Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform

Ali Mosleh[1]
mos_ali@encs.concordia.ca
Nizar Bouguila[2]
bouguila@ciise.concordia.ca
A. Ben Hamza[2]
hamza@ciise.concordia.ca

[1] Department of Electrical and Computer Engineering
[2] Concordia Institute for Information Systems Engineering
Concordia University
Montréal, QC, Canada

Figure 1: Several steps of the text detection technique. (a) Original image. (b) Edge map using bandlet transform. (c) SWT of the image. (d) Text and non-text CCs classification. (e) Merging text CCs to produce the final result.

A slew of semantic image content analysis techniques are specialized in extracting text embedded in images since it is a vital source of semantic information. A robust text detection step is the basic requirement for a scheme designed to extract text information from images. Text detection is still a challenging issue due to unconstrained color, sizes, alignments of characters, lighting and also various shapes of fonts, even though various methods have been proposed in the past years [2]. Existing text detectors can be broadly classified into two main groups: texture (also called region) based and connected component (CC) based methods.

The general scheme of our proposed method consists in producing (Fig. 1) the image edge map and then finding CCs based on stroke width transform (SWT) [1] guided by the generated edge map. Next, precise feature vectors are formed using the properties of CCs from SWT and pixel domain. An unsupervised clustering is performed on the image CCs to detect the candidate text CCs. Finally, text candidate components are linked to form text-words. The method is considered as a CC-based technique and the contribution is twofold: 1) Since accurate edge maps drastically enhance SWT results, a precise edge detection approach adaptive to text-regions is proposed by employing the bandlet transform. 2) A feature vector based on text properties and stroke width values is employed in k-means clustering in order to detect text CCs.

Bandlet transform [3] effectively represents the geometry of an image. The image coefficients are dyadically segmented in squares S for polynomial flow approximation of the geometry before the bandletization process. Since the image coefficients are all warped along local dominant flows in the bandlet transform, the final bandlet coefficients generated for each segmentation square S have the form of approximation, and high-pass filtering values appear in the wavelet transform of a 1D signal. We benefit from the bandlet-based resulting 1D high-pass frequency coefficients that are adapted to the directionality of the edge that exists in each segmentation square S in order to find a binary map of the edge positions in the image. Since the approximation part of the bandlet transform resulting coefficients consists of coarse information of the original signal, we discard it and only process the high-pass coefficients. The first-order derivatives of the fine-detail bandlet coefficients are computed. Then, local maxima of the resulting gradient signal are found using a contextual filter as follows:

$$M_i = \begin{cases} 1 & \text{if } g_i > T \ \wedge \ g_i > g_j, \ \forall j \in [i-L, i-1] \ \wedge \\ & g_i > g_j, \ \forall j \in [i+1, i+L] \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where each point in the gradient signal is denoted by x_i. g_i represents the gradient value for x_i and g_j indicates gradient value of neighboring pixels of x_i that exist is a window with size $2L+1$ centered at x_i. T is a threshold value and M is a map of local maxima of the gradient signal. The corresponding locations of 0's of M in the bandlet fine (high-pass) coefficients are set to 0, for all the bandlet squares S. Then, the inverse bandlet transform is performed in order to have the final edge locations of the original image. The quality of the edge map depends on the value of the threshold T. Therefore, the edge detection process is performed with two different values of T. Once, a low value is chosen to assure that no reasonable edge information is lost. Then, a higher value is assigned to T to only capture the significant edges. Filially, the two results are combined to produce the finial edge map (Fig. 1(b)).

Element	Description
V_{SWT}	variance of stroke width values
μ_{SWT}	mean of stroke width values
M_{SWT}	median of stroke width values
R_s	ratio of the component diameter and M_{SWT}
V_G	variance of gradient directions of all the CC's edge pixels
SK_G	skewness estimation for the gradient directions of all the CC's edge pixels
C_L	contrast value of the CC and the background in the CC's bounding box
R_{asp}	aspect-ratio of the bounding box of the CC

Table 1: Elements of the feature vector generated or each CC.

Our text detection approach obtains features for CCs produced by SWT, then decides which CC is a text candidate using k-means clustering. In the first step, we find the edges of the input image by means of the proposed bandlet-based edge detection method. A ray shooting process is performed from each edge pixel along its gradient direction. The number of pixels which lie on the ray between two edge pixels with opposite gradient directions is considered as the stroke width for those pixels. Using a 4-neighboring pixels search, adjacent pixels are grouped if the ratio of their stroke width values is higher than 0.3 and lower than 3 (Fig. 1(c)). Then, features of the produced CCs are extracted and used to find text candidates. Table 1 summarizes the elements of the feature vector generated for each CC which has the following form:

$$\overrightarrow{F} = \{V_{SWT}, \ \mu_{SWT}, \ M_{SWT}, \ R_s, \ V_G, \ SK_G, \ C_L, \ R_{asp}\} \quad (2)$$

The produced $\overrightarrow{F}$ of all the CCs of the image are fed to a k-means scheme and consequently clustered into two groups, non-text and text components (Fig. 1(d)). In order to identify which cluster is associated to the texts and which is not, at the beginning of the process we append a sample text to the end of each input image. Hence, the resulting cluster that contains the sample text components is considered as the group of text components and the rest of the components are discarded. In the last step, the remaining text components which are horizontally aligned and have reasonable distance to each other, for example as far as a character width, are grouped together and form the word components (Fig. 1(e)).

[1] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2963–2970, June 2010.

[2] K. Jung, K. I. Kim, and A. K. Jain. Text information extraction in images and video: a survey. *Pattern Recognition*, 37(5):977 – 997, 2004.

[3] S. Mallat and G. Peyre. A review of bandlet methods for geometrical image representation. *Numerical Algorithms*, 44:205–234, Mar. 2007.