Unsupervised Segmentation of Text Fragments in Real Scenes

Leonardo M. B. Claudino[1], Antônio de P. Braga[1],
Arnaldo de A. Araújo[2], and André F. Oliveira[2]

[1] Centro de Pesquisa e Desenvolvimento em Engenharia Elétrica,
Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
{claudino, apbraga}@cpdee.ufmg.br
[2] Depto. de Ciência da Computação, Universidade Federal de Minas Gerais,
Belo Horizonte, Minas Gerais, Brazil
{arnaldo, fillipe}@dcc.ufmg.br

Abstract. This paper proposes a method that aims to reduce a real scene to a set of regions that contain text fragments and keep small number of false positives. Text is modeled and characterized as a texture pattern, by employing the QMF wavelet decomposition as a texture feature extractor. Processing includes segmentation and spatial selection of regions and then content-based selection of fragments. Unlike many previous works, text fragments in different scales and resolutions laid against complex backgrounds are segmented without supervision. Tested in four image databases, the method is able to reduce visual noise to 4.69% and reaches 96.5% of coherency between the localized fragments and those generated by manual segmentation.

1 Introduction

Text fragments are blocks of characters (e.g. words and phrases) that often appear isolated from one another in scenes containing objects such as traffic signs, billboards, subtitles, logos, or car license plates. Such fragments are visually salient, especially due to high-contrast against the background, spatial properties, and occurrence of vertical borders.

This paper is particularly motivated by the problem of finding vehicular license plates in a scene, for plate recognition. The authors of [1], for instance, introduce a technique for finding license plates based in the supposition that the lines containing the plate have regular gray scale intervals and produce a signature of the plate. In [2], it is noted that there is a significant amount of vertical edges in the region of the license plate. The image is split in equally spaced horizontal lines and, for each line, the vertical edges are tagged when the difference of values is above a given threshold. The regions are formed by merging vertically adjacent tags and each region is a candidate plate.

The work of Mariano et al. [3] looks for evidences of text in a vehicle and is intended to support surveillance and security applications. The method produces clusters in L*a*b* space and each group is tested to decide whether it has pixels

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 399–406, 2005.

that belong to text. Neighbor lines are streaked to indicate the occurrence of pixels in candidate text clusters.

Clark et al. [4] look especially for paragraphs and large blocks of text, and propose five local statistical metrics that respond to different text attributes. The metrics are combined in a three-layer neural network trained with patterns from 200 text and 200 non-text regions extracted from 11 manually labeled images. The paper presents only a few qualitative results and shows that text is found in different scales and orientations.

In the work of Wu et al. [5], the image suffers texture segmentation by applying three second order gaussian derivatives in different scales followed by a non-linear function. Then, k-means (with $k = 3$) is applyed. Post-processing consists in forming and clustering regions in three different scales, false detection reduction and text refinement. Finally, the text is binarized so that it can be fed to a character recognizer. The work assumes that text appears horizontal in the image.

In the following section, it is presented a new unsupervised text segmentation technique. Here, to be considered a text fragment, a region must satisfy three main conditions:

- being at most constituted from vertical edges or borders;
- being long, with two or more characters;
- presenting regularity of vertical borders throughout it's extention.

Like [5,6,7], it models and characterizes text as a texture pattern. In the novel approach, segmentation is done keeping only the vertical detail coefficients images, which are blurred and then binarized. After that, eight-connected binary regions are selected according to its spatial behavior. The resulting regions are mapped back to spatial domain, becoming candidate text fragments. The expected text fragments are selected according to two proposed content-based features, in a final step.

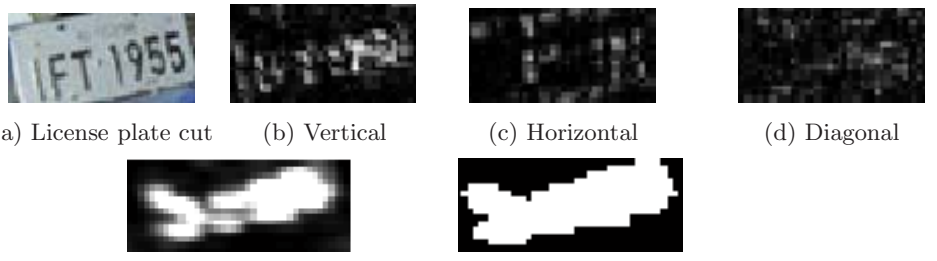
2 Proposed Method

The proposed solution is divided in three procedures. After applying the wavelet transform to the input image and extracting the vertical detail in three different scales, regions are segmented and then selected into candidate fragments, based on spatial aspects. Finally, only those fragments satisfying certain content requirements are considered valid text fragments and output by the method.

2.1 Wavelet Transform

This work adopted the discrete wavelet transform (DWT) and thus brought the process of fragment segmentation to the spatial-frequency (s-f) domain.

A Quadrature Mirror Filter (QMF) bank [8] with three levels of resolution is applied to the input image, with Daubechies-4 being used as the basis function.



(a) License plate cut (b) Vertical (c) Horizontal (d) Diagonal

(e) Cut (b) after blurring. (f) Cut (e) after thresholding.

Fig. 1. Cuts at the text area (license plate) of one of the sample vehicle images: original image (a), vertical (b), horizontal (c), diagonal (d) normalized sub-bands, at the higher decomposition level corresponding image. Vertical coefficients after blurring (e) and thresholding (f).

Only the images of vertical (sub-band) detail coefficients are kept, as they give more information and less noise, compared to the horizontal and diagonal ones (Fig. 1). The vertical coefficients image output from the $p-th$ decomposition level, $\mathbf{V}(p)$, of size $m \times n$ is normalized w.r.t. the sub-band energy $E(\mathbf{V}(p))$, yielding $\mathbf{V}^N(p)$.

$$\mathbf{V}_{ij}{}^N(p) = \frac{[\mathbf{V}_{ij}(p)]^2}{E(\mathbf{V}(p))}, \quad where \tag{1}$$

$$E(\mathbf{V}^N(p)) = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} [\mathbf{V}_{ij}(p)]^2 \tag{2}$$

2.2 Region Segmentation

Region segmentation is performed as an unsupervised clustering on the coefficient images generated in the previous step. This is done by first convolving two unidimensional gaussian masks (a vertical and a horizontal one) and the image, resulting in the filtered image $\mathbf{F}$. The size of the two gaussian windows are kept the same for the three images, which favors the production of larger regions, as decomposition level increases.

The role of the filtering is to group neighboring detail coefficients. This procedure is important because it merges, for instance, fragment coefficients from defective or incomplete characters, or also coefficients from characters that belong to different lines. A binarization threshold computation follows, taking into account the global response of the image to the vertical details. Thus, the threshold θ tells whether a point in $\mathbf{F}$ is salient or not, and is defined in (3).

$$\theta = \mu(\mathbf{F}) + \frac{\sigma(\mathbf{F})}{2} \tag{3}$$

The value $\mu(\mathbf{F})$ stands for the mean of filtered image $\mathbf{F}$, and $\sigma(\mathbf{F})$ for it's standard deviation. Fig. 1 also shows that after gaussian blurring and further binarization the vertical detail coefficients corresponding to the license plate text were properly grouped. The regions are tracked according to the binary connectivity of the component pixels to its eight neighbors. Each of them has its position and bounding-box calculated.

2.3 Spatial Selection of Segmented Regions

First of all, the orientation (or rotation angle) of each segmented region is estimated by performing a linear regression (least-squares) on its binary image pixels, each contributing a pair (x, y).

The bounding-box is rotated about its center, producing the corrected one. The measures w and h are the greatest intervals of the region along x and y, thus corresponding to its real width and height, respectively. The aspect ratio of the region is calculated from the ratio between w and h.

In the next step, spatial coordinates x (first column), y (first line), w and h are mapped back to spatial domain (x^e, y^e, w^e and h^e) following Eqs. 4 and 5, where p corresponds to the current decomposition level.

$$x^e(x) = 2^p \cdot x - \sum_{i=1}^{p} 2^i, \quad y^e(y) = 2^p \cdot y - 3 \cdot \sum_{i=1}^{p} 2^i \tag{4}$$

$$w^e(x) = 2^p \cdot w, \quad h^e(y) = 2^p \cdot h \tag{5}$$

The procedure represented by (4), for each already performed decomposition, doubles the coordinates and subtracts $0.25 \cdot f = 2$ for x and $0.75 \cdot f = 6$ for y, being f the lenght of the decomposition filters adopted ($f = 8$). Values w and h, however, are just doubled.

After spatial characterization, each region is evaluated and must have at least 10 pixels of h^e, values of w^e greater than h^e, and an aspect ratio > 2.

2.4 Content-Based Selection of Candidate Text Fragments

Fragments of text usually have high density of vertical edges, regularly placed from line to line [1,2]. Actually, the second observation is true only if the text is aligned with the capturing device. So, being avaliable well-bounded, horizontally oriented (or rotation-fixed) regions, it is reasonable to suppose that the lines containing text are those with greatest edge density and take a considerable area of the candidate region. The edges in those lines should be also distributed regularly, that is, their central position should not change much from line to line.

In order to describe a candidate fragment according to those text content hypothesis, a simple technique is proposed. It starts by determining the occurrence of relevant transitions in the rotated candidate fragment image. A transition is relevant if the absolute difference of intensity between two co-linear pixels is greater than a percent threshold, k_{MIN}, relative to the greatest difference observed in the whole image. For each line, the transitions are inspected and stored. If the transitions in a line are mostly concentrated before or after the middle of the fragment, that line is discarded, since text characters must occur along the whole line and so must do the corresponding transitions. The algorithm then groups valid, consecutive lines into blocks. The blocks are separated by valleys in the transition profile (Fig. 2) and the block containing more lines is considered to be where the text fragment more probably is. After the most probable block

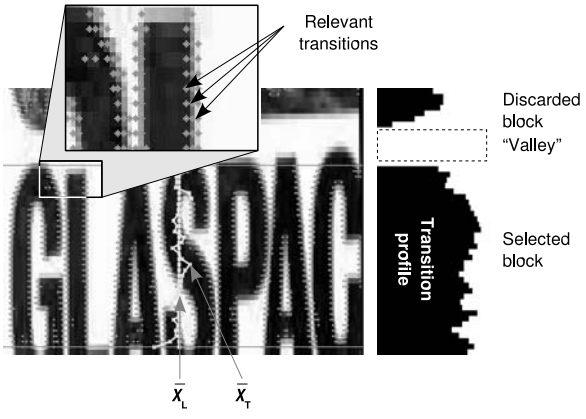


Fig. 2. Relevant transitions and each line transition profile, in a candidate fragment present in one of the test samples

is pointed out, a measurement of the regularity of its transitions, the ρ feature, is calculated, according to (6).

$$\rho = min(\frac{\sum_i |\bar{x}_T(i) - \bar{x}_L|}{\sum_i n_T(i)}, 1), where \tag{6}$$

$$\bar{x}_L = \frac{\sum_{i=1}^{n_L} \bar{x}_T(i)}{n_L} \tag{7}$$

$$\bar{x}_T(i) = \sum_{k=1}^{n_T(i)} \frac{t_i(k)}{n_T(i)} \tag{8}$$

In (8), $x_T(i)$ is calculated as the average x-coordinate of the transitions in line i. In (7), $\bar{x}_L$, the central x-coordinate, is the average of $\bar{x}_T(i)$, for each of the n_L lines as shown in Fig. 2. The value ρ is the integral of the differences of the average position of each line and the central line. It is divided by the total number of transitions in the block, n_T, that appears as the area of the profile also shown in Fig. 2, to quantify the importance of the difference. Since $n_T(i)$ is not an exact normalizer, ρ is saturated at 1.

Another extracted feature is the ratio between the total transitions in the block and total transitions in the fragment it belongs to, n_F, calculated according to (9).

$$a = \frac{\sum_{i=1} n_T(i)}{n_F} \tag{9}$$

The final step in content-based selection is to decide whether the pairs of features (ρ, a) extracted from each of the candidate fragments are to be considered as belongig to a text fragment. For simplicity, here they are only compared to pre-defined thresholds ρ_{MAX} and a_{MIN}. The final results of the method are illustrated in Fig. 3.

(a) Oriented selected regions.

(b) Image reduction.

Fig. 3. Content-based selection of candidate fragments into text fragments. Parameters used: $\rho_{MIN} = 0.2$ and $a_{MIN} = 0.95$).

3 Results and Conclusions

A total of 580 images from four databases were tested (Tab. 1). The results produced by the method were compared to manual segmentation made by three collaborators that marked the bounding boxes of text fragments (or text blocks) in the scenes. The parameters employed were $\rho_{MIN} = 0.35$ and $a_{MIN} = 0.85$. Results were evaluated according to two indicators, calculated after the execution of each of the three phases of the method. The first, true positives (I_{TP}), quantifies the accuracy of the method in terms of fragment finding: the returned fragments are intersected with the manually selected region and the resulting area is divided by the total marked area of the manually selected region. The second indicator, false positives (I_{FP}), evaluates the capacity of removing distractors from the image: the area of returned fragments that do not correspond to the manually selected region is divided by the total area of the image.

Table 1. Summary of image databases employed in the experiment

A. Images of 363 vehicles with visible license plates. Acquired using two digital cameras under different environmental conditions. Plate text in different scales and orientations. Sampling resolution of 320 × 240 pixels. Compressed JPEGs.
B. 100 images of vehicles with visible license plates. Acquired from various websites. Diverse image sizes and file formats, mostly compressed JPEGs.
C. 88 images of vehicles with visible license plates. Acquired from campus surveillance MPEG-compressed videos, the camera is sometimes out of focus. Sample resolution of 640 × 480 pixels.
D. 29 images with diverse text fragments. Includes advertisements, banners, historical documents and office room number plates. Various sizes and file formats.

For database A, 98.21% of the regions of interest are detected by the first phase of the method and 91.82% remain after the region selection stages. Meanwhile, the false positives drop from initial 26.42% to about 5.37%, an average reduction of 20% of the image area. Figs. 4 (a) and (b) present the results for an example from database A and it can be seen that the method deals well with varying perspective. Figs. 4 (c) and (d) show the results for a scene from image

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4. Qualitative results after testing the method in the four image databases

database B. The amount of true positives for this database falls from 96.65% to 91.04%, while false positives decrease from 34.83% to 5.79%. In Figs. 4 (d) and (e), the method finds both the vehicle's license plate and the taxi mark. Keeping 96.50% from the 98.06% of true positives is a very good output for database C,

since it is negatively affected by high video compression and bad focusing of the camera. The decrease in false positives is also high, reaching 4.69%. Figs. 4 (f) and (g) depict the results for one from the 29 available images of database D, a low resolution flatbed scanned historical document image. In that database, text appears with greatest variability. The indicators show a regular true positive rate (73.92% after all) and spurious regions removal around 7.78%.

The method presented here succeeds the proposed goals, since it is demonstrated by I_{TP} and I_{FP} that it reaches 96.50% of true positives and reduces the visual noise to 4.69%, according to manual segmentation. Now that text fragments in arbitrary scenes are efficiently detected by the presented method, it will be integrated to a character recognition system that will operate on the fragments it outputs.

References

1. J. Barroso, A. Rafael, E.L.D., Bulas-Cruz, J.: Number plate reading using computer vision. In: IEEE-International Symposium on Industrial Electronics ISIE'97. (1997) 761–766
2. Setchell, C.J.: Applications of Computer Vision to Road-traffic Monitoring. PhD thesis, Faculty of Engineering, Department of Electrical and Electronic Engineering of the Univesity of Bristol (1997)
3. Mariano, V.Y., Kasturi, R.: Detection of text marks on moving vehicles. In: 7th International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK. (2003) 393–397
4. Clark, P., Mirmehdi, M.: Finding text regions using localised measures. In Mirmehdi, M., Thomas, B., eds.: Proceedings of the 11th British Machine Vision Conference, BMVA Press (2000) 675–684
5. Wu, V., Manmatha, R., Riseman, E.M.: Textfinder: An automatic system to detect and recognize text in images. IEEE Transactions on Pattern Analysis and Machine Intelligence **21** (1999) 1224–1229
6. K. Etemad, D.D., Chellapa, R.: Multiscale segmentation of unstructured document pages using soft decision integration. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 92–96
7. Jain, A.K., Yu, B.: Automatic text location in images and video frames. Pattern Recognition **31** (1998) 2055–2076
8. Smith, J.R.: Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression. Phd. thesis, Graduate School of Arts and Sciences, Columbia University, New York, NY. (1997)