Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments
Assistive Technology for All Ages
CVHI 2007, M.A. Hersh (ed.)

TEXTURE BASED TEXT DETECTION IN NATURAL SCENE IMAGES: A HELP TO BLIND AND VISUALLY IMPAIRED PERSONS

Shehzad Muhammad Hanif, Lionel Prevost

Institut des Systèmes Intelligents et Robotique CNRS – FRE 2507
Université Pierre et Marie Curie, Paris 06,
BC 252, 4 Place Jussieu, 75252 Paris CEDEX 05, France
Tel: +33 1 44 27 96 73. Fax: +33 1 44 27 44 38.
Email: shehzad.muhammad@lisif.jussieu.fr, lionel.prevost@upmc.fr

Abstract: In this paper, we propose a texture based technique to detect text in grey level natural scene images. This work is a part of the project called Intelligent Glasses. It is a wearable system to facilitate navigation and to assist the blind and visually impaired persons in real world. It has three parts, a bank of stereovision, a processing unit for visual perception and a handheld tactile surface. In its original form, it will be able to provide information about different types of obstacles and their position with respect to user. Our textual/symbolic information interpretation module to the vision system of the Intelligent Glasses will recognize the text from the captured scene and textual and/or symbolic information will be displayed on the handheld tactile. Initial results are encouraging with a text detection rate of 64%.

Keywords: Intelligent Glasses, text detection, image analysis, handheld tactile

1. Introduction

In this article, we have proposed a texture based technique to detect text in grey level natural scene images. This work is a part of the project called “Intelligent Glasses” (Velazquez et al., 2003). The aim of the project is to help blind and visually impaired persons to know their environment in a better way. The Intelligent Glasses is a man-machine interface which translates visual data (such as a 3D global information) onto its tactile representation as shown in figure 1. It has three parts, a bank of stereovision, a processing unit for visual perception and a handheld tactile of Braille surface type. The visual data is acquired and processed by a vision system, while its tactile representation is displayed on a touch stimulating surface.

(a) (b) (c) (d)

Figure 1 Intelligent Glasses : (a) concept (b) scene (c) environment perception (d) representation

One of the salient feature of this tactile surface of the system is its compact representation of the environment and space. Using the visual (acquired by stereovision system) and inertial information, it

changes with the time and with the scene observation point. With the help of these two dimensions, this interface can represent a simplified 2D version of 3D world. It is also possible to estimate the relations between real world and its tactile representation such as the size of an object, an obstacle distance with respect to user etc. Moreover, it can represent different geometrical shapes (square, rectangle, circle, arcs, curves….) making it feasible to represent contours, edges and arcs etc which are generally present in a scene.

However, all these information are not sufficient to give enough autonomy to a blind/visually impaired person. As an important form of human beings language, visual texts are widely used in our daily life. For example, different sign boards, directions, shop names etc contain textual and/or symbolic information that is perceived by a human being to facilitate his knowledge of environment and perhaps also help in his navigation. The need to interpret textual and/or symbolic information becomes evident in the case of blind or visually impaired persons. With this perspective, we want to add textual/symbolic information interpretation module to the vision system of the Intelligent Glasses that will recognize the text characters from the captured scene text and finally, textual and/or symbolic information will be displayed on the handheld tactile. However, before recognition, we have to know about the text existence (text detection) and text location in an image (text localisation). As the tactile surface is reprogrammble, it is possible to adapt it for the representation of textual and symbolic information persent in the scene.

This paper addresses the problem of scene text detection in grey-level images and presents our preliminary work in this domain. Our algorithm generates potential/candidate text regions that can later be verified by a validation/verification scheme. We follow a general framework of candidate regions generation and validation as employed by various researchers. Our detection method is based on six texture features computed from grey-level co-occurrence matrix (GLCM), proposed by Haralick (Haralick et al., 1973). These features capture texture information present in the image. Based on the fact that text is a distinct texture, we can distinguish between text and non-text regions. Our classification scheme classifies image pixels as text or non-text. Probabilistic model and neural network based classifiers have been employed for the classification task.

The rest of the paper is organized as follows. In section 2, we persent some of the previous work in this domain. We describe the proposed texture coding and classification technique in section 3. Six features namely contrast, homogeneity, dissimilarity, entropy, energy and correlation from grey-level co-occurrence matrix (GLCM) are computed over a small region of image. Hence capturing local textural details present in the image, we can build a system which can differentiate text and non-text regions. Image database, experimental setup and results are included in section 4. Finally section 5 concludes this paper.

2. Previous works

2.1 Existing systems for blind and visually impaired persons incorporating text detection

Most of the existing systems use voice synthesizer to help a blind or visually impaired person. The system proposed by Ezaki et al. (Ezaki et al., 2004) consists of a PDA, a CCD camera and a voice synthesizer. The system is destined to work in two scenarios: walk around mode and sitting mode. In ”walk around mode”, the camera placed on the shoulder captures images and algorithm on PDA searches text in the image. Once found, the text is zoomed in and high resolution characters recognized and read out to user via sound synthesizer. The other mode is used for reading restaurant menu or book covers. The Visual Integration and Dissementation of Information (VIDI, Silapachote et al., 2005) system is also a wearable device with a head mounted camera attached to a mobile computer allowing a visually impaired person to receive information about the presence of signs present in the environment. The system is destined to deal with a variety of signs such as traffic, government, public and commercial signs. In an other work (Ferreria et al., 2003), the authors proposed a system to read out text from a wide range of printed documents like newspaper, books, restaurant menus etc. The system has three modules: image acquisition module to capture images, a text detection/recognition module and a text to speech converter. The algorithms are run on a PDA or smartphone.

2.2 Text detection methods

In recent years, digital cameras/ camcoders and PDA are increasingly popular and they have shown potential as alternative imaging devices. The researchers working in document analysis and recognition have changed their orientation and instead of working with traditional scanner captured document images, they are concentrating on analysis of images taken from camera. This change accompanies with lot more challenges such as low resolution, uneven lightining, perspective distortion, nonplanar surfaces, wide angle lens distoration, complex background (Liang et al., 2005).

Existing methods for text detection, localisation and extraction can broadly be classified as gradient features based, color segmentation based and texture features based (Liang et al., 2005). Here, we will concentrate on texture methods. Text is viewed as a unique texture that exhibits a certain regularity that is distinguishable from background. Humans can identify text of foreign languages even when they do not understand them largely due to its distinct texture. Various researchers have exploited this fact to detect text in images. The texture methods are largely used for text detection. Texture features can be extracted directly from the pixel's spatial relationships or from frequency data. However, these methods are often computationally expensive and are greatly dependant on contrast of the text in an image, but lead to good results.

3. Proposed Method

3.1 Texture Coding Scheme

We have proposed a simple texture coding method to detect scene text in grey-level natural scene images. We have used spatial histograms computed from grey-level co-occurrence matrix (GLCM) for texture coding. Grey level co-occurrence matrix $M_{(x, y)}(d, \theta)$ or second order histogram (which consider the relationship between groups of two pixels in the original image) was initially defined by Haralick (Haralick et al., 1973). Since then, GLCM has been widely used in remote-sensing and analyzing satellite images. In most of the cases, this method is used in texture segmentation.

By simple definition, GLCM is a tabulation of how often different combinations of pixel values (grey levels) occur in an image. When divided by the total number of neighboring pixels $R_{(x, y)}(d, \theta)$ in the image, this matrix becomes the estimate of the joint probability $p_{(d, \theta, x, y)}(i,j)$ or $p(i,j)$ of two pixels, a distance d apart in a direction θ having particular (co-occurring) grey values i and j. Moreover, x and y represent the spatial position of matrix. The dimension of GLCM is GxG where G is the number of grey-levels used to construct the matrix.

Generally, GLCM is computed over a small square window of size N centered at a certain pixel (x, y) and then window is moved by one pixel in the same manner like convolution kernel. Fine texture description requires small values of d and/or small window size, whereas coarse texture requires large values of d and/or large window size. An average over all orientations is taken so that these matrices are rotation invariant. Figure 2 (a) shows an example of construction of grey-level co-occurrence matrix for d = 1 and θ = {0°, 180°} and {90°, 270°}. The matrix $M_{(0, 0)}$ (1, 180°) is just the transpose of $M_{(0, 0)}$ (1, 0°). So to cover all orientations (8 in this case), we need only to compute first four orientations.

It is known that feature based algorithms are generally more stable than raw data (grey levels) based algorithms so a number of features can be calculated using the co-occurrence matrix (containing G^2 elements) for texture discrimination. Haralick defined 14 such features. We have used six out of these fourteen features in our work. These features are: contrast, homogeneity, dissimilarity, energy, entropy and correlation and are listed in figure 2(b). As we can see, to calculate different features, the joint probability density of grey level co-occurrence computed by GLCM is weighted differently. The first three features (contrast, dissimilarity, homogeneity) can be grouped and can be named as *"Contrast Group"* .They compute quantity of contrast in a window. The second group called *"Orderliness Group"* contains features (energy and entropy) which indicate how regular (orderly) the pixel values are within the window. C*orrelation* measures the linear dependency of grey levels on those of neighboring pixels.

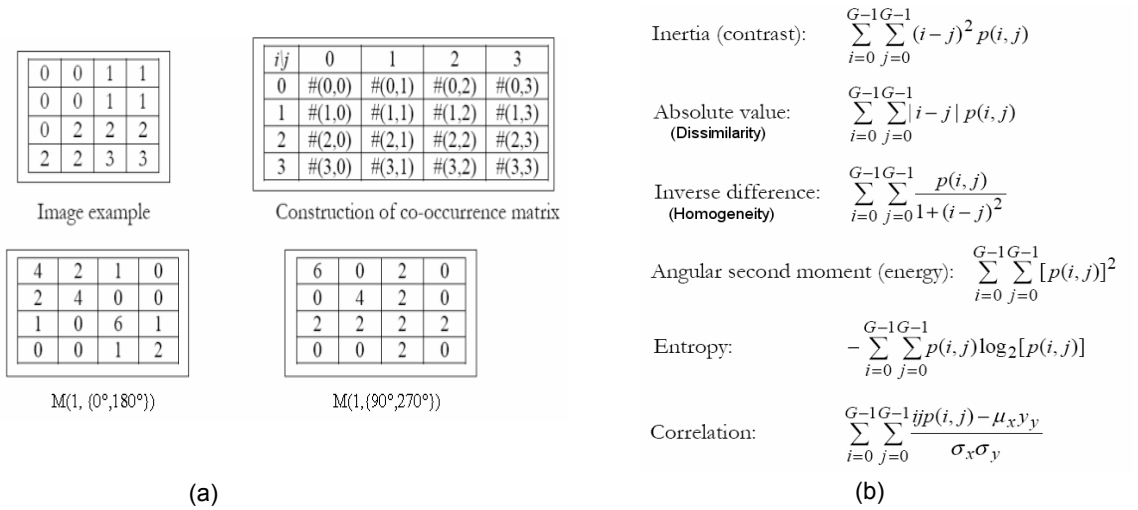

Figure 2 Haralick Features (Haralick et al. 1973): (a) Construction of co-occurrence matrix (b) Texture features

3.2 Classification

To classify pixels as text or non-text, we tested various generative and discriminative classifiers.

The generative classifiers are parametric, based on the assumption that features joint probability distribution for a certain class is multi-variate gaussian. In our experiments, we used one or more (two in our case) gaussians for each class. Given two classes – text and non-text and number of gaussians, the classifiers are: mono gaussian for text and non-text class (TNTSG), mono gaussian for text class (TSG), multi gaussian for text and non-text class (TNTMG), multi gaussian for text class (TMG). In multi gaussian models, a gaussian mixture model containing two gaussians, is trained using Expectation-Maximization (EM) algorithm. Gaussian parameters i.e. mean and covariance matrix for each class are estimated on a training database by using maximum likelihood estimator. In test, Mahalanobis distance is used to compute likelihood estimate.

The discriminative classifier (NNC) is a two-layer perceptron with 6 inputs, N_H hidden cells and 2 outputs. The number of hidden cells (N_H) used in our analysis are 20, 40 and 60, resulting in three discriminative classifiers. Texture features values from training database are normalized between –1 and 1 before feeding to neural classifier and the desired outputs are example labels: 1 for text, -1 for non-text. The network is trained for 10000 epochs with cross-validation stopping. One fourth of the training database is used in cross-validation.

4. Experimental Results

4.1 Database

We have used ICDAR 2003 robust reading and text locating database (ICDAR database., 2003) in our experimentation. The publically available trial database is already divided into two parts: TrialTrain and TrialTest. However, in our experimentation, we have used a total of 100 images taken from TrialTrain part. These images contain text with various font sizes, word lengths, orientations and colors. The size of images varies from 640x480 to 1024x768. There are 433 text segments in the images and font size varies from 10 pixels to 300 pixels. Out of these 100, 50 images are used for training and other 50 for test. For training different classifiers, 100,000 text examples (pixels in this case) and 100,000 non-text examples (pixels) are taken randomly from 50 images. As a preprocessing step, images are converted to grey scale. No other preprocessing is employed.

4.2 Computation of grey-level co-occurrence matrices and texture features

We compute GLCMs over a small square window of size N centered at a certain pixel (x, y) and then window is moved by one pixel in convolution kernel manner. GLCMs are computed in 8 directions (E, NE, N, NW, W, SW, S, SE) or (d = 1, θ = 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°) and an average is taken so that these features are rotation invariant. In actual implementation only four orientation matrices are needed to be calculated and the other four orientations can be computed using transpose of these matrices. Moreover, five different square windows with size N = 5x5, 7x7, 9x9, 11x11, 13x13, 17x17 are used.

Due to intensive nature of computations, reduction of number of intensity levels (quantizing the image to a few levels of intensity) helps increase the speed of computation with negligible loss of textural information. The grey-levels are uniformly quantized to G levels between minimum and maximum grey levels of a window. We choose 32 grey-levels in our implementation. Once GLCMs are computed, texture features can be calcluted by the equations in figure 2(b). Average feature calculation time for an image of 640x480 pixels using un-optimized C++ code running on Pentium IV 3.2GHz for various windows is given in table 1.

Window Size (N)	**5x5**	**7x7**	**9x9**	**11x11**	**13x13**	**17x17**
Time (seconds)	43	51	67	88	133	196

Table 1 Features calculation time

4.3 Text detector evaluation method

To evaluate the performance of a certain text detector, we adopt a pixel based evaluation mechanism. The target images (binary in nature) in ICDAR 2003 database contains one or more black (pixel values = 0) rectangular areas representing text regions. The pixel value for non-text pixels is 1. The database is designed for text localisation. However, in our scheme, due to absence of localisation step which generates rectangles around text strings, we have to evaluate performance of text detector with the given target images where text regions are represented by rectangular regions and figure-ground data is missing. Hence, such an estimate is biased and the actual detection rate is higher than the calculated. The text detector generates either 0 (for text) or 1 (for non text) for each pixel of the input image. In pixel based evaluation mechanism, the output of text detector is compared with the target and a confusion matrix is created. For evaluation, two quantities, text detection rate and false alarm rate are computed.

4.4 Text detector results

Table 2 summarize the performance of different text detectors. We observed that that two class model (TNTSG or TNTMG) is better than the single class model (TSG or TMG). Moreover, mono gaussian works better than two gaussians model. The neural classifier with 20 hidden celles gives the best results: text detection rate is 64% and false alarm rate is 25%. Text detection results on some of the images are shown in figure 3.

Classifier Type	**TNTSG**	**TSG**	**TNTMG**	**TMG**	**NNC**		
					N_H = 20	**N_H = 40**	**N_H = 60**
Best Window Size	17x17	17x17	5x5	17x17	17x17	17x17	17x17
Text Detection Rate (%)	48.8	38.0	61.8	33.0	64.1	64.0	64.5
False Alarm Rate (%)	14.4	14.4	44.5	14.4	25.1	25.5	25.6

Table 2 Comparison of various text detectors

5. Conclusions and Prospects

In this paper, we have employed a texture features computed from grey-level co-occurrence matrices text detection in natural scene images. Our text detectors work on a wide range of text font sizes and fonts (see figure 3). Although, the performance is evaluated on a small test database of 50 images but the results are encouraging and we hope that performance evaluation of these text detectors on a larger database will validate these results and conclusions. Different classifiers have been employed and we have found that mono gaussian models are more robust than multi-gaussian models and two class (text and non-text class) model is better than single class (text class) model. The best classifier is multi-layer perceptron which gives a text detection rate of 64% and false alarm rate is 25%. Till now, we have not filtered any detected text region by applying validation methods e.g. geometrical and spatial constraints, baseline detection, character alignment etc. We believe that such validation schemes will lower the false alarm rate. Furthermore, we are exploring gradient methods as they can differentiate text and non-text regions. Gradient methods are rapid in calculation so one such method can be used to generate candidate text regions which can further be processed by our proposed texture scheme, thus making overall process fast.

Figure 3 Text detection results: original images (row 1) and classification results (row 2). Neural classifier (6 inputs, 20 hidden cells and 2 outputs) using window size 17x17.

References

Haralick, R., K. Shanmugam and I. Dinstein (1973). Textual features for image classification, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610-621.

Ferreria, S., C. Thillou and B. Gosselin (2003). From picture to speech: an innovative application for embedded *environment,* Proceedings of 14th PRORISC Workshop on Circuits, Systems and Signal Processing.

ICDAR database (2003). *ICDAR Robust Reading and Text Locating Competition*, http://algoval.essex.ac.uk/icdar/RobustReading.html

Velázquez, R., F. Maingreaud and E.E. Pissaloux (2003). Intelligent glasses: a new man-machine interface concept integrating computer vision and human tactile perception, *EuroHaptics 2003*, Dublin, Ireland.

Ezaki, N., M. Bulacu and L. Schomarker (2004). Text detection from natural scene images: towards a system for visually impaired persons, *Proceedings of 17th International Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 683-686.

Liang, J., D. Doermann and L. Huiping (2005). Camera-based analysis of text and documents : a survey*, International Journal on Document Analysis and Recognition (IJDAR)* vol. 7, pp. 84-104.

Silapachote, P., J. Weinman, R. Weiss and M.A. Mattar (2005). Automatic sign detection and recognition in natural scenes, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol 3. pp. 27.