Localization of Caption Texts in Natural Scenes Using a Wavelet Transformation*

Javier Jiménez and Enric Martí

Centre de Visió per Computador – Dept. Informàtica, UAB
Edifici O – Campus UAB, 08193 Bellaterra (Barcelona) – Spain
{Javier.Jimenez,Enric.Marti}@cvc.uab.es

Abstract. Automatic extraction of text from multimedia contents is an important problem that needs to be solved in order to obtain more effective retrieval engines. Recently, Crandall, Antani and Kasturi have shown that a direct analysis of certain DCT coefficients can be used to locate potential regions of caption text in MPEG-1 videos. In this paper, we extend their proposal to wavelet-coded images, and show that localization of text superimposed in natural scenes can also be effectively and efficiently performed by a wavelet transformation of the image followed by an analysis of the distribution of second order statistics on high frequency wavelet bands.

Keywords: Natural-scene statistics, text localization, text segmentation, wavelets, texture analysis, image analysis, computer vision.

1 Introduction

Digital video and still images have become popular, with an increase of media that content-based image retrieval systems are being required to access. Automatic extraction of text from multimedia images is an important problem that needs to be solved in order to obtain more effective high-level retrieval engines.

Previous proposals for text segmentation mostly assumed graphic documents or were developed to work with very structured images and simple distributions of colors (a survey of classical techniques that still are up-to-date can be found in [1]). New functional categorizations [2] recognize photographic images as a second major group, characterized by more colors, less structure and smooth transitions between elements.

To face the problem of extracting text that has been superimposed over natural images, methodologies from texture-segmentation contexts view the text as a texture and apply classical texture segmentation techniques, making very few assumptions about the image. Though textural analysis can be done directly over the input image [3, 4], most of these techniques apply some kind of transformation to the data in order to provide a more efficient representation, for example

* This work has been supported by the Spanish *Ministerio de Ciencia y Tecnología* under grants TIC2003-09291 and TIC2000-0399-C02-01.

A. Sanfeliu et al. (Eds.): CIARP 2004, LNCS 3287, pp. 100–107, 2004.

using Fourier [5], co-occurrent matrices [6], Gabor filters [7], or wavelets [8]. Recently, Crandall, Antani and Kasturi [9] have shown that a direct analysis of certain DCT coefficients can be used to locate potential regions of caption text in MPEG-1 video frames.

In this paper, we extend the proposal by Crandall, Antani and Kasturi [9] to wavelet-coded images, and show that localization of caption text superimposed in natural scenes can also be efficiently performed by a wavelet transformation of the image followed by an analysis of the distribution of second order statistics on high frequency wavelet bands. Examples of caption text addressed in this work can be seen in Fig. 1.

(a) Hateren composition [10] (b) Brodatz composition [11]

(c) Movie credits (d)Web banner

Fig. 1. Examples of caption texts over natural backgrounds

The rest of this paper is organized as follows. In Sect. 2, the previous approach of Crandall, Antani and Kasturi [9] for text localization, which is the base of our proposal, is summarized. In Sect. 3, the new method for text localization is described. Experimental results are reported and discussed in Sect. 4. Conclusions are set out in Sect. 5.

2 A Previous Approach for Text Localization in DCT-Coded Images

The DCT became popular because of its energy packing capabilities while approaching a statistically optimal transform (the KLT) in decorrelating a signal governed by a Markov process; and because efficient algorithms for fast computation of the DCT are known, based on one-dimensional filtering transformations [12].

Recently, Crandall, Antani and Kasturi [9] have proposed a method to detect and locate caption text of arbitrary size and orientation in 8x8 DCT-coded

images, like JPEG for still images and MPEG-1 for video. Specifically, the DCT F_b of an 8x8 image block f_b is given by

$$F_b[u,v] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f_b[m,n] \widetilde{\psi}_{u,v}[m,n] \,, \tag{1}$$

with

$$\widetilde{\psi}_{u,v}[m,n] = \cos(\frac{(2m+1)u\pi}{2M}) \cos(\frac{(2n+1)v\pi}{2N}) \,. \tag{2}$$

The m and n variables stand for the discrete spatial coordinate positions within each input image block f_b; u and v variables are the coordinate positions within the corresponding transformed block F_b; and $M = N = 8$ stand for the horizontal and vertical dimensions of each block.

In [9], the coefficients $F_b[u,v]$ of transformed blocks are directly analyzed in order to locate potential elements of text. A postprocessing is also suggested in order to integrate the detected regions of text and reject false positives while, in parallel, an analysis at different resolutions allows to detect letters of different sizes.

3 A New Method for Text Segmentation Using Wavelets

The method proposed by Crandall, Antani and Kasturi [9] is based on Eq. 1, which is equivalent to an inner product in each sub-block image space $\Omega = \{0,\ldots,7\} \times \{0,\ldots,7\}$ with a base given by an orthornormal system of cosinus functions (Eq. 2).

Analogously, a wavelet transformation (described in Sect. 3.2) can be viewed as an inner product with a biorthonormal system of wavelet and scale functions [13]. The fundamental difference between both approaches is the selection of a specific representation. This common formulation allows us to extend the text localization method proposed in [9] from DCT-coded images to wavelet-coded images.

The next subsections describe step by step our proposed method, which consists in a color space transformation to obtain a gray-level image, followed by a wavelet transformation, and finally a statistical analysis of high frequency wavelet bands to locate potential regions of text.

3.1 Color Space Transformation

This step only needs to be done in the case of color images, and consists in a transformation from RGB to YCbCr, which decorrelates the dependencies between color components into luminance and chrominance bands. Since the chrominance components are less sensitive to the human visual system (HVS) than the luminance component, we will work with the luminance band alone.

3.2 Wavelet Transformation

In this step, a wavelet transformation is used to decorrelate spatial dependencies and obtain a more sparse representation of the gray image.

Olshausen et al. [14, 15] found that localization, orientation and band-pass properties of certain wavelets agree with processes of the human visual system (HVS) and at the same time provide sparse representations for natural images. The Daubechies family of wavelets is indicated as suitable for this kind of image and, in order to choose a specific one, Wang [16] recommends *db8* and *db4*, which give better results than *Haar* [13, 17]. Other approaches make use of these wavelets for text segmentation and recognition tasks, but differ from our proposal in the target context or in the overall methodology: Menoti et al. [8] use Haar wavelets to locate text in postal envelopes; Bhattacharya et al. [18] use Daubechies *db4* wavelets to recognize handprinted numerals; Li and Doermann [19] suggest the use of *db4* wavelets, combined with a neural network, to classify text regions in digital video.

Based on these previous studies, in our experiments (reported in the next section) we choose transforming the input image by Mallat's algorithm with a *db4* wavelet filter [13, 17]. Let $\psi(x)$ be an orthonormal wavelet and let $\varphi(x)$ be its associated scaling function, such that there exists two square-summable functions, g and h, with

$$\psi(x/2) = 2\sum_k g[k]\varphi(x-k) \tag{3}$$

$$\varphi(x/2) = 2\sum_k h[k]\varphi(x-k) \tag{4}$$

Then Mallat's algorithm defines four sub-bands F^{LL}, F^{HL}, F^{LH} and F^{LL}, to form the global transform F:

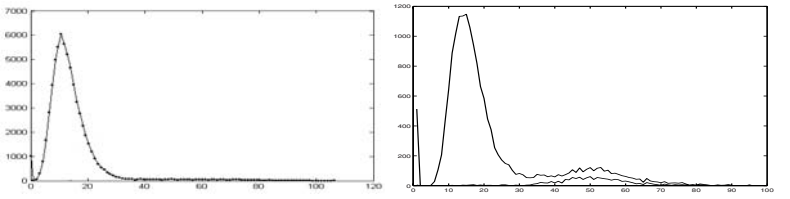
$$\begin{aligned} F^{LL}[m,n] &= \sum_j h[j]\sum_i h[i]f[2m-i,2n-j] \\ F^{HL}[m,n] &= \sum_j h[j]\sum_i g[i]f[2m-i,2n-j] \\ F^{LH}[m,n] &= \sum_j g[j]\sum_i h[i]f[2m-i,2n-j] \\ F^{HH}[m,n] &= \sum_j g[j]\sum_i g[i]f[2m-i,2n-j] \end{aligned} \tag{5}$$

Notice that Mallat's algorithm decimates the signal, but in order to improve the localization one can obviate the decimation or previously double the size of the gray image.

3.3 Feature Extraction and Classification

Our approach extends naturally to wavelet-coded images the text texture energy (TTE) descriptors from DCT coefficients proposed by Crandall, Antani and

Kasturi [9]. To this end, a map of local standard deviations on the LH and HL bands[1] is calculated, having the same dimensions as any of the four wavelet sub-bands. Fig. 2 shows typical histogram distributions of these feature maps.



(a) Unimodal but positively skewed (b) Bimodal

Fig. 2. Feature distributions of Figs. 1a and 1b. Two curves are represented, the major one is the overall histogram (background and text classes) while the other refers only to the text

We have found through different tests with natural images that the global distribution commonly appears as the composition of one or two bell-shaped non-gaussian distributions. In the unimodal case the text becomes blurred into the overall distribution (as in Fig. 2a). In the bimodal case each heap corresponds with one class (see Fig. 2b). In either case, the right part of the global distribution mostly corresponds with the text, and a first separation between classes in this space with a threshold umbralization is similar to using the TTE with DCT coefficients [9]. The two classes overlap only with certain non-natural distributions, making the separation more difficult or impractical (as in Fig. 6). Fig. 3 shows binarization results with Fig. 1 samples.

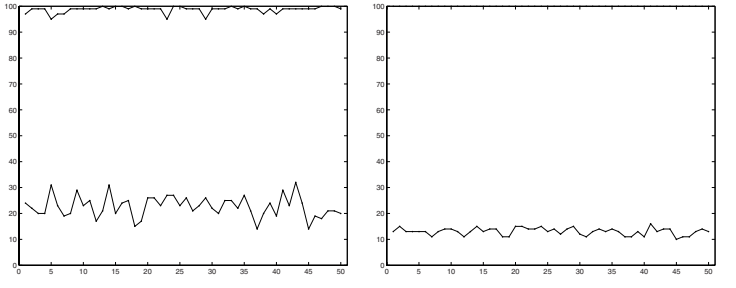
4 Experimental Results

To assess the overall performance of the method proposed in Sect. 3, we have built a test set using the first 50 natural images of the publicly available van Hateren's collection [10], where we have superimposed a constant caption text as in Fig. 1a. A detected pixel is counted as *correctly detected* if it is marked in the ground truth, or as *false positive alarm* if it is not. Non-detected pixels which have been marked in the ground truth are counted as *missed*. To perform the evaluation, the number of correctly detected, false positive alarms, and missed pixels are counted. The results are expressed as recall and false positive (FP) error rates: The recall is defined as the number of correctly detected pixels divided by the number of correctly detected and missed pixels, while the FP rate is defined as the number of false alarms divided by the number of pixels in the image. In this way, good performance corresponds to high recall with low FP

[1] Empirically we found that the inclusion of the HH band in this map does not improve the results.

Fig. 3. Localizations without further postprocessing (see Fig. 1)

rate. Note that a recall index alone is insufficient to measure actual performance, since a recall of 100% could be trivially obtained by classifying every pixel as text.

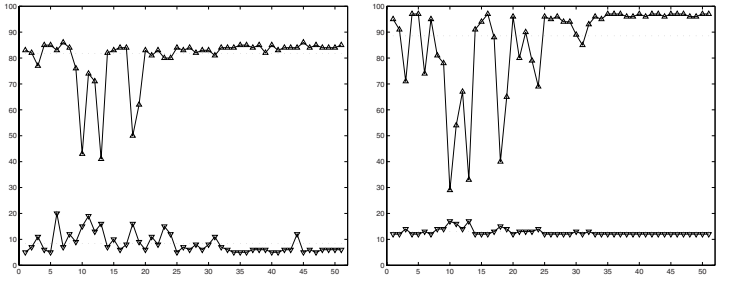


(a) Localization of bounding boxes (b) Extraction of text letters

Fig. 4. Performance evaluation using the van Hateren's collection

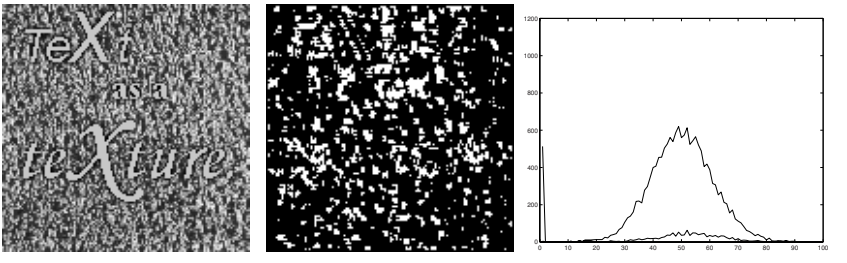
Two performance evaluations are commonly done using the precision and recall indices introduced above: localization of bounding boxes and extraction of text letters. For localization, Fig. 4a shows an average recall rate of 98.78% (upper curve) at the cost of 22.68% of false positives (lower curve). For text extraction, Fig. 4b shows a maximum recall (100%) at the cost of a 13.06% of false positives.

Fig. 5 exhibits similar results with a second test set of 50 images, formed using Brodatz textures [11] with caption text that has been superimposed as in Fig. 1b. An evaluation by bounding boxes (Fig. 5a) show a recall rate of 81.74% (upper curve) with an FP rate of 8.48% (lower curve). Let us highlight the fact that there appear four *outsiders* – that is, four images that sit far apart from the rest – as the one in Fig. 6.



(a) Localization of bounding boxes (b) Extraction of text letters

Fig. 5. Performance evaluation with a second, more difficult, test set of Brodatz textures. One of the compositions appears in Fig. 6



(a) Brodatz composition (b) Localization (c) Feature's distribution

Fig. 6. Unusual test result. Fig. 6a is a particular Brodatz composition where the superimposed text is difficult to be perceived visually. The localization given in Fig. 6b is one of the worst results obtained. Fig. 6c shows how the background's distribution overlaps the text's distribution

5 Conclusions

In this paper we have shown that localization of caption text can be done using wavelet transformations in an analogous way to the method proposed by Crandall, Antani and Kasturi [9] to analyze DCT-coded images. The proposed method has a linear complexity (with respect to the size of the input image) and experimental results show its viability for fast localization of text that has been superimposed over natural scene backgrounds. This method could be particularly well suited to deal with images that have been coded using modern systems based on wavelets, like JPEG2000 and MPEG-4.

References

1. O'Gorman, L., Kasturi, R., eds.: Document Image Analysis. IEEE Computer Society Press (1997) Published as Technical Briefing.
2. Hu, J., Bagga, A.: Categorizing images in web documents. IEEE Trans. on Multimedia (2004) 22–30

3. Allier, B., Duong, J., Gagneux, A., Mallet, P., Emptoz, H.: Texture feature characterization for logical pre-labeling. In: Proc. of Int. Conference on Document Analysis and Recognition. (2003) 567–571
4. Zhong, Y., Karu, K., Jain, A.K.: Locating text in complex color images. In: Proc. of Int. Conference on Document Analysis and Recognition. (1995) 146–149
5. Patel, D.: Page segmentation for document image analysis using a neural network. Optical Engineering **35** (1996) 1854–1861
6. Payne, J.S., Stonham, T.J., Patel, D.: Document segmentation using texture analysis. In: Proc. of Int. Conference on Pattern Recognition. (1994) 380–382
7. Jain, A.K., Bhattacharjee, S.K.: Address block location on envelopes using Gabor filters. Pattern Recognition **25** (1992) 1459–1477
8. Menoti, D., Borges, D.L., Facon, J., Britto, A.S.: Segmentation of postal envelopes for address block location: an approach based on feature selection in wavelet space. In: Proc. of Int. Conf. on Document Analysis and Recognition. (2003) 699–703
9. Crandall, D., Antani, S., Kasturi, R.: Extraction of special effects caption text events from digital video. Int. Journal on Document Analysis and Recognition **5** (2003) 138–157
10. van Hateren, J.H., Ruderman, D.L.: Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. Proc. of the Royal Society of London, Series B **265** (1998) 2315–2320
11. Brodatz, P.: Textures: A photographic Album for Artists and Designers. Dover Publications, N.Y. (1966)
12. Rao, K.R., Yip, P.: Discrete Cosine Transform. Algorithms, Advantages, Applications. Academic Press (1990)
13. Mallat, S.: A Wavelet Tour of Signal Processing. Academic Press (1998)
14. Olshausen, B.A., Field, D.J.: Sparse coding with an overcomplete basis set: a strategy employed by V1. Vision Research **37** (1997) 3311–3325
15. Olshausen, B.A., Field, D.J.: Natural image statistics and efficient coding. Network Computation in Neural Systems **7** (1996) 333–339
16. Wang, J.Z.: Integrated Region-based Image Retrieval. Kluwer Academic Publishers, The Netherlands (2001)
17. Mallat, S.: A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. on Pattern Analysis and Machine Intelligence **11** (1989) 674–693
18. Bhattacharya, U., Chaudhuri, B.B.: A majority voting scheme for multiresolution recognition of handprinted numerals. In: Proc. of Int. Conference on Document Analysis and Recognition (ICDAR). (2003) 16–20
19. Li, H., Doermann, D.S.: Automatic identification of text in digital video key frames. In: Proc. of Int. Conference on Pattern Recognition. (1998) 129–132