International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

Localization and Detection of Multi-lingual Text in Scenery Using an Ensemble of Gabor, FIR and Gaussian Filters

Suryadipto Sarkar
Electronics and Communication Engineering, Manipal Institute of Technology, Manipal, Karnataka, India

ABSTRACT

Texts have become an integral part of urban as well as rustic landscapes. Starting from directions, street names, milestones, public notices and traffic signboards to hoardings and advertisements—we can find text everywhere. Text detection in natural scenery, more popularly known as "Scene Text Detection", has been an extensive field of study in the domain of Image Processing. Yet, owing to the varying alignment, quality, clarity, distance and other parameters of non-uniformity that are present in such types of text, researchers have not been able to completely saturate this aspect of Image Analysis. As more work is being done, we get a better understanding in this regard. But, it is still far from complete resolved. In this paper, the author has proposed a technique for scene text detection using time-frequency analysis.

Keywords: Short-Time Fourier Transform, Gabor transform, Finite Impulse Response(FIR) filter, Gaussian filter, MSRA-TD500 database.

I. INTRODUCTION

Scene text recognition, like any other computer vision problem, aims at representing amplitude information from digital images in a more comprehensive, informative manner. Such representations enhance subsequent stages to processing. Oftentimes, the images are of such bad quality or are so vastly non-uniform that it is a challenge to spatially modify all the images using a particular algorithm.

Also, spatial analysis depends directly on the intensity information that the image possesses. There are innumerable techniques of spatial analysis, each suitable in serving a particular aspect of image processing and resolution enhancement. Though we can good results in certain cases with the application of contrasting, thresholding, histogram equalisation and various other spatial analysis techniques, that is not always the case. For example, in an image that has been blurred beyond repair or lost complete clarity due to external noise, there is no way we can retrieve it back by intensity analysis. But that does not mean that the image is devoid of information. It is full of useful information, just that they are not detectable by the human eye.

For this purpose, we make use of frequency analysis. One of the innumerable applications of frequency analysis comes in the form of edge detection. It is a known fact that high frequency components include the edges in the image. Therefore, we can just deploy a high-pass filter to extract image edges. Or a low-pass filter to get the plain areas. Depending on our requirements, we can make use of the different types of filters and control their responses as well as bandwidths(frequencies).

But that is not to say that we can completely ignore the intensity information that was present in the original image. We need a method of analysing images taking into account both its time and frequency components for best results during pre-processing and minimal information loss. The way

we achieve this is by making use of what is known as time-frequency analysis of images. The most primitive type of time-frequency analysis is the Short-Time Fourier Transform(STFT), but Wavelet transforms are more effective, sophisticated and in higher demand.

The main benefit of using Wavelet transforms is that they provide greater temporal resolution.

In case of STFT, the image is continually multiplied by a window function of uniform dimensions which is non-zero for only a short instance of time. The Fourier transform of the resultant signal is calculated, while the window slides along the time axis. Because the window dimensions remain unchanged for every single computation, STFT gives uniform frequency resolution.

But that is not how wavelet transformation works. It makes use of a dynamic basis function rather than a constant one like in STFT. Wavelet transformation works based on the principle of uncertainty which states that it is not possible to measure both the frequency and time information of a signal at the same given instant of time. Therefore, basis functions are chosen in such a manner that they can take a decision based on the trade-off between time and frequency information at any particular point in time. That is why, wavelet transforms allow changes only in time extension but not shape, of the basis function. Changes in time extension would therefore automatically result in a change in the frequency of analysis of the basis function. This conforms with the uncertainty principle which is given by the following equation:

$$\Delta t \Delta w \geq \frac{1}{2},$$

where t represents time and w represents angular frequency

$$\Rightarrow \Delta t \times 2\pi\Delta f \geq \frac{1}{2},$$

where f represents temporal frequency of basis function

$$\Rightarrow \Delta t \Delta f \geq K,$$

where K is a positive, real constant

II. GABOR FILTER

Gabor filters are band-pass, orientation-sensitive filters that work exceptionally well on unidimensional signals. A Gabor filter which has been designed to orient in a certain direction provides a strong response for locations of the image which have structures aligned in that particular direction. Based on the lines, edges and gratings on the image, the orientation of the Gabor filter is often set up to extract optimal response.

Gabor filters are used in edge detection, feature extraction, texture segmentation, target detection, and stereo disparity estimation. It has acted as a medium for achieving unsupervised classification in many researches. From text detection, handwriting recognition, medical image analysis to image processing applications for the army, medicine and other domains—Gabor has proved to be one of the strongest time-frequency analysis techniques.

By definition, a Gabor filter is obtained on multiplication of a Gaussian kernel function and a complex sinusoid. In other words, it is nothing but a complex sinusoidal carrier modulated by a Gaussian envelope. This results in a Gabor wavelet given by the equation,

$$g(x,y) = s(x,y)wr(x,y)$$

where, $s(x,y)$ is the complex sinusoidal carrier
and, $w_r(x,y)$ is the Gaussian modulation function

The carrier is of the form:

$$s(x,y) = e^{j(2\pi(u0x+v0y)+\oplus)}$$

where, u_0 is the frequency of the horizontal sinusoid
v_0 is the frequency of the vertical sinusoid
and $\oplus$ is the phase shift

The real part of the carrier signal is given by,

$$Re(s(x,y)) = \cos(2\pi(u_0x + v_0y) + \oplus)$$

The imaginary part of the carrier signal is given by,

$$Im(s(x,y)) = \sin(2\pi(u_0x + v_0y) + \oplus)$$

The Gaussian envelope is defined by the following equation,

$$w_r(\mathrm{x},\mathrm{y}) = \frac{1}{2\pi\sigma_x\sigma_y}\mathrm{e}^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x{}^2}+\frac{y^2}{\sigma_y{}^2}\right)}$$
$$where, \sigma_x \; and \; \sigma_y \; are \; constants$$

On multiplication of the complex sinusoidal carrier signal with the Gaussian modulator, the resultant signal is a Gabor wavelet defined by the following equation,

$$\mathrm{g}(\mathrm{x},\mathrm{y}) = \mathrm{K}\mathrm{e}^{-\pi(\mathrm{a}^2(\mathrm{x}-\mathrm{x}_0)^2{}_\mathrm{r}+\mathrm{b}^2(\mathrm{y}-\mathrm{y}_0)^2{}_\mathrm{r})}$$
$$where, (x-x_0)_r = (x-x_0)\cos(\theta) + (\mathrm{y}-y_0)\sin(\theta)$$
$$and, (y-y_0)_r = -(x-x_0)\sin(\theta) + (\mathrm{y}-y_0)\cos(\theta)$$

III. Finite Impulse Response(FIR) filter

The output of an FIR filter is calculated as the weighted sum of a finite number of terms, of past, present, and sometimes future values of the input. Therefore, it has impulse response of finite period. It is different from Infinite Impulse Response(IIR) filters in that IIRs often have internal feedbacks from past or future outputs, and they respond infinitely. FIRs, on the other hand, have their response damping down and finally becoming zero at a finite interval. Unlike IIRs, their responses do not carry on indefinitely. Hence, FIRs are more effective low-pass, band-pass and high-pass filters as compared to IIRs.

Mathematically, it is defines by the following equation:

$$y[n] = \sum_{k=-M_1}^{M_2} b_k x[n-k]$$
$$where, M_1 \; and \; M_2 \; are \; finite \; values$$

An example of a simple FIR filter is a 3-term moving median filter:

$$y[n] = \frac{1}{3}(x[n-1] + x[n] + x[n+1])$$

FIR filters are feed-forward filters, meaning there is no feedback of future or past outputs in current output computation. Only calculations(generally simple addition or subtraction) on inputs(which are often weighted) are taken into consideration while computing the present output of the filter.

IV. GAUSSIAN FILTER

A Gaussian filter is a filter which has its impulse response as a Gaussian curve. It is a low-pass filter used in image processing for the purpose of smoothening, noise reduction and computing of derivatives of the image.

The Gaussian filter is a convolution-based filter, where the Gaussian matrix serves as the kernel. It is a linear filter, which means it replaces every pixel by a linear combination of its neighbouring values(i.e., with weights specified by the Gaussian kernel matrix). It is also a localised filter, which means that it produces values of output pixels based on the neighbouring pixel values as has been determined by the Gaussian convolution kernel.

Gaussian filters are capable of producing exact simulations of optical blur. Therefore, Gaussian filters are used for any blurring application on images. An integral property of Gaussian filters is that they always possess non-negative values. Hence, convolution with a Gaussian filter always produces a non-negative result. Hence the Gaussian function serves in forming a one-to-one mapping of values that are non-negative. And the result is always a valid image.

It is used to blur images and remove detail and noise. Thus, it is somewhat similar to the mean filter, but it uses the Gaussian(bell-shaped) kernel. Since it is used for the purpose of blurring and noise reduction, if we use two of them and subtract then they can be used for very effective edge detection. But one filter alone will just reduce contrast and blur the image.

The advantage of Gaussian filter over median filter is that it is fast to compute(especially on huge images), because multiplication and addition take lesser computational time than sorting. Another reason why Gaussian filters are especially preferred in image processing is that, the Fourier transform of a Gaussian function is also a Gaussian curve. Hence, they are

very close in resemblance in both their time-domain and frequency-domain values.

Some properties of Gaussian functions are given as follows:

Impulse response of $1-D$ *Gaussian filter*:

$$g(x) = \sqrt{\frac{a}{\Pi}} e^{-ax^2}$$

$$=> g(x) = \frac{1}{\sqrt{2\Pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

$$where, \sigma \text{ stands for standard deviation}$$

Frequency response is given by its Fourier transform

$$G(f) = e^{-\frac{\Pi^2 f^2}{a^2}}$$

$$=> G(f) = e^{-\frac{f^2}{2\sigma_f{}^2}}$$

$$where, \sigma.\sigma_f = \frac{1}{2\Pi}$$

2D Gaussian filter is nothing but a combination of tw one $-$ *dimensional filters*(*one in each direction*):

$$g(x,y) = \frac{1}{2\Pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

V. Maximally Stable Extremal Regions(MSER)

Maximally Stable Extremal Regions(MSER) is a technique for blob detection, also known as region detection, from images. The MSER technique computes a set of co-variant regions called MSERs from an image. It takes into consideration areas which stay uniform through a wide variation of threshold values. All pixels with intensity values below the threshold are white whereas those above the threshold are black. MSER is used in wide-baseline stereo matching and image recognition applications. In our application, we have used it for the purpose of character detection in text.

VI. PROPOSED METHODOLOGY

In this paper, the author has deployed a combination of different types of filters in order for better time-frequency analysis of text in natural images. Figure 1 shows the workflow followed for the purpose of this research.

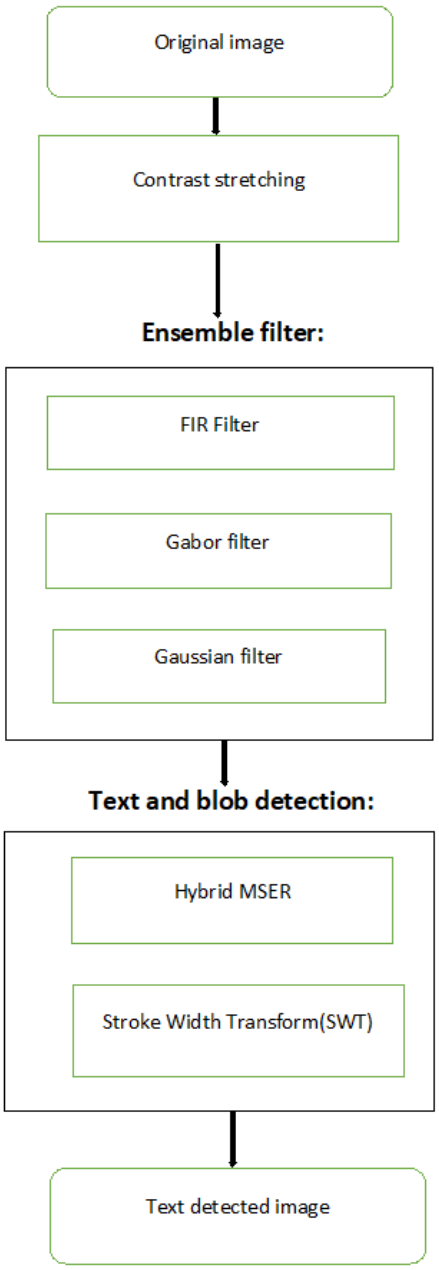

Fig. 1 Flowchart depicting the proposed technique

VII. RESULTS AND SIMULATIONS

For the purpose of this experiment, MATLAB R2018a has been used as the main platform for simulation and image analysis. All processing has been done on a 64-bit Windows 10 machine.

Figure 2 given below shows the entire process of text detection for a particular image from the MSRA Text Detection 500(MSRA-TD500) public database. Figure 3 depicts a comparison between the text and non-text clusters. Figure 4 shows the difference between region image and stroke width image. Finally, figure 5 shows the particular image before and after text detection.

Fig. 2(a) Original image

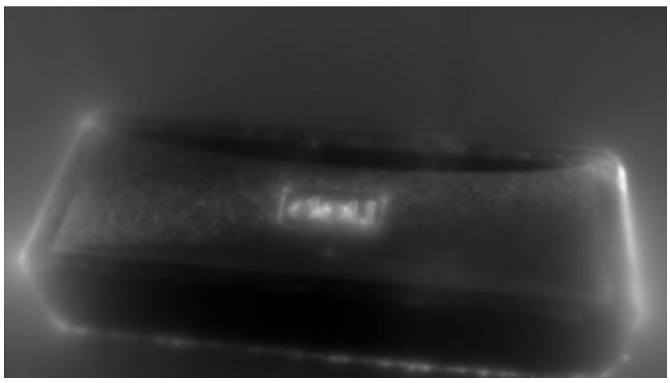

Fig. 2(b) Image after application of ensemble filter

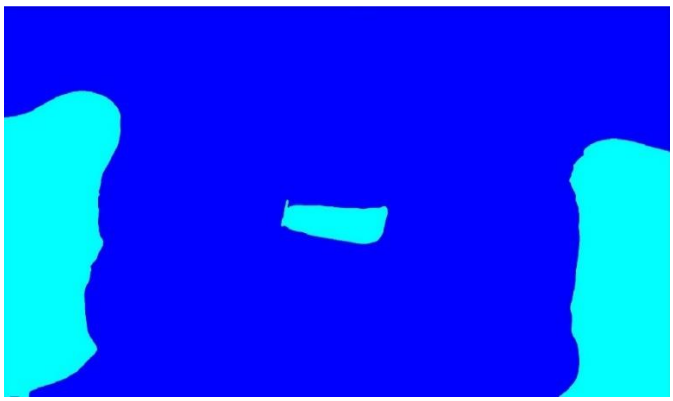

Fig. 2(c) Image after clustering into text and non-text regions

Fig. 2(d) Image of text cluster

Figure 2(e) Multiple expanding bounding boxes on application of Stroke Width Transform

Fig. 2(f) Final text detected image

Fig. 2 Process of text detection on an image from the MSRA-TD500 database

Fig. 3 Comparison between the text and non-text clusters

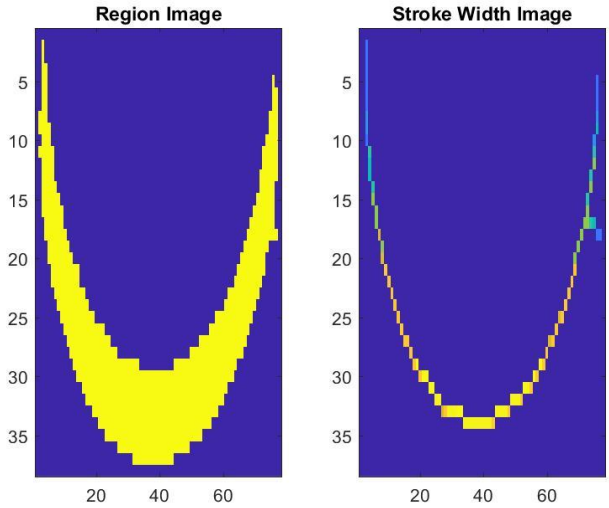

Fig. 4 Region image and stroke width image

Fig. 5 Original image and text detected image

VIII. CONCLUSION AND FUTURE SCOPE

The proposed method is capable of detecting texts that are too small to even be recognised by the naked eye. Also, they can detect areas with text irrespective of the language of text. The method deployed in this research helps to draw a perfect balance between time as well as frequency variations in images. This technique can be used in related domains like number-plate and facial emotion recognition as well, owing to its minimal pre-processing and complete dependence on generic, time-frequency analysis.

IX. REFERENCES

[1] Asif Shahab, Faisal Shafait and Andreas Dengel, "ICDAR 2011 Robust Reading Competition Challenge2: Reading Text in Scene Images," in 2011 International Conference on Document Analysis and Recognition, DOI 10.1109/ICDAR.2011.296

[2] Adam Coates, Blake Carpenter, Carl Case, Sanjeev Satheesh, Bipin Suresh, Tao Wang, David J. Wu and Andrew Y. Ng "Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning,"

in 2011 International Conference on Document Analysis and Recognition, DOI 10.1109/ICDAR.2011.95

[3] SeongHun Lee, Min Su Cho, Kyomin Jung and Jin Hyung Kim, "Scene Text Extraction with Edge Constraintand Text Collinearity," in 2010 International Conference on Pattern Recognition, DOI 10.1109/ICPR.2010.969

[4] Yi-Feng Pan, Cheng-Lin Liu and Xinwen Hou, "Fast Scene Text Localization by Learning-Based Filtering And Verification," in Proceedings of 2010 IEEE 17th International Conference on Image Processing, September 26-29, 2010, Hong Kong

[5] Jerod J. Weinman and Erik Learned-Miller, and Allen Hanson, "Fast Lexicon-Based Scene Text Recognition with Sparse Belief Propagation."

[6] Joo-Hwee Lim, Yiqun Li, Yilun You and Jean-Pierre Chevallet, "Scene Recognition with Camera Phones for Tourist Information Access."

[7] Shangxuan Tian, Shijian Lu, Bolan Su and Chew Lim Tan, "Scene Text Recognition using Co-occurrence of Histogram of Oriented Gradients," in 2013 12th International Conference on Document Analysis and Recognition, DOI 10.1109/ICDAR.2013.186

[8] Josef Sivic and Andrew Zisserman, "Efficient Visual Search of Videos Cast as Text Retrieval," in IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 31, NO. 4, APRIL 2009

[9] Yu Zhong, Hongjiang Zhang and Anil K. Jain, "Automatic Caption Localization in Compressed Video," in IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 4, April 2000

[10] Xilin Chen, Jie Yang, Jing Zhang and Alex Waibel, "Automatic Detection and Recognition of Signs From Natural Scenes," IEEE Transactions on Image Processing, Vol. 13, No. 1, January 2004, DOI 10.1109/TIP.2003.819223

[11] Yingying Zhu, Cong Yao and Xiang Bai, "Scene text detection and recognition: recent advances and future trends," DOI 10.1007/s11704-015-4488-0

[12] Huiping Li, David Doermann and Omid Kia, "Automatic Text Detection and Tracking in Digital Video," in IEEE Transactions on Image Processing, Vol. 9, No. 1, January 2000

[13] Tomoyuki Saoi, Hideaki Goto and Hiroaki Kobayashi "Text Detection in Color Scene Images based on Unsupervised Clustering of Multi-channel Wavelet Features," in Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)