Thresholding Video Images for Text Detection

Eliza Yingzi Du and Chein-I Chang
Remote Sensing Signal and Image Processing Laboratory.
Department of Computer Science and Electrical Engineering,
University of Maryland Baltimore County
1000 Hilltop Circle, Baltimore, MD 21250
Paul D. Thouin
Department of Defense, Fort Meade, MD 20755

Abstract

Thresholding video images is very challenging due to the fact that image background generally has low resolution and is also more complicated and highly distorted than document images. As a result, thresholding methods that work well for document images may not work effectively for video images in some applications. This paper investigates the issue of thresholding video images for text detection and further develops a relative entropy-based thresholding approach that can effectively extract text from complicated video images. In order to demonstrate its performance a comparative study is conducted among the proposed thresholding methods that are developed for gray scale document images. In particular, the two commonly used objective criteria for thresholding, uniformity and shape measures are shown to be inappropriate measures for video image thresholding. In fact, the one that yields the best text detection results generally has very low uniformity and shape values, which are supposed to be very high for gray scale document images.

1. Introduction

Information retrieval from video images has become an increasingly important research area in recent years. The rapid growth of digitized video collections is due to the widespread use of digital cameras and video recorders combined with inexpensive disk storage technology. Textual information contained in video images can provide one of the most useful keys for successful information indexing and retrieval. Keyword searches for scene text of interest within video images can provide additional capabilities to the search engines. In video images, text characters generally have much lower resolution and dimmer intensity than binary document characters. In addition, video text characters may also have various colors, sizes, styles, and orientations within the same image. Furthermore, the video background is generally much more complex than document images. A combination of this complex background and a large variety of low-quality characters cause thresholding algorithms that are designed for document image processing to perform poorly on video images. In this paper, we investigate this issue and further develop a relative entropy-based approach, referred to as joint relative entropy (JRE), to thresholding video images. It uses relative entropy, also known as Kullback-Leibler information distance, to measure uncertainty resulting from gray level transitions between background and foreground where the co-occurrence matrix is used to account for spatial correlation. As demonstrated by experiments, the proposed JRE performs significantly better for video images than those thresholding methods that are developed for gray scale document images. In particular, the two commonly used objective criteria for thresholding, uniformity and shape measures are shown to be inappropriate measures for video image thresholding. In fact, the one that yields the best text detection results generally has very low uniformity and shape values, which are supposed to be very high for gray scale document images.

2. Relative Entropy-based Thresholding Method

Entropy-based and relative entropy-based thresholding have received considerable interests recently [1,2]. Of particular interest is the relative entropy approach developed by Chang et al. in [2]. Let $W = [p_{ij}]_{i,j}$ be the co-occurrence matrix with the $(i,j)$ element given by gray-level transition probability from gray level $i$ to $j$, $p_{ij}$. Let $t$ be a threshold used to threshold an image. It partitions the co-occurrence matrix defined by Eq. (1) into four quadrants, namely, $A$, $B$, $C$, and $D$ given by

$$P_{A} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{ij}, \quad P_{B} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{ij},$$
$$P_{C} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{ij}, \quad P_{D} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} p_{ij}$$

(1)

The probabilities in each quadrant can be further obtained by normalization, called "cell probabilities",

$p_{ijA} = p_{ij} / P_{A}$, $p_{ijB} = p_{ij} / P_{B}$, $p_{ijC} = p_{ij} / P_{C}$, $p_{ijD} = p_{ij} / P_{D}$.

(2)

Now assume that $t$ is a selected threshold value. Let $h_{ij}$ be the transition probability of the $t$-thresholded binary image in response to $p_{ij}$. We can define their corresponding cell probabilities of a thresholded image in each of four quadrants given by

$$h_{ijA} = p_{ij} / P_{A}$$
$$h_{ijB} = p_{ij} / P_{B}$$
$$h_{ijC} = p_{ij} / P_{C}$$
$$h_{ijD} = p_{ij} / P_{D}$.

(3)
\[ h_{ij}^t = q_A^t / (t + (t + 1)), \]
\[ h_{ij}^t = q_B^t / (t + (t + 1)), \]
\[ h_{ij}^t = q_C^t / (L - t - 1)(L - t - 1), \]
\[ h_{ij}^t = q_D^t / (L - t - 1)(t + 1). \]

The relative entropy between the probability distributions \( \{ p_{ij} \} \) and \( \{ h_{ij}^t \} \) is then defined by
\[
J_R(p_{ij}; h_{ij}^t) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{h_{ij}^t}
\]
\[
= -H(p_{ij}) - \sum_{i,j} p_{ij} \log h_{ij}^t = -H(p_{ij})
\]
\[
- \left( p_{ij}^B \log q_{ij}^B + p_{ij}^C \log q_{ij}^C + p_{ij}^D \log q_{ij}^D \right) \tag{3}
\]
where \( H(p_{ij}) \) is the entropy of \( \{ p_{ij} \} \) and independent of \( t \). The thresholding algorithm developed by Chang et al. in [2] finds a threshold value that minimizes
\[
J_{RE}(t) = -p_{ij}^B \log q_{ij}^B - p_{ij}^C \log q_{ij}^C - p_{ij}^D \log q_{ij}^D.
\tag{4}
\]
That is, \( t_{RE} = \arg \min_{t \in G} H_{RE}(t) \).

In Eq. (4), the optimal threshold minimizes the relative entropy in all the four quadrants, \( A, B, C, \) and \( D \). However, since we are interested in text detection, a more logical approach is to focus on the gray level transitions between text and background which occur in quadrants \( B \) and \( D \). In this case, we need to introduce a normalization factor \( p_{ij}^{BD} = p_{ij} / (p_B^t + p_D^t) \) to normalize the probabilities in quadrants \( B \) and \( D \). The resulting relative entropy is referred to as joint relative entropy (JRE) given by
\[
J_{RE}(p_{ij}^{BD}; h_{ij}^t) = \sum_{i,j} p_{ij}^{BD} \log \frac{p_{ij}^{BD}}{h_{ij}^t}
\]
\[
= -H_{BF+BD}(t) - \sum_{i,j} p_{ij}^{BD} \log h_{ij}^t
\]
\[
= -H_{BF+BD}(t) - q_{ij}^B \left( p_B^t + p_D^t \right) \log \left( q_B^t / (p_B^t + p_D^t) \right)
\]
\[
- q_{ij}^D \left( p_B^t + p_D^t \right) \log \left( q_D^t / (p_B^t + p_D^t) \right) \tag{5}
\]
where \( H_{BF+BD}(t) = -\sum_{i,j} p_{ij}^{BD} \log p_{ij}^{BD} \) is the entropy of quadrants \( B \) and \( D \). The JRE approach finds a threshold \( t_{RE} \) that minimizes \( J_{RE}(p_{ij}^{BD}; h_{ij}^t) \), that is, \( t_{RE} = \arg \min_{t \in G} J_{RE}(p_{ij}^{BD}; h_{ij}^t) \).

3. Objective Measures

In order to avoid human interpretation, two objective measures, uniformity and shape, suggested in [3] have been used to evaluate thresholding performance analysis.

3.1. Uniformity Measure

The uniformity measure is generally used to describe the homogeneity of regions in an image. For a given threshold \( t \), it is defined by
\[
U(t) = 1 - \frac{\sigma_B^2 + \sigma_F^2}{C},
\]
where \( C = \frac{\sum_{x,y} f(x,y)}{N_B}, \sigma_B = \sum_{x,y} f(x,y) - \mu_B^2, \sigma_F = \sum_{x,y} f(x,y) - \mu_F^2, \mu_B = \sum_{x,y} f(x,y) / N_B, \mu_F = \sum_{x,y} f(x,y) / N_F, N_B = |B| \) and \( N_F = |F| \) are the number of pixels in background and foreground region respectively.

3.2. Shape Measure

The shape measure is generally used for measurement of the shapes of objects in an image. It is defined by
\[
S_P(t) = \frac{\sum_{x,y} sign(f(x,y) - t) |\Delta(x,y)|}{C_p},
\]
where \( C_p \) is a normalization constant, \( sign(x) \) is the sign function of \( x \), and \( \Delta(x,y) = \sum_{x=1}^2 D_k^2 + \sqrt{2} D_k(D_k + D_A) - \sqrt{2} D_k(D_k - D_A) \).

4. Experiments

Two sets of data were used for experiments, synthetic images and real video images. Two commonly used thresholding methods, Pun-Kapur et al.’s maximum entropy (ME) [4,5] and Otsu’s method [6] were used for comparison. Also included for comparison were joint entropy and local entropy method developed by Pal and Pal in [1] from which Chang et al.’s algorithm and JRE are derived from their algorithms.

4.1. Synthetic Images

Two 256 × 256 synthetic images were created in Figs.1(a) and 2(a) to simulate two scenarios, one with a bright square located in the center of a simple background and another with a light square located at the bottom of a complicated background. Both synthetic images had an identical histogram shown in Figs. 1(b) and 2(b), but had very different spatial correlation which can be visualized in 3-D representations in Figs. 1(c) and 2(c) where \( x,y \)-axis is the spatial coordinates of the image pixels with their corresponding gray level values specified by the z-axis. In Fig. 1, the foreground and background were well-separated by two Gaussian distributions using any threshold value between 164 and 206, whereas, in Fig. 2 the background was simulated using gray levels from 0 to 229. In the latter case, the background was simulated in such a way that the background was gradually moved towards the
foreground with smooth transition as shown in Fig. 2(c). Their uniformity and shape values are plotted in Figs. 1(d-e) and 2(d-e). Since ME and Otsu's method are solely based on image histograms shown in Figs. 1(b) and 2(b), the threshold values produced by these methods for images in Figs. 1(a) and 2(a) were identical and their respective results are shown in Figs.1(f-g) and 2(f-g). Figs. 1(h)-1(k) and 2(h)-2(k) show the results produced by LE, JE, GRE and JRE respectively. Since the image background in Fig. 1(a) is well separated from the foreground as shown in Fig. 1(c), we can expect that Otsu’s method would perform very well which was indeed the case. According to Fig. 1, the best performance was produced by Otsu's method and JRE. Due to gradual transition from background to foreground shown in Fig. 2(c), ME, LE and GRE completely missed the square, while Otsu’s method and JE became having difficulty with thresholding and could only extracted half of the square. JRE was the only one, which could segment the full square out the background. In Table 1, we tabulate the threshold values, uniformity and shape values generated by ME, Ostu’s method, LE, JE, GRE and JRE for both synthetic images where JRE produced the lowest values of uniformity and shape for synthetic image 2. This contradicts the general understanding in image thresholding that a good threshold value should also produce higher values of uniformity and shape. This interesting experiment demonstrates that thresholding video images is very different from that for gray scale images such as document images.

4.2. Video Images

Fig. 3 shows experimental result of a video image which has a single column text in the center of the image. As we can see from Fig. 1(a), the image background is very complicated and contained various low resolution objects. For text detection, the best performance was the one produced by JRE, which completely extracted the text in the video image. Like the experiments of image 2, JRE also produced the lowest values of uniformity and shape in Table 1. Interestingly, except GRE, other entropy-based thresholding methods performed reasonably well by extracting the text along with some portion of the background.

5. Conclusion

This paper develops a relative entropy-based thresholding method, joint relative entropy (JRE) for video images for text detection. It shows that the thresholding techniques that are commonly used for document and gray scale images may not work due to complicated background of video images. Additionally, it also demonstrates that the uniformity and shape measures are not appropriate criteria to measure for thresholding video images. Despite only one real image experiment presented in this paper, many other video images were also conducted for experiments. According to our results, the proposed JRE seems to work effectively in many cases of text detection.

Acknowledgment

The authors would like to thank Department of Defense to support his work through the contract number MDA-904-00-C2120.

References

Figure 1. Thresholding results produced by ME, Otsu’s method, LE, JE, GRE and JRE

Figure 2. Thresholding results produced by ME, Otsu’s method, LE, JE, GRE and JRE

Figure 3. Thresholding results produced by ME, Otsu’s method, LE, JE, GRE and JRE

<table>
<thead>
<tr>
<th></th>
<th>Threshold Value</th>
<th>Uniformity</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>image 1</td>
<td>Image 2</td>
<td>Video image</td>
</tr>
<tr>
<td>ME</td>
<td>115</td>
<td>115</td>
<td>172</td>
</tr>
<tr>
<td>OTSU</td>
<td>161</td>
<td>161</td>
<td>69</td>
</tr>
<tr>
<td>LE</td>
<td>117</td>
<td>118</td>
<td>156</td>
</tr>
<tr>
<td>JE</td>
<td>149</td>
<td>154</td>
<td>177</td>
</tr>
<tr>
<td>GRE</td>
<td>127</td>
<td>126</td>
<td>87</td>
</tr>
<tr>
<td>JRE</td>
<td>212</td>
<td>228</td>
<td>205</td>
</tr>
</tbody>
</table>

Table 1. Threshold value, uniformity and shape values generated by ME, Otsu’s method, LE, JE, GRE and JRE