Combining Statistical Measures to Find Image Text Regions

P. Clark and M. Mirmehdi
Department of Computer Science,
University of Bristol,
Bristol, UK, BS8 1UB,
{pclark,majid}@cs.bris.ac.uk

Abstract

We present a method based on statistical properties of local image pixels for focussing attention on regions of text in arbitrary scenes where the text plane is not necessarily fronto-parallel to the camera. This is particularly useful for Desktop or Wearable Computing applications. The statistical measures are chosen to reveal charactersitic properties of text. We combine a number of localised measures using a neural network to classify each pixel as text or non-text. We demonstrate our results on typical images.

1. Introduction

To automatically enter the contents of a text document into a computer, one can place it on a flatbed scanner and use state of the art Optical Character Recognition (OCR) software to retrieve the characters. However, automatic segmentation and recognition of text in arbitrary scenes, where the text may or may not be fronto-parallel to the viewing plane, is an area of computer vision which has not been extensively researched previously. The problems involved are to first locate the text, then align it correctly to obtain a fronto-parallel view, and finaly pass it to an OCR system or a human observer for higher level interpretation. In this paper we are concerned with the first stage of this task.

The research into retrieval of text from 3D scenes also has applications for intelligent robots which gain information from text in their surroundings, replacing the document/photograph scanner with a point and click camera, aid for the visually impaired, general Wearable Computing tasks benefiting from knowledge of local text, and automated tasks requiring the ability to read where it is not possible to use a scanner.

Wu et al. [4] used K-means clustering of the average local energy of an image's derivatives to differentiate the text regions from the rest of the image. This segmentation assumed that text has a large number of edges along the horizontal direction. Jain et al. [3] used a three-layer neural network for text classification in a document. With different training sets, they applied their network to classify between English and Chinese scripts, and to classify between photos, text and line-art regions. They also attempted to find text in images of real scenes, by looking for rectangular regions with a large horizontal spatial variance. Chen and Chen [1] tried to differentiate text regions from graphics on journal covers and found that across a text region there is a low variance in the spatial density or the ratio of text to background pixels. This is due to the even spacing of text, and the fact that most characters have a similar ratio of their area to the amount of space they occupy.

The research mentioned above are amongst many other examples which assume the text to be face-on in the image, usually a scanned document. Here, we want to be able to identify text which may be at an orientation to the camera. In [2] we presented a method to locate and recover all regions of text in the image by first extracting local information such as page borders and edges around text. While these provided good results, the edge extraction and line finding stages involved relied on thresholds that can vary from one scenario to another. The method in [2] also assumed that each document in the scene has borders to be recovered. However, some documents may be overlaid or their edges may not contrast well enough against their background. Here, we therefore report on an alternative method to locate regions of text which eliminates such problems and which is based around the local image statistics. We combine a number of measures using a neural network to classify the text regions.

We have directed our work to finding and recovering paragraphs and blocks of text rather than single words or lines. As well as readable text, we desire recognition in the case when the text is too small to read and location of text at an unreadable angle, to facilitate an autonomous robot to decide to move into a suitable position to read the text, or a computer controlled camera (wearable or otherwise) which can zoom in on the text in order to read it. The advantage

this facility gives these applications is that the resolution of the camera may be minimised.

2. Statistical Measures

The human visual system can quickly identify text-like regions without having to examine individual characters, even when the text is too far away to read. This is because text has textural properties that differentiate it from most of the rest of a scene. We now present four texture-like statistical measures, $M_s, s \in \{1..4\}$, which are applied to an input image to determine its regions of text.

Each of the statistical measures responds differently to different properties of text. We apply the measures M_s in small neighbourhoods across the image. For each measure a new image is generated where each pixel in the new image represents the result of the measurement applied to the neighbourhood of the corresponding pixel in the original image. The values chosen for the radii of the neighbouhood masks employed vary for each measure and are discussed later.

• Measure M_1: We use the variance of the greylevel histogram over a circular neighbourhood of radius 3 (total size N pixels) at each pixel as a measure of how much local information there is.

$$M_1 = \sum_{i=1}^{N} (H(i) - \bar{H})^2 \tag{1}$$

where $\bar{H}$ is the mean intensity of histogram H. We are interested in areas of medium variance since text has information, but small and medium scale text undergoes aliasing at the boundaries where text and background greylevels mix, which results in regions of not vastly contrasting intensities. High variance regions generally indicate extreme high frequency changes, such as a single sharp edge. The output for this measure is shown in Figure 1(b).

• Measure M_2: Text regions have a high density of edges. We measure this density in a circular neighbourhood of radius 6 centred at each pixel by summing all edge magnitudes located with a Sobel filter. Although this measure is similar to variance, the output for this measure shown in Figure 1(c) demonstrates that it is more invariant to changes in lighting.

• Measure M_3: Chen and Chen's [1] continuous spatial density assumption (given a flat-bed scanner view of a document) states that the ratio of text to non-text intensity greylevels should not vary greatly as we pass over a text region. We apply the principle that there will be only a small change in local greylevel histograms across a text region (we reuse the histograms computed for measure M_1). We compute the distance between histogram H and its eight-connected neighbouring histograms G_i as:

$$M_3 = \sum_{i=1}^{8} \sum_{j=1}^{B} (H(j) - G_i(j))^2 \tag{2}$$

where B is the number of histogram bins. By evaluating the difference between one region and its neighbours, we find the stability of spatial density. This generates the image in Figure 1(d) which shows little change across the text regions.

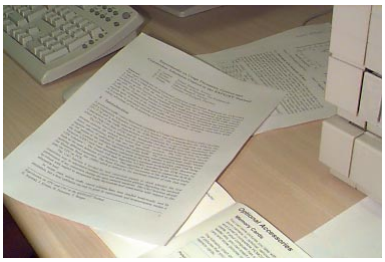
• Measure M_4: In high resolution images one expects to find a high number of edges in a text region, and the angles of the edges to be well distributed due to the presence of curves on many characters. However, this will not be the case at low resolution, where individual characters merge and edges follow the tops and bottoms of text lines. Figure 2 shows the distribution of edge angles in the large text region of Figure 1(a). We observe that there is a tendency for the magnitude of edges in one direction to be matched by edges in the opposite direction of equal magnitude. More specifically, each edge of a character is likely to be accompanied by an edge in the opposite direction, found on the opposite side of the text character or stroke. We draw the hypothesis that over a text region the histogram of the edge angles has rotational symmetry. Hence, M_4 is a measure of the strength of *asymmetry* using a localised edge angle histogram, A:

$$M_4 = \frac{1}{E} \sum_{\theta=0}^{\pi} (A(\theta) - A(\theta + \pi))^2 \tag{3}$$

where $A(\theta)$ is the total magnitude of edges in direction θ, and E is the overall edge magnitude which normalises the result. θ is incremented in steps of $\frac{\pi}{8}$ which is an adequate resolution. We perform this across the image in a circular neighbourhood of radius 16 centred at each pixel. This produces the image in Figure 1(e).

None of the measures $M_s, s \in \{1..4\}$ uniquely identifies a text region. Each one also responds to some non-text areas of the image. The measures are designed to complement each other, so that incorrect decisions by one of them can be corrected by others. In Section 3 we show how the measures are combined to classify text regions.

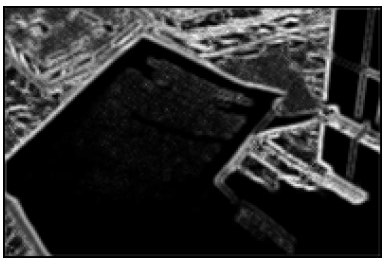
Circular masks are employed for generating the histograms, finding means and searching for edges. The radii of these masks are important. If they are too small then text regions may be broken up where there are gaps between words and paragraphs. If they are too large, different text regions may overlap, small text regions may be missed, and processing time is wasted. The different measures also operate with different mask sizes. For example, measure M_4 requires a larger area of the image than the other measures because it is sensitive to overlapping one half of a text line. The optimum size of the masks depends on the size of the



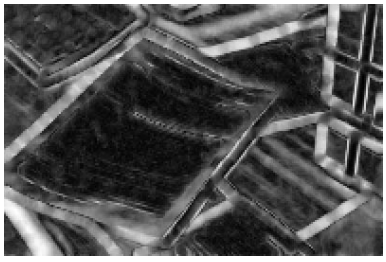
(a) Original image

(b) Measure M_1 output

(c) Measure M_2 output



(d) Measure M_3 output



(e) Measure M_4 output

Figure 1. Example image and associated statistical measures.

text we are looking for. Multiresolution methods (performing processing at different scales) such as in [4] offer one solution to this problem. Alternatively, to scan at a higher or lower scale we can change the size of our masks. For the experiments reported here the radii were determined empirically to work for small and medium size text and are kept constant throughout.

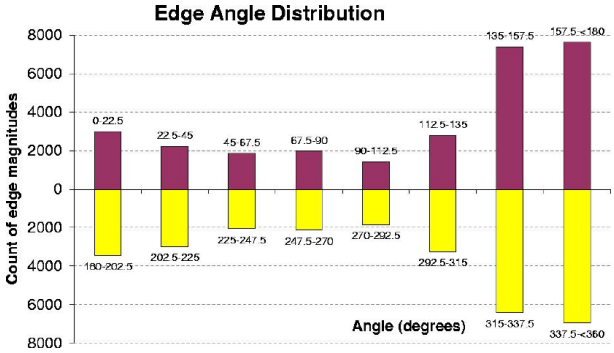


Figure 2. Histogram of edge angle values between $0^\circ - 360^\circ$

3. Combining measures

The outputs of the four measures can be thresholded and then combined with a boolean AND operation to produce a new image with all the text regions classified. This would not be a stable approach and it is preferable to avoid the use of thresholds. Instead, we have introduced a three-layer neural network to use the data from all of the measures simultaneously and make a classification based on the combination of measure values for each pixel. The measures are normalised before input to the network to have zero mean and standard deviation of 1. The final result is a total classification of the image into text and non-text regions.

Four nodes are provided in the hidden layer to find consistencies and relationships in the distribution of the measures. The network has two output nodes, which compete to classify a pixel as text or non-text. We trained the network with 11 hand-labeled images. From each image, the measures from 200 positive (text) and 200 negative (non-text) regions were used as training data, resulting in 4400 training patterns. The desired outputs were given as $\{1, 0\}$ for a text region, and $\{0, 1\}$ for a non-text region. Learning was performed using standard back-propogation algorithm for 300 iterations. During testing, each image is scanned using our circular windows. The measures for each pixel are put into the neural network, and each output node returns a probability value (see Figure 3(a)), the maximum of which is chosen as the text or non-text classification (see Figure 3(b)). Some more classified images are shown in

(a) Neural network output visualisation

(b) Final located text regions

Figure 3. Output from the neural network and final classification.

Figure 4 with two images from a sequence. In the examples shown our over-confidence has resulted in some non-text areas being classified as text. We regard these false-positives as an acceptable pay-off to ensure that no real text regions are missed. These regions could be rejected by higher level processes.

4. Conclusions

We presented a novel method of finding text regions in images where the document is not aligned fronto-parallel to the camera view, and the size and greylevel of the text is unknown. Four complementary local pixel neighbourhood measures were introduced. These were fed as input features into a neural network to classify pixels as text. By focussing attention on text regions we can direct higher level processing steps more efficiently. We have avoided the use of thresholds and the parameters we employ, such as circular masks radii, are kept constant throughout.

For large text, we could increase our mask-size but that would be computationally expensive. We are currently investigating other approaches, such as multiresolution analysis. However, it can be noticed in Figure 4 that some of the larger text regions have been successfully selected. We attribute this to the introduction of larger text regions in the neural network training and to the scale invariance of measure M_4.

Figure 4. More example images (with the top two from a sequence) and located text.

References

[1] W.-Y. Chen and S.-Y. Chen. Adaptive page segmentation for color technical journals cover images. *Image and Vision Computing*, 16(12):855–877, August 1998.

[2] P. Clark and M. Mirmehdi. Location and recovery of text on oriented surfaces. *Proc. of SPIE Conference on Document Recognition and Retrieval VII*, pages 267–277, Jan 2000.

[3] A. Jain, B. Yu, Y. Zhong, O. Trier, and N. Ratha. Document processing research at Michigan State University. *Proc. 1995 Symposium on Document Image Understanding Technology*, pages 126–140, 1995.

[4] V. Wu, R. Manmatha, and E. M. Riseman. Finding text in images. In *Proc. 2nd ACM Int. Conf. on Digital Libraries*, pages 3–12. ACM Press, July 23–26 1997.