A Character Recognition Method in Natural Scene Images

Álvaro González, Luis M. Bergasa, J. Javier Yebes, Sebastián Bronte
Dept. of Electronics, University of Alcalá, Spain
{*alvaro.g.arroyo, bergasa, javier.yebes, sebastian.bronte*}*@depeca.uah.es*

Abstract

Reading text from scene images is a challenging problem that is receiving much attention, especially since the appearance of imaging devices in low-cost consumer products like mobile phones. This paper presents an easy and fast method to recognize individual characters in images of natural scenes that is applied after an algorithm that robustly locates text on such images. The recognition is based on a gradient direction feature. Our approach also computes the output probability for each class of the character to be recognized. The proposed feature is compared to other features typically used in character recognition. Experimental results with a challenging dataset show the good performance of the proposed method.

1. Introduction

Commercial OCR (Optical Character Recognition) systems have a good performance when recognizing machine-printed text in camera-based document analysis. However, they do not work well for reading text in natural scenes, where text is usually embedded in complex backgrounds and many problems arise due to geometric distortions, partial occlusions, changes in illumination, different font styles, font thickness, font color and texture, among others. Therefore, the task of recognizing text in natural images still remains an active research topic. Proof of this is the few works that have competed in the Robust Reading Competitions held in the ICDAR 2003 and 2011 conferences in the challenges of text recognition, where no work was presented in 2003 [8] and only four works competed in 2011 [11].

This paper focuses on the recognition of individual characters in scene images. We propose to use gradient direction features and a classification method that gives different solutions with output probabilities. We compare our proposal to other features. Section 2 describes the features used as well as the classification algorithm. Section 3 provides the experimental results while section 4 concludes the paper.

2. Feature extraction and object classification

Most text-reading systems are composed of a text location algorithm in first place and a text recognition method in second place. Our location approach is explained in [5]. It gives as result binarised objects as the one shown in Fig. 1. Our recognition stage works as follows. It takes each binarised character as input, then it computes its feature vector and the object is classified into a class using a KNN (K-Nearest Neighbors) approach.

We have named the feature used in this paper as Direction Histogram (DH) and it is slightly inspired by [6]. We propose to detect the edge pixels of the binarised objects and then to compute the direction of the gradient for each edge pixel. As it is a binarised image, there is only gradient on the edge pixels, so it is faster to compute. Later we quantize the direction of the gradients in the edge pixels into 8 bins: $\{-135°, -90°, -45°, 0°, 45°, 90°, 135°, 180°\}$, and we compute the histogram for each bin. The image is divided into 16 blocks in order to have spatial information, and the histograms for each block are concatenated into a 128-dimensional vector. As this method is based exclusively on the direction of the edge pixels, it is not affected by color neither intensity. An overview can be seen in Fig. 1.

The classification is based on a KNN approach. The training dataset is composed of 5482 character samples extracted from the train set of the ICDAR 2003 Robust Reading Competition dataset, which has a wide diversity of fonts. Instead of giving only one solution, we propose to give different solutions with output probabilities. Firstly, the nearest K neighbors in the training dataset of the character to be classified are extracted. Each neighbor belongs to a class, *i.e.* each neighbor votes for a certain candidate $S = \{s_1, s_2, \ldots, s_K\}$,

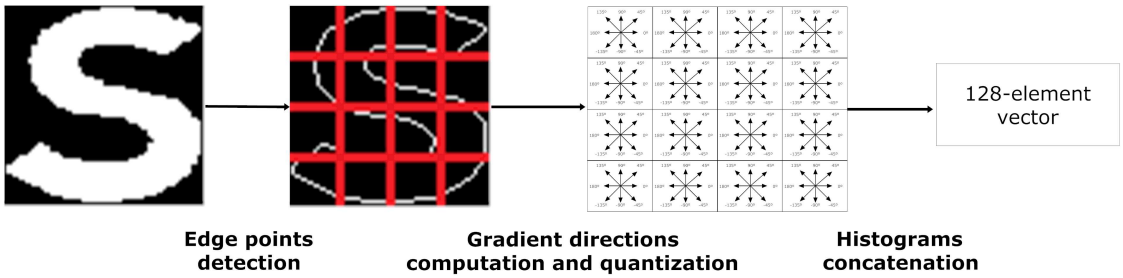

Figure 1. Feature detection

where $s_i \in \{$'A', 'B', ..., 'Z', 'a', 'b', ..., 'z', '0', ..., '9'$\}$ (62 classes). The set of distances from the object to each neighbor is $D = \{d_1, d_2, \ldots, d_K\}$. We define the ratio between each distance to the minimum one as in (1).

$$R = \{r_1, r_2, \ldots, r_K\} = \{1, \frac{d_1}{d_2}, \ldots, \frac{d_1}{d_K}\} \quad (1)$$

We define p as the output probability of the nearest neighbor. We assume that the output probabilities of the following $K-1$ nearest neighbors are related to p by the distance ratios defined in (1). Therefore, it must be fulfilled (2).

$$\sum_{i=1}^{K} r_i \cdot p = p + \frac{d_1}{d_2} \cdot p + \ldots + \frac{d_1}{d_K} \cdot p = 1 \quad (2)$$

The value of p can be easily computed from (2). The output probabilities of the object for every class can be computed using (3). Equation (3) means that the probability of the object of belonging to class 'A' is computed only from the neighbors that correspond to this class. The same is done for class 'B', 'C' and so on.

$$\begin{aligned} p_A &= \sum_{j=1}^{K} r_j \cdot p \qquad \forall j/s_j = A \\ p_B &= \sum_{j=1}^{K} r_j \cdot p \qquad \forall j/s_j = B \\ &\vdots \\ p_9 &= \sum_{j=1}^{K} r_j \cdot p \qquad \forall j/s_j = 9 \end{aligned} \quad (3)$$

With this method, when the object to be recognized is clearly a certain letter, there are many minima that vote for the same class, thus it will have a high output probability for that class. When it is not a clear case, the highest output probability tends to be low, and the worst case would be when each neighbor is at a similar distance and votes for a different class, thus there would be K outputs with comparable probability. Therefore, it must be found a compromise in the value of K. A low value for K could be insufficient to have reliable output probabilities, but a high value could lead to errors, as the solutions with highest output probabilities would tend to those classes with a bigger number of samples. In our case, in which the training dataset is asymmetric, *i.e.* there are classes with a number of elements much higher than other classes, the number of nearest neighbors K has been set empirically to 25.

As the feature proposed is a distribution represented by histograms, it is natural to use the χ^2 test statistic. Therefore, the distances in the classification are computed using (4), where $h_i(k)$ and $h_j(k)$ denote the N-bin normalized histogram for objects i and j respectively.

$$D_{ij} = \frac{1}{2} \sum_{k=1}^{N} \frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)} \quad (4)$$

3. Experimental results

In order to evaluate the robustness of the chosen feature, we evaluate other six different types of local features:

- Shape Context (SC) [2].
- Geometric Blur (GB) [3].
- Scale Invariant Feature Transform (SIFT) [7].
- Gauge Speeded Up Robust Feature (G-SURF) [1].
- Histogram of Oriented Gradients (HOG) [4].
- Local Binary Patterns (LBP) [10].

Table 1 shows the character recognition rate using each kind of feature. Three cases have been analysed. The first one only takes into account the hit rate for the output class with highest probability. The second analysis computes the hit rate for those cases in which the recognition succeeds for either the first or the second solution. Similarly it is done for the first, second and third candidates. It can be clearly seen that DH is the best feature and this method successfully recognizes more than 90% of characters as first or second solution. On the other hand, Fig. 2 shows the character recognition rate as a function of the training dataset size. It can be seen that the hit rate for DH feature tends to an asymptote for a training dataset size of 2000 samples, while the asymptote for other features is reached for a major number of samples.

Table 1. Individual character recognition on ICDAR 2003 dataset.

Features	Hit rate 1st candidate	Hit rate 1st/2nd candidate	Hit rate 1st/2nd/3rd candidate
DH	**76.3%**	**91.4%**	**95.6%**
LBP	67.5%	82.7%	90.0%
SC	59.6%	77.0%	83.4%
SIFT	58.9%	66.8%	68.4%
GB	56.1%	70.1%	75.4%
G-SURF	52.2%	64.0%	70.2%
HOG	48.8%	66.8%	75.4%

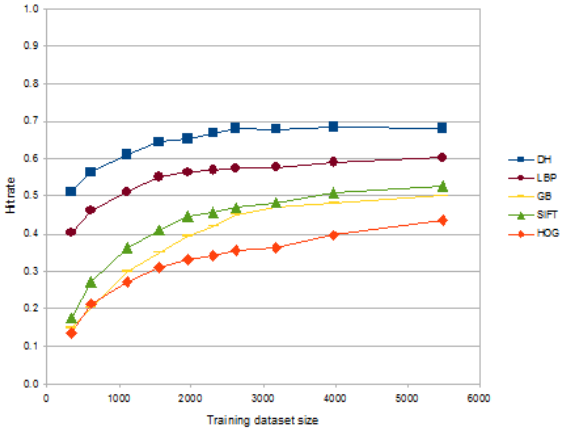

Figure 2. Recognition rate vs Training dataset size ($1^{st} cand.$)

The proposed method has been evaluated on the ICDAR 2003 test dataset, which contains more than 5000 letters in 250 pictures. We compare our approach to the Neumann and Matas' method [9], which was tested with the same dataset. Their method is based on a chain-code bitmap that codes the orientation of the boundary pixels of each binarised object. Table 2 shows the comparison of our method to Neumann's technique. Since Table 1 does not take into account the number of non-detected objects, we have incorporated the non-detection rate in Table 2 in order to make a fair comparison. It can be seen that we get a similar performance to the Neumann's method, even slightly better in terms of hit rate, but we get a really good performance if we take into account the second candidate for this analysis. The mismatched rate for the first two candidates is reduced almost to one third of the mismatched rate with only one candidate and it is much lower than the Neumann's mismatched percentage. Actually, it has been observed that there is a set of pairs and threes of letters that cannot be differentiated between upper-case and lower-case: {'Cc', 'Iil', 'Jj', 'Oo', 'Pp', 'Ss', 'Uu', 'Vv', 'Ww, 'Xx', 'Zz'}. The only way to distinguish these letters in their upper-case and lower-case variants is to use as reference the height of the other unambiguous letters in the same line. In principle, we are just interested in character recognition in a raw way, but if we compute the character recognition rate joining both classes of the undistinguishable letters as only one class for each pair, we get the results shown in Table 3. It can be clearly noticed that the hit rate for the first candidate greatly increases, as it achieves a matched rate higher than 80% and the mismatched rate reduces to 9%.

Table 2. Individual character recognition on ICDAR 2003 dataset.

Algorithm	Matched	Mismatched	Not found
Neumann & Matas [9]	67.0%	12.9%	20.1%
Our method (1st candidate)	68.2%	21.2%	10.6%
Our method (1st/2nd cand.)	81.7%	7.7%	10.6%
Our method (1st/2nd/3rd candidate)	85.4%	4.0%	10.6%

4. Conclusions

A character recognition method based on a simple and fast-to-compute feature has been proposed in this paper. The feature has been baptised as Direction His-

(a) Source image (b) Segmented image

Figure 3. Segmented image.

Table 3. Individual character recognition on ICDAR 2003 dataset, taking indistinguishable pairs of letters as one class for each pair.

Algorithm	Matched	Mismatched	Not found
Our method (1st candidate)	80.4%	9.0%	10.6%
Our method (1st/2nd cand.)	84.1%	5.3%	10.6%
Our method (1st/2nd/3rd candidate)	85.8%	3.6%	10.6%

togram as it consists of histogramming the gradient directions of the contour pixels of a segmented object. It has been compared to other well-known features such as Shape Context or Local Binary Patterns and the results show the robustness of the proposed feature for recognizing characters in complex natural images. In addition, the proposed recognition method does not give only one solution as most systems do. Our approach gives different solutions with output probabilities. Their applications can be various. Typically, a language model and probabilistic methods are applied after the OCR in order to correct the errors made in the character recognition phase. Therefore, those output probabilities could be helpful for this purpose. Another useful application could be for splitting those characters that were not possible to separate in the segmentation step. Fig. 3 shows an example where the objects U and T are treated as only one because they are 8-connected in the binary image. Therefore, initially only 4 objects (R, O, UT, E) have been detected and the output probabilities of the first candidate for each object, identified as 'R', 'O', 'M' and 'E', are $p_1 = 0.97$, $p_2 = 1.0$, $p_3 = 0.42$ and $p_4 = 0.74$. It can be clearly seen that the third object has a lower probability respect to the others. It suggests that something is wrong with it. Therefore, we have developed an algorithm to use this evidence together with others (the width of the object with respect to the others and the existence of minima in the region projection on the horizontal axis), in order to deal with this kind of situations. Actually, with this method we are able to solve the example shown above and the first solution for each object is 'R', 'O', 'U', 'T' and 'E' with output probabilities $p_1 = 0.97$, $p_2 = 1.0$, $p_3 = 0.61$, $p_4 = 1.0$ and $p_5 = 0.74$ respectively.

Acknowledgments

Work funded through the projects ADD-Gaze (TRA2011-29001-C04-01), Ministerio de Economía y Competitividad, and Robocity2030 (CAM-S-0505/DPI000176), Comunidad de Madrid.

References

[1] P. F. Alcantarilla. *Vision Based Localization: From Humanoid Robots to Visually Impaired People*. PhD thesis, University of Alcalá, Alcalá de Henares, Madrid, Spain, October 2011.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

[3] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *CVPR (1)*, pages 26–33, 2005.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[5] A. González, L. M. Bergasa, J. J. Yebes, and S. Bronte. Text location in complex images. In *ICPR*, 2012.

[6] F. Kimura, T. Wakabayashi, S. Tsuruoka, and Y. Miyake. Improvement of handwritten japanese character recognition using weighted direction code histogram. *Pattern Recognition*, 30(8):1329–1337, 1997.

[7] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[8] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young. Robust reading competitions. In *ICDAR*, 2003.

[9] L. Neumann and J. Matas. A method for text localization and recognition in real-world images. In *ACCV*, 2010.

[10] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[11] A. Shahab, F. Shafait, and A. Dengel. ICDAR 2011 Robust Reading Competition. Challenge 2: Reading Text in Scene Images. In *ICDAR*, 2011.