determining the position at which a relatively small match area best matches a portion of an input image. Correlation measures are used to measure the degree of similarity or disagreement between the match area and an equivalent size area on the input image. Let the symbols x and y denote to single elements of arrays X and Y, where X is the match image and Y is an area of the input image which has the same dimensions as X. [4]

Let M be the number of elements in the match area X. Two representative correlation measures are:

$$\mathrm{SXY} = \frac{\sum \mathrm{xy} - \sum \mathrm{x} - \sum \mathrm{y}}{\mathrm{M}} \qquad \ldots (1)$$

$$\mathrm{PXY} = \frac{\mathrm{SXY}}{(\mathrm{SXX} - \mathrm{SYY})^{1/2}}$$

Correlation measure SXY is the covariance of the match area with a portion of the input area. Large positive values indicate similarity, while large negative values indicate similarity between a positive and a negative image. Values near zero indicate little or no similarity. Correlation measure RXY is the linear correlation coefficient of statistics. This measure is a normalized version of SXY, with values ranging between +1 and -1.

The value of +1 indicates exact similarity, while values near zero indicate little similarity. In general, a correlation value will be computed for every possible position where the match area will fit on the input image. The match position where the correlation measure is maximized corresponds to the best placement of the match area on the image.

The computation time for image correlation is dominated by the time to compute the Σxy, Σy, and (for measure RXY) the Σy^2 values for all possible match positions. The Σx and Σx^2 values involve only the match area elements, and need to be computed (or precomputed) only once. The way in which data elements are combined to obtain the Σxy values is similar to operations performed in a variety of important image processing tasks, including convolution and filtering. For an input image having R rows and C columns and a match area having r rows and c columns, there are (R - r + 1) (C - c + 1) match positions. Serial computation of the Σxy terms over the entire image, performed by simply sliding the match area over the image and calculating the value of Σxy for each overlap position, requires (R - r + 1) (R - r + 1) rc multiplications and (C - c + 1)(C - c + l)(rc - 1) additions, see figure (1) [5].

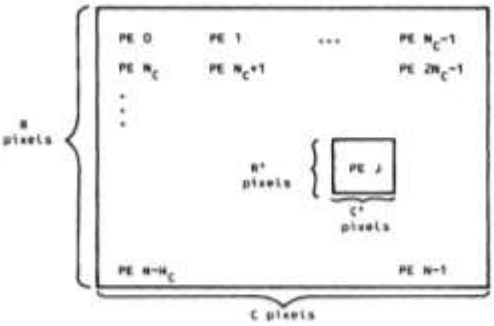


Figure (1): Data assignment of R×C image [5]

In computing the Σxy values, each match position generates a new set of terms to be summed. No terms from one match position can be reused in a different match position. In computing the Σy and Σy^2 values, two (or more) input image elements summed for one match position may also be summed for another match position. The algorithms considered for calculating the Σy and Σy^2 values therefore attempt to avoid "redundant" operations, e.g., performing a sum for one match position which has already been performed for another. The operations performed in computing the Σy and Σy^2 values, i.e., the summing of elements under a window where the window moves over an image, are typical of operations required for a variety of image processing tasks. These include image smoothing, edge enhancement, and convolution using a rectangular window.

Consider the following serial (uniprocessor) algorithm for computing the Σy's, i.e., summing the pixel values in each match area. This algorithm will be used as a basis for parallel algorithms.

If the Σxy, Σy, and Σy^2 values for a given match position are computed together, the correlation measure for that match position can be calculated, and is saved only if it is the current maximum over the correlation measure values computed so far. Thus, the Σxy, Σy, and Σy^2 values for each position do not have to be saved. [5]

III. Cross correlation

Cross correlation is a standard method of estimating the degree to which two series are correlated. Consider two series x(i) and y(i) where i=0,1,2...N-1. The cross correlation r at delay d is defined as: [6, 7, 8]

$$r = \frac{\sum_i [\,(x(i) - mx) * (y(i-d) - my)\,]}{\sqrt{\sum_i (x(i) - mx)^2}\sqrt{\sum_i (y(i-d) - my)^2}}$$

Where mx and my are the means of the corresponding series. If the above is computed for all delays d=0, 1, 2, ... N-1 then it results in a cross correlation series of twice the length as the original series.

$$r(d) = \frac{\sum_i [\,(x(i) - mx) * (y(i-d) - my)\,]}{\sqrt{\sum_i (x(i) - mx)^2}\sqrt{\sum_i (y(i-d) - my)^2}}$$

There is the issue of what to do when the index into the series is less than 0 or greater than or equal to the number of points. (i-d < 0 or i-d >= N) The most common approaches are to either ignore these points or assuming the series x and y are zero for i < 0 and i >= N. In many signal processing applications the series is assumed to be circular in which case the out of range indexes are "wrapped" back within range, i.e. x(-1) = x(N-1), x(N+5) = x(5) etc.

The range of delays d and thus the length of the cross correlation series can be less than N, for example the aim may be to test correlation at short delays only. The denominator in the expression above serves to normalize the correlation coefficients such that -1 <= r(d) <= 1, the bounds indicating maximum correlation and 0 indicating no correlation. A high negative correlation indicates a high correlation but of the inverse of one of the series.

The maximum correlation is achieved at a delay of 3. Considering the equations above, what is happening is the second series is being slid past the first, at each shift the sum of the product of the newly lined up terms in the series is computed. This sum will be large when the shift (delay) is such that similar structure lines up. This is essentially the same as the so called convolution except for the normalization terms in the denominator.

IV. 2D Pattern Identification using Cross Correlation

It is one of the approaches used to identifying a pattern within an image uses cross correlation of the image with a suitable mask. Where the mask and the pattern being sought are similar the cross correlation will be high. The mask is itself an image that needs to have the same functional appearance as the pattern to be found.

Consider the image below in black and the red mask shown figure (2). The mask is centered at every pixel in the image and the cross correlation calculated, this forms a 2D array of correlation coefficients.

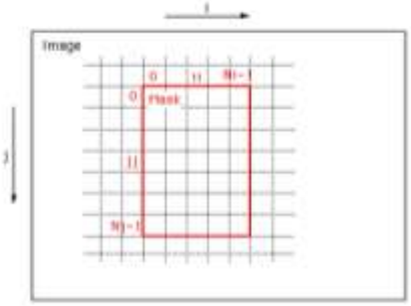


Figure (2): The mask is centered at every pixel in the image and the cross correlation calculated

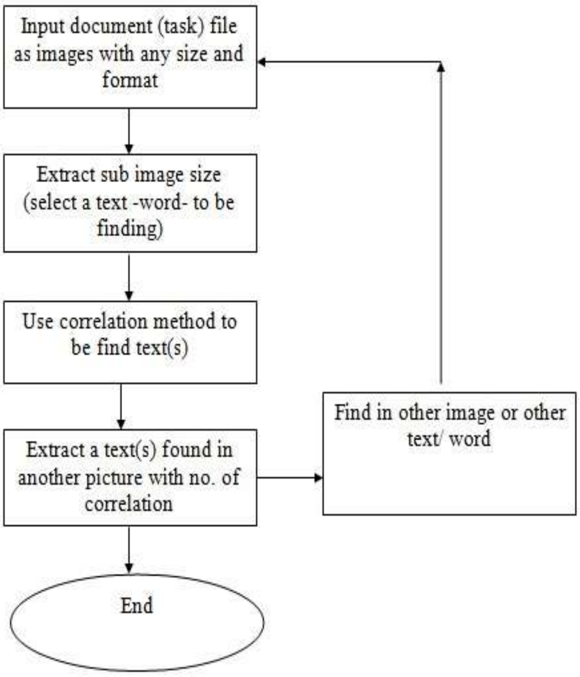
The form of the un-normalized correlation coefficient at position (i,j) on the image is given by: [6]

$$r[i][j] = \sum_{jj=-Nj/2}^{jj<Nj/2} \sum_{ii=-Ni/2}^{ii<Ni/2} (mask[ii+Ni/2][jj+Nj/2] - \overline{mask})\,(image[i+ii][j+jj] - \overline{image})$$

Where $\overline{\text{mask}}$ is the mean of the masks pixels and $\overline{\text{image}}$ is the mean of the image pixels covered by the mask.

V. Algorithm of present work:

Figure (3) shows the Algorithm of present work.


Input document (task) file as images with any size and format
Extract sub image size (select a text -word- to be finding)
Use correlation method to be find text(s)
Extract a text(s) found in another picture with no. of correlation
Find in other image or other text/ word
End

VI. Interface of program and results:

The figure (4) shows the Interface of program (V.B Language), with the examples of many samples to find (search) of text, equation and symbol, in document images.

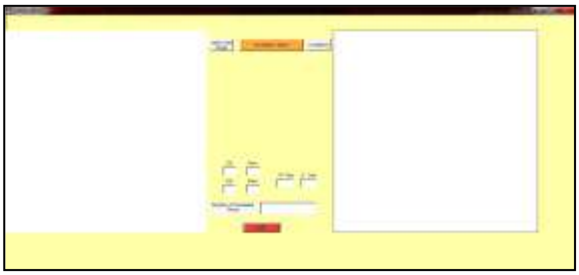
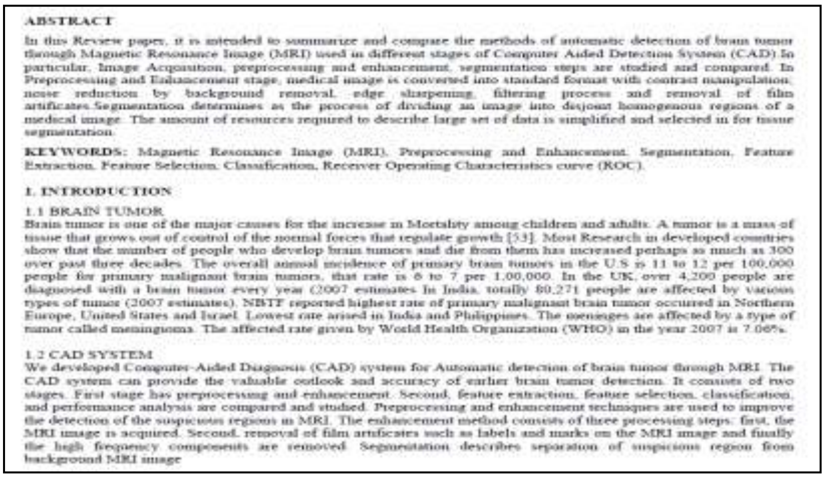


Figure (4): The Interface of program in (V.B Language)

- Sample (1): find the text "MRI" in flowing image:



ABSTRACT

In this Review paper, it is intended to summarize and compare the methods of automatic detection of brain tumor through Magnetic Resonance Image (MRI) used in different stages of Computer Aided Detection System (CAD).In particular, Image Acquisition, preprocessing and enhancement, segmentation steps are studied and compared. In Preprocessing and Enhancement stage, medical image is converted into standard format with contrast manipulation, noise reduction by background removal, edge sharpening, filtering process and removal of film artificates.Segmentation determines as the process of dividing an image into disjoint homogenous regions of a medical image. The amount of resources required to describe large set of data is simplified and selected in for tissue segmentation.

KEYWORDS: Magnetic Resonance Image (MRI), Preprocessing and Enhancement, Segmentation, Feature Extraction, Feature Selection, Classification, Receiver Operating Characteristics curve (ROC).

1. INTRODUCTION

1.1 BRAIN TUMOR

Brain tumor is one of the major causes for the increase in Mortality among children and adults. A tumor is a mass of tissue that grows out of control of the normal forces that regulate growth [53]. Most Research in developed countries show that the number of people who develop brain tumors and die from them has increased perhaps as much as 300 over past three decades. The overall annual incidence of primary brain tumors in the U.S is 11 to 12 per 100,000 people for primary malignant brain tumors, that rate is 6 to 7 per 1,00,000. In the UK, over 4,200 people are diagnosed with a brain tumor every year (2007 estimates In India, totally 80,271 people are affected by various types of tumor (2007 estimates). NBTF reported highest rate of primary malignant brain tumor occurred in Northern Europe, United States and Israel. Lowest rate arised in India and Philippines. The meninges are affected by a type of tumor called meningioma. The affected rate given by World Health Organization (WHO) in the year 2007 is 7.06%.

1.2 CAD SYSTEM

We developed Computer-Aided Diagnosis (CAD) system for Automatic detection of brain tumor through MRI. The CAD system can provide the valuable outlook and accuracy of earlier brain tumor detection. It consists of two stages. First stage has preprocessing and enhancement. Second, feature extraction, feature selection, classification, and performance analysis are compared and studied. Preprocessing and enhancement techniques are used to improve the detection of the suspicious regions in MRI. The enhancement method consists of three processing steps: first, the MRI image is acquired. Second, removal of film artificates such as labels and marks on the MRI image and finally the high frequency components are removed. Segmentation describes separation of suspicious region from background MRI image.

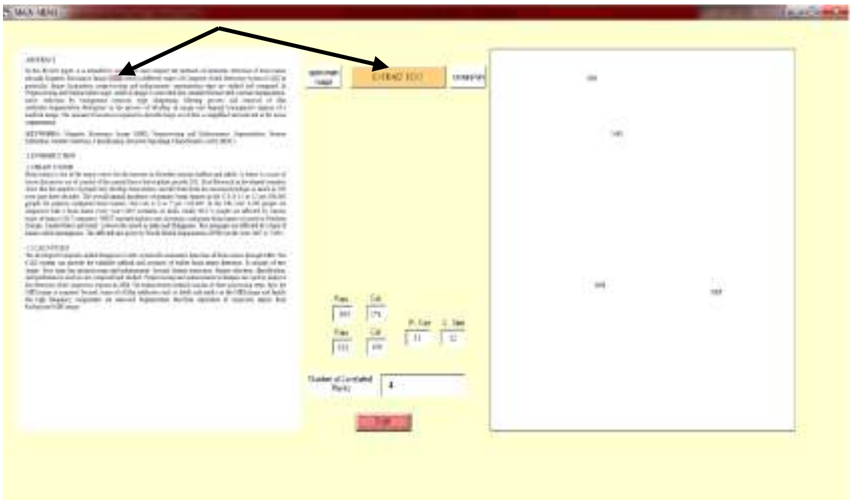
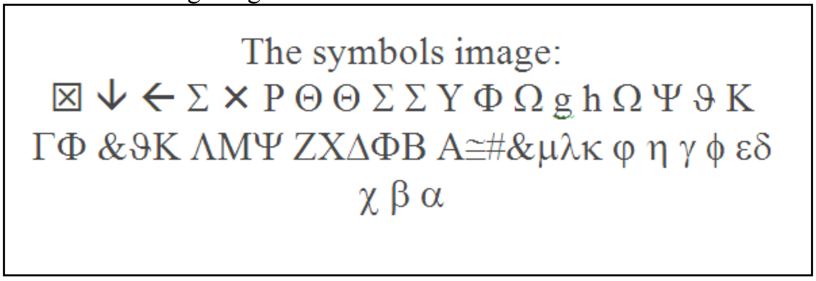


Figure (5): The result of applying correlation program on sample (1)

- Sample (2): Find the text "Ω" in flowing image



The symbols image:

⊠ ↓ ← Σ × Ρ Θ Θ Σ Σ Υ Φ Ω g h Ω Ψ ϑ Κ ΓΦ &ϑΚ ΛΜΨ ΖΧΔΦΒ Α≅#&μλκ φ η γ ϕ εδ χ β α

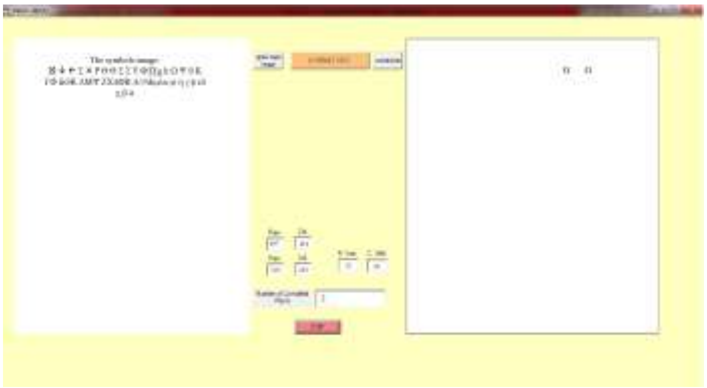


Figure (6): The result of applying correlation program on sample (2)

VII. Conclusions

1. The current method is easy and more effective to find and count words (texts) to be searched for in the image files.
2. It is Possible to be applied on “symbols" and "equations", because the programs such as MS-word dose not have the possibility to search for code within a given equation.
3. This method can be applied to search for any part of the picture.
4. It could be applied to any language: Arabic, Chinese, Japanese … etc.
5. From defects in the current way, it is searching for a word in other words; does not differentiate between large and small letters or words in color because this method looking for the shape and the color of the word as it is exactly.

References:

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[2] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*. New York: Academic, 1976.

[3] W. K. Pratt, *Correlation techniques of image registration*, IEEE Trans. Aerosp. Electron. Syst., vol. AES-10, pp. 353-358, May 1974.

[4] H. C. Andrews, *Digital Image Processing*, New York: IEEE, 1978.

[5] Leah J. Siegel, Howard Jay Siegel, And Arthur E. Feather, Parallel Processing Approaches To Image Correlation, IEEE, Transactions On Computers, vol. C-31, No. 3, March 1982.

[6] PaulBourke, *CrossCorrelation,Autocorrelation- 2DPatternIdentification*,August 1996.

[7] David Jacobs, Correlation and Convolution, Class Notes for CMSC 426, fall 2005.

[8] Soo-Chang Pei, Jian-Jiun Ding, Jahan Chang, *Color Pattern Recognition by Quaternion Correlation*, Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, IEEE.