Printed and Handwritten Mixed Kannada Characters Recognition using Template Matching Method

[1]Dr. H.S. Mohana, [2]Pradeepa R, [3]Rajithkumar BK, [4]Dr. M Shivakumar
[1]Dept. of Instrumentation Technology, Malnad College of Engineering, Hassan, Karnataka, INDIA
[2]Dept. of ECE, Malnad College of Engineering, Hassan, Karnataka, INDIA
[3]Dept. of Electronics and Communication Engineering, RVCE, Bangalore, Karnataka, INDIA
[4]Dept. of Electronics and Communication Engineering, GSSSIT, Mysore, Karnataka, INDIA
Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India

Abstract
Rapid growth of technology and prevalent use of computer in the business and other areas, more and more organization are converting their paper document into electronic documents that can be processed by computer. Kannada, the native language of south western state in India has more than 60 million in Karnataka and outside. Much research has been done on the identification of English, Arabic, Japanese and Chinese. It is observed that the research on recognition of Kannada characters is still an open research problem. In this paper, an attempt has been made to develop a simple and efficient method for the Recognition of Kannada printed and handwritten characters. Here in the present work, the images of Kannada printed and handwritten characters are captured from an ordinary mobile camera of 5Mega pixel resolutions and images are pre-processed to remove noise. Here it recognizes the Kannada characters based on template matching, and a character is identified by analyzing its shape and comparing its features that distinguish each character. Cross correlation technique is implemented in matching the characters. Experimental results shows, it demonstrates relatively high accuracy in recognizing both printed and handwritten characters and with better time efficiency when compared to previous methods.

Keywords
Template Matching, Cross Correlation, 2-D Correlation Coefficient

I. Introduction

A character can be written in a number of ways differing in shape and properties. Recognition of any hand written character with respect to any language is difficult [2]. Kannada the native language of south western state in India has several speakers in Karnataka. It is estimated that there are between 50 and 200 million books ever published. A significant fraction of these books are only available in non-digital forms. Furthermore, most of the estimated 100,000 new books published yearly are published exclusively on paper for many reasons including copyright [6]. This prohibits the automatic indexing and searching of these documents and consequently limits their availability and accessibility on digital medium like the Internet, PDA and mobile phones. Optical Character Recognition (OCR) technology has been used for decades to convert scanned images of documents to editable text [7].

Here in the present work, the image is processed such that its character is recognized. The major problem which arises while identifying the characters in printed or handwritten is the difference in the style in literature [5]. Template matching, or matrix matching, is one of the most common classification methods. In template matching, individual image pixels are used as features [1]. Classification is performed by comparing an input character image with a set of templates from each character class. Each comparison results in a similarity measure between the input character and the template. One measure increases the amount of similarity when a pixel in the observed character is identical to the same pixel in the template image. If the pixels differ the measure of similarity may be decreased. After all templates have been compared with the observed character image, the character's identity is assigned as the identity of the most similar template. Structural classification methods utilize structural features and decision rules to classify characters. Structural features may be defined in terms of character strokes, character holes, or other character attributes such as concavities. For a character image input, the structural features are extracted and a rule-based system is applied to classify the character. Template matching for character recognition is straightforward and reliable. This method is more tolerant to noise than structural analysis method.

II. Related Works

The review of the literature pertaining to the present topic is presented to the readers. In [1] authors concentrate on Template Matching method for Recognition Musnad characters based on correlation analysis. In this paper, it extended that work and applied that algorithm for recognize printed and handwritten Kannada characters. In [3] authors concentrate on the Era Identification and Recognition of Stone In-scripted Kannada Characters Using Artificial Neural Networks. In this paper it use same Gaussian filter for filtering, the Gaussian filter smoothing the image and it helps find edges of characters accurately. In [5] authors concentrate Printed Number Recognition using MATLAB. In this paper it uses same thinning and cropping procedure for to extract desired characters shape. In [2] authors concentrate on Extraction of Kannada characters using SIFT. In this paper it extended that work and applied that algorithm.

III. Proposed Algorithm

The proposed method consists of following steps see in fig 1

A. Create a Template of Kannada characters and each image in a template is in size of 24x42 dimensions.

B. Capture Different style, size Kannada Printed and Handwritten images using ordinary mobile camera of 5Mega resolution

C. Pre-processing

In this it perform

1. Removing noise content in an captured images using Gaussian filter
2. Resizing of all pre-processed images into fixed pixel size and dimension
3. Cropping of characters based on user wish
4. Finding Edge in an image using Sobel edge detection
5. Perform dilation

6. Use top-hat filtering to correct uneven illumination
7. Remove all objects in the image containing less than 30 pixels
8. Reconstruction of image by reconstructing its boundary and
9. Filling its holes
10. Thinning of characters and take complement of image for clear visibility

D. Characters Cropping

1. This is a user block; here user can crop any characters in an image for Recognition

E. Extraction of each Kannada characters in an image

Steps involved this is

1. The pre-processed image is smoothened to reduce the number of connected components
2. Then calculating the connected components
3. Then by reading the connected components it get the required Kannada characters
4. In this it divide Kannada characters into lines and it divide input Kannada characters into first line and remain line

F. Cross correlation

1. In this method it perform Cross correlation between Template and Classified captured images

G. Recognition of Kannada Printed and Handwritten Characters

1. Based on Cross correlation Analysis value it recognize Captured Printed and handwritten Kannada characters

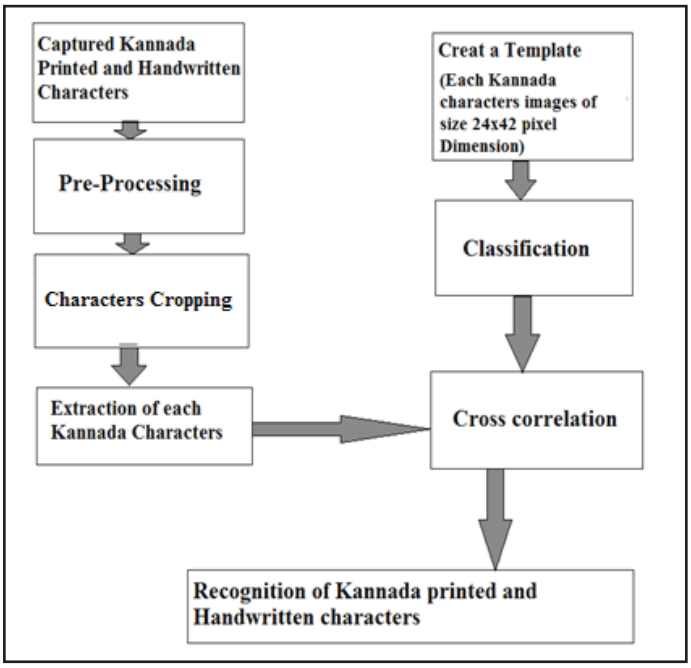

Fig. 1: A Block Diagram of the Proposed Algorithm

IV. Methodology and Implementation

A. Pre-processing

In this it first removing the noise using Gaussian filter

1. Gaussian Filtering

Gaussian filtering is used to blur images and remove noise and detail.

In one dimension, the Gaussian function is

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

Where σ is the standard deviation of the distribution. It have also assumed that the distribution has a mean of zero (i.e. it is centered on the line x=0). The distribution is illustrated in fig. 2

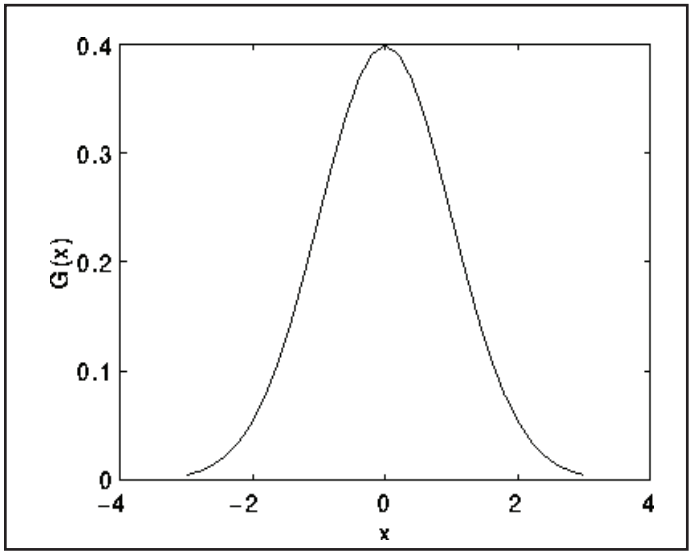

Fig. 2: 1-D Gaussian Distribution With Mean 0 and σ = 1

2. Edge Detection

Here it Find edges of objects in images using Sobel. The Sobel operator is used in image processing, particularly within edge detection algorithms. Technically, it is a discrete differentiation operator, computing an approximation of the gradient of the image intensity function. At each point in the image, the result of the Sobel operator is either the corresponding gradient vector or the norm of this vector.

3. Dilation

The dilation operator takes two pieces of data as inputs. The first is the image which is to be dilated. The second is a (usually small) set of coordinate points known as a structuring element (also known as a kernel). It is this structuring element that determines the precise effect of the dilation on the input image. Note that in this and subsequent diagrams, foreground pixels are represented by 1's and background pixels by 0's.

4. Removing of Small Object

The dilated image contains some small object, so it remove all small object using BWAREAOPEN operation, this operation will remove all small pixel object and it remove small pixel object based on user need and here in our work it removing all small object whose size less than 80 pixel

5. Reconstruction of Image

The dilated image contain some breaking border so In this it reconstructs the character by eliminating its breaking border, for reconstruction it use IMRECONSTRUCT for reconstruct border and IMFILL operation for fill holes in images

6. Thinning

Thinning process removes selected parts of foreground pixels of a binary image. It is somewhat like eroding the selected image to leave just the skeleton of it. It can be used to process images of

handwritten or printed characters. This process helps us extract the basicshape of the image.

B. Extraction of Kannada characters in an image

Steps involved

1. The edge detected image is smoothened to reduce the number of connected components
2. Then calculating the connected components
3. Then by reading the connected components it get the required objects

Connected Components

A connected component in a binary image is a set of pixels that form a connected group. For example, the binary image below in fig. 3 has three connected components.

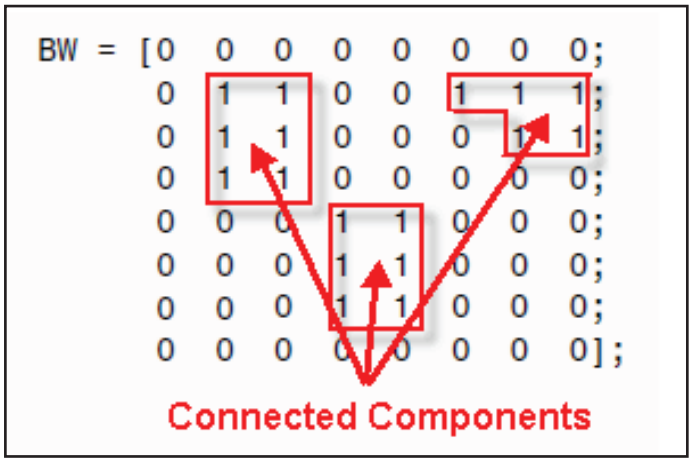

Fig. 3: Calculating Connected Components

Connected component labeling is the process of identifying the connected components in an image and assigning each one a unique label as shown in fig. 4:

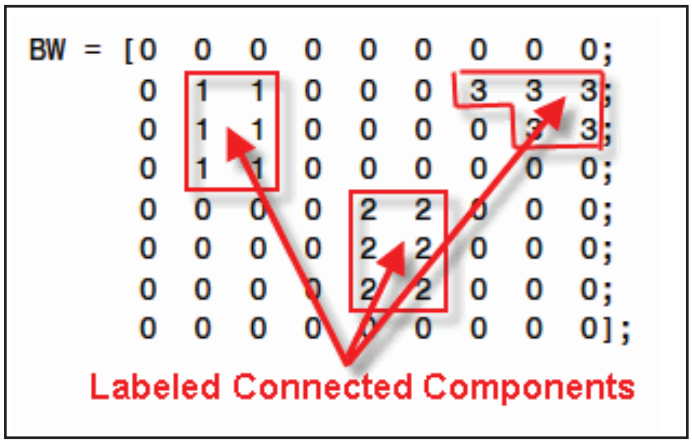

Fig. 4: Label Matrix

By using this connected components value and label value it extract each character in an image.

C. Template Matching Method

Template matching is one of the Character Recognition techniques. It is the process of finding the location of a sub image called a template inside an image. Once a number of corresponding templates is found, their centers are used as corresponding points to determine the registration parameters. Template matching involves determining similarities between a given template and windows of the same size in an image and identifying the window that produces the highest similarity measure. It works by comparing derived image features of the image and the template for each possible displacement of the template.

This process involves the use of a database of characters or templates. There exists a template for all possible input characters. For recognition to occur, the current input character is compared to each template to find either an exact match, or the template with the closest representation of the input character. If I(x, y) is the input character, TN(x, y) is the template n, then the matching function s (I, TN) will return a value indicating how it template n matches the input character. Some of the more common matching functions are based on the following

Formulas

$$S(I.Tn) = \sum_{I=0}^{W} \sum_{J=0}^{H} |(i,j) - Tn(i,j)| \quad (2)$$

$$S(I.Tn) = \sum_{I=0}^{W} \sum_{J=0}^{H} |(i,j) - Tn(i,j)|\,2 \quad (3)$$

$$S(I.Tn) = \sum_{I=0}^{W} \sum_{J=0}^{H} |(i,j)Tn(i,j)| \quad (4)$$

$$S(I.Tn) = \frac{\sum_{I=0}^{W} \sum_{J=0}^{H} |(i,j) - |I|(Tn(i,j) - Tn||}{\sqrt{\sum_{I=0}^{W} \sum_{J=0}^{H} |((i,j) - |I|)2((Tn(i,j) - Tn)2|}} \quad (5)$$

Matching approaches: (2) City block, (3) Euclidean distance, (4) Cross Correlation, (5) 2-D Normalized Correlation.

1. Implementation of Kannada Character Recognition

The implementation of Kannada character recognition is done by firstly refining the extracted characters to fit them into a window without white spaces on all the four sides and creating the template for each extracted character. The templates are normalized to 42x24 pixelsand stored in the database. Normalization is done using window to view port transformation. This mapping is used to map every pixel of the original image to the corresponding pixel in the normalized image. The extracted character of the input test image, after normalization, is matched with all the characters in the database using 2-D normalized correlation coefficients approach to identify similar patterns between a test image and the standard database images. This approach is shown in Equation (4).

V. Results

Step 1: Create a **Template of Kannada characters** and each image in a template is in **size of 24x42** dimensions as shown in fig. 5.

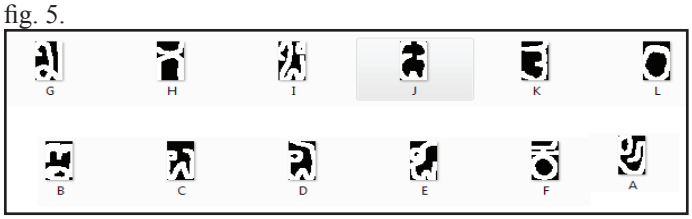

Fig. 5: Kannada Characters Template Images

Step 2: Capture Different style, size Kannada Printed and Handwritten images using ordinary mobile camera of 5Mega Pixels resolution, as shown in fig. 6

Fig. 6: Captured Kannada Handwritten and Printed Characters Found in News Paper or a Book

Step 3: Pre-preprocessing
In this step it extract the each Kannada Characters and steps involved are shown in fig. 7.

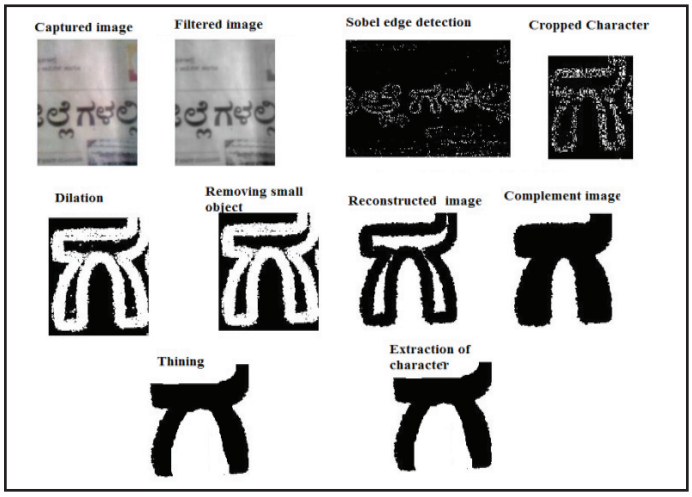

Fig. 7: Image Processing and Extraction of Each Kannada Character

In this it removing noise content in an captured images using Gaussian filter and it Resizing of all pre-processed images into fixed pixel size and dimension and it Finding Edge in an image using Sobel edge detection and then it Perform dilation and Use top-hat filtering to correct uneven illumination. It remove all objects in the image containing less than 30 pixels and it reconstruction of image by reconstructing its boundary and filling its holes, finally by calculating its connected components it extract the character in an image. This process is similar to handwritten Kannada characters also.

Step 4: Cross correlation
In this it perform cross correlation betiten pre-processed image with template. Based on their result value it recognize the characters as shown in fig. 8, 9, 10.

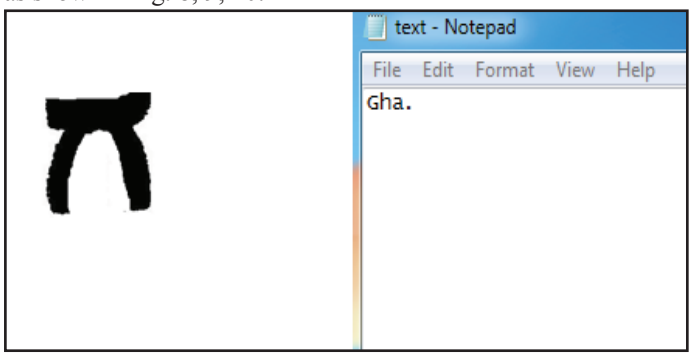

Fig. 8: Recgonization of Printed Kannada Character

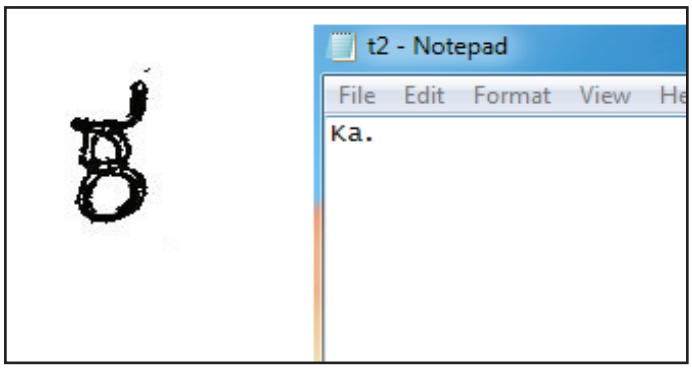

Fig. 9: Recgonization of Handwritten Kannada Character Ka.

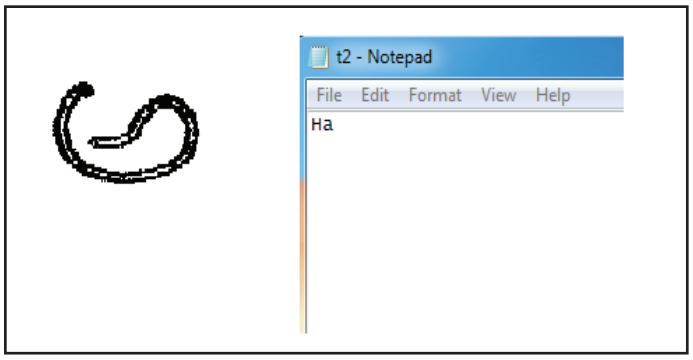

Fig. 10: Recgonization of Handwritten Kannada Character Ha.

VI. Experimental Results

Experiments have been performed to test the proposed method. MATLAB (R2009a) is the software tool that was used for Recognition of Kannada Characters. The Experiments are performed on many printed and handwritten test images of Different size, style. Table 1, 2 gives the results of Recognition rate between the character images and their Templates images. The Recognition rate can be obtained by formula

$$\sigma = \frac{\alpha - \gamma}{\delta} \qquad (6)$$

Where σ = Recognition rate, α = sum of correct match, γ = sum of incorrect match, δ = Number of test samples, Here it are considering 2% error for handwritten character and 3% error for printed character.

Table 1: The Recognition Rate Analysis of Kannada Handwritten Characters

Test image	Recognized Character	Sum of correct Match (α)	Sum of incorrect Match (γ)	Number of test Samples (δ)	Error (2%)	Recognition rate (σ)
	Ka	10	1	10	0.2	88%
	Ha	10	0	10	0.2	98%
	Ta	10	0	10	0.2	98%
	Gha	10	1	10	0.2	88%
	Tha	10	1	10	0.2	88%

Table 2: The Recognition Rate Analysis of Kannada Printed Characters

Test image	Recognized Character	Sum of correct Match (α)	Sum of incorrect Match (γ)	Number of test samples (δ)	Error (3%)	Recognition rate (σ)
ಕ	Ka	10	0	10	0.3	97%
ಘ	Gha	10	1	10	0.3	87%
ಥ	Tha	10	0	10	0.3	97%
ರ	Ra	10	0	10	0.3	97%

From Table 1 it is shown that our proposal recognition rate should be 92% accuracy for Kannada handwritten characters and from Table 2 it is shown that our recognition rate should be 95% accuracy for Printed characters

VII. Conclusion

A simple and effective template matching method for identification of Kannada printed and handwritten characters method introduced in this paper. Here it are used ordinary mobile camera of 5Mega pixel resolution to capture characters so it is very easy to implement and it is costless when compare to other methods. For recognition process, the extracted character was compared to each template in the database to find the closest representation of the input character. The matching metric was computed using 2-D correlation coefficients approach to identify similar patterns bitten the test image and the database images. Experimental results show that the proposed method is efficient for identification Kannada printed and handwritten characters.

References

[1] Mohammed Ali Qatran,"Template matching method for recognition Musnadcharacters based on correlation analysis", Department of Computer Science, Amran University, Yemen.

[2] Dr. H.S. Mohana, et al.,"Extraction of Stone In-scripted Kannada Characters Using Sift Algorithm Based Image Mosaic", International Journal of Electronics & Communication Technology, IJECT Vol. 5, Issue 2, April - June 2014.

[3] Dr. H.S. Mohana, et al.,"Era Identification and Recognition of Stone In-scripted Kannada Characters Using Artificial Neural Networks", 2nd National Conference on Innovation in Computing and Communication Technology, March 22, 2014.

[4] Maalinee Ramu,"Printed Number Recognition using Matlab", University Teknologi, Malaysia.

[5] G. Rajput, Rajeswari Horakeri, Sidramappa Chandrakant, "Printed and Handwritten Mixed Kannada Numerals Recognition Using SVM", Department of Computer Science, Gulbarga University, Gulbarga, Karnataka-India.

[6] Ahmad Abdulkader Matthew R. Casey,"Low Cost Correction of OCR Errors Using Learning in a Multi-Engine Environment", IEEE10.1109/ICDAR.2009.242

[7] H. S. Mohana,"Machine Vision Based Non-Magnetic Object Detection and Removal on Moving Conveyors in Steel Industry through Differential Techniques", International Journal ofComputer Vision and Image Processing (IJCVIP) 2 (2012): 3, accessed (March 14, 2014).

[8] Mori, S., Suen, C. Y., Yanamoto, K.,"Historical review of OCR research and development", Proc. of the IEEE, Vol.80, pp. 1029- 1058, 1992.

[9] Izakian, H., Monadjemi, S. A., Ladani, B. T., Zamanifar, K.,"Multi-Font Farsi/Arabic Isolated Character Recognition Using Chain Codes", World Academy of Science, Engineering and Technology, 43, 2008.

[10] Rafel C Gonzalez, Richard E.woods,"Digital Image processing (2nd edition).

[11] Tappert, C., Suen C., Wakahara, T.,"The stateof the art in online handwriting recognition", IEEETrans. On Pattern Anal. And Machine Intel. Vol. 12, Issue 8, pp. 787 - 808, 1990.

[12] Serge Belongie, Jitendra Malik,"Shape Matching and Object Recognition Using Shape Contexts", IEEE transactions on patternanalysis and machine intelligence, Vol. 24, No. 24, April 2002.

[13] Dipti Deodhare, et al.,"Preprocessing and image enhancement algorithms for a Form-based Intelligent Character Recognition System", IJCSA 2(2), pp. 131-144, 2005.

[14] [Online]Available: http://www.en.wikipedia.org/wiki/Gaussian_blur

[15] [Online]Available: http://www.mathwork.com