2012 10th IAPR International Workshop on Document Analysis Systems

Combining Multi-Scale Character Recognition and Linguistic Knowledge for Natural Scene Text OCR

Khaoula Elagouni
Orange Labs R&D
Cesson-Sévigné, France
khaoula.elagouni@orange.com

Christophe Garcia
LIRIS, Insa de Lyon
Villeurbanne, France
christophe.garcia@liris.cnrs.fr

Franck Mamalet
Orange Labs R&D
Cesson-Sévigné, France
franck.mamalet@orange.com

Pascale Sébillot
IRISA, Insa de Rennes
Rennes, France
pascale.sebillot@irisa.fr

***Abstract*—Understanding text captured in real-world scenes is a challenging problem in the field of visual pattern recognition and continues to generate a significant interest in the OCR (Optical Character Recognition) community. This paper proposes a novel method to recognize scene texts avoiding the conventional character segmentation step. The idea is to scan the text image with multi-scale windows and apply a robust recognition model, relying on a neural classification approach, to every window in order to recognize valid characters and identify non valid ones. Recognition results are represented as a graph model in order to determine the best sequence of characters. Some linguistic knowledge is also incorporated to remove errors due to recognition confusions. The designed method is evaluated on the ICDAR 2003 database of scene text images and outperforms state-of-the-art approaches.**

***Keywords*-scene text recognition; multi-scale character recognition; convolutional neural networks; language model.**

I. Introduction and Related Work

Texts present in natural scene images generally contain high level semantic information which can be useful for many applications such as image indexing and retrieval, robotic vision and intelligent navigation systems. However, the access to these textual clues involves several complex challenges related to the nature of texts (written, printed or painted on papers, boards, walls, storefronts, etc.) and to their acquisition conditions (low quality, non uniform illumination, shadows, and occlusions). Moreover, characters can be of various colors, fonts, and sizes on very complex and textured backgrounds. These issues make the problem of recognizing text in natural scene images more difficult than document text recognition. (In some cases, even humans can fail reading them.)

Therefore, research in OCR (Optical Character Recognition) systems specifically designed to recognize scene texts has recently received a great interest. A lot of work has been dedicated to the problem of recognizing isolated characters in scene images. Several robust recognition methods [1], [2], [3], able to deal with large numbers of extremely variable characters, have been designed and achieved outstanding performance. However, systems developed to recognize words (*i.e.*, the term word is used loosely here to mean any string of characters in a scene image) still obtain unsatisfactory results. Indeed, the majority of classical approaches proposes to segment the text into characters, generally relying on a pre-processing step that distinguishes text from background [4], (a complete survey of character segmentation methods is presented in [5]) and then recognize extracted characters. However, the different distortions in scene text images make the segmentation very hard, leading to poor recognition results. To improve performance, some authors [6], [7], [8] propose hybrid approaches that combine image processing techniques and recognition. The idea is to build concurrent segmentations and to rely on a character recognizer to identify the correct ones. Further information, namely language properties and lexicons, can also be integrated to improve the performance of OCR schemes [9], [8], [10]. In all these approaches the initial selection of candidate segmentations relies only on image processing which is unreliable in the case of complex scene images.

Other authors [11], [12] propose methods that do not rely on conventional segmentation techniques. Words are determined as sequences of characters identified and localized in text images based on extracted features. However, the major issue of these approaches is to determine the discriminant features to represent extremely variable characters.

In this work, we consider the problem of recognizing words in natural scene images in a different way, avoiding any explicit character segmentation. A multi-scale scanning scheme using windows with non-linear borders, well adapted to the local morphology of the image, is first used to cover the entire text image and several possible sizes of characters. A robust character recognizer relying on a convolutional neural networks, is designed and trained to classify each window and identify characters. A graph model is then built in order to represent spatial constraints between sliding windows and determine the recognized word as the sequence of characters corresponding to the best path in the graph. We also show that linguistic knowledge (namely a language model) can be incorporated within the best path search to improve the performance of the recognition system.

This paper is organized as follows. First, our method to scan the full text image, classify valid / non valid characters and recognize valid ones is detailed in section II. Then, a

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.26

description of the graph model and the integration of the language model are explained in section III. Finally, after presenting experiments and discussing obtained results in section IV, conclusions are drawn in section V.

II. Multi-Scale Character Recognition in Natural Scene Images

In this work, we propose to address the problem of natural scene text recognition without any explicit character segmentation. This section presents our approach to scan the scene text image and classify sliding windows.

A. Multi-Scale Image Scan Scheme

In contrast to the dominant methodology that aims at segmenting the image into separated characters and inspired by works dedicated to handwriting word recognition, our first proposal consists in scanning the full text image with a sliding window at regular and close positions (typically, a step of one eighth of the image height, in our experiments). The purpose is to be sure that at least one window will be aligned with each character. Furthermore, since characters belonging to a same word can be of different sizes depending on their labels (*e.g.*, "Y" and "o" in Fig. 1) and their fonts, text images are scanned at several scales using sliding windows of variable widths. Experiments have shown that good results are obtained with four windows of sizes $\frac{h}{4}$, $\frac{h}{2}$, $\frac{3h}{4}$ and h, where h is the height of the image. A classification step (see subsection II-B) is applied on every window to identify non valid characters and recognize valid ones. Fig. 1 describes this process and shows examples of well framed characters at their corresponding scales (*e.g.*, "Y" and "o" are framed with windows equal to h and $\frac{h}{2}$ respectively).

Figure 1. Multi-scale sliding window scheme: images on the left and on the right are respectively scanned with windows of sizes h and $\frac{h}{2}$ and extracted sub-images correspond to windows placed at positions $\frac{3h}{4}$ and h.

Although the multi-scale scanning scheme helps to cover different scales and positions of characters, vertical borders of windows can extract also parts of their neighbors (*e.g.*, a part of the "o" is present in the first window in Fig. 1) which reduces the performance of window image classification. To overcome this limit and increase the accuracy of the recognition, we propose to adapt the window borders to the local morphology of the image and hence, when possible, clean the neighborhood of characters (see Fig. 2). For each window position and scale, these borders within the full image are computed as follows. (Interested readers can refer to [8] to get more details regarding the computation of non-linear borders.) First assuming that pixels of scene texts are of two classes, "text" and "background", a pre-processing step generates a fuzzy map which encodes for each pixel in the full scene image its membership degree to "text". Using a shortest path algorithm within the obtained map, non-linear vertical borders are afterwards computed following pixels that have a low probability to belong to the class "text". Obtained borders are thus characterized by a score, calculated from their pixel probabilities and encoding their probability to correspond to a right separation between characters. In case of important image distortion or non separated characters, the shortest path algorithm induces straight vertical borders since pixels in the local area have the same probability. Fig. 2 shows new obtained windows with their non-linear borders and illustrates an example of a window with straight vertical borders (the rightmost one).

Figure 2. Multi-scale sliding windows with non-linear borders: from the left to the right, window widths are respectively h, $\frac{h}{2}$, $\frac{3h}{4}$ and $\frac{h}{2}$.

B. Isolated Window Classification

Every window at each position and scale has then to be classified as a character or not. In this context Convolutional Neural Networks (hereafter ConvNets) [13] have shown to be well adapted to our recognition problem [3], [14]. A ConvNet is a bio-inspired hierarchical multi-layered neural network able to learn visual patterns directly from the image pixels without any pre-processing step. Relying on specific properties (local receptive fields, weight sharing and sub-sampling), this neural model is particularly robust to noise, geometric transformations and distortions and has proved a great ability to deal with a large number of extremely variable patterns. This model was tested on several pattern recognition tasks ranging from handwritten characters recognition [14] to face detection [15] and generally outperformed other classification models such as SVM (Support Vector Machine) models [8].

In our application, several network architectures were first tested to classify windows as "valid character" or "garbage" (*i.e.*, window misaligned with a character, part of a character or interstice between characters). The best configuration, hereafter WCConvNet for Window Classifier ConvNet, takes as input a color window image mapped into a 48×48 input map, containing values normalized between -1 et 1, and returns a value encoding the probability of the window to correspond to a valid character. Windows identified as not containing a character are labeled as "garbage", while others

are presented to a second ConvNet, hereafter CRConvNet for Character Recognizer ConvNet, whose task is to recognize the character present in the window. CRConvNet takes as input the normalized color window image and returns a vector of values encoding probabilities of the image to belong to each character class (36 classes are considered: 26 Latin letters and 10 Arabic numbers). The architectures of WCConvNet and CRConvNet are similar to the one presented in [8]. As shown in Fig. 3, each well framed character (such as "o") is identified as valid and recognized, while interstices between characters are identified as "garbage". Nevertheless, some parts of characters can introduce recognition confusions (*e.g.*, the part of "u" recognized as a "i"). In the next section, we detail the proposed method to remove these ambiguities and reduce recognition errors.

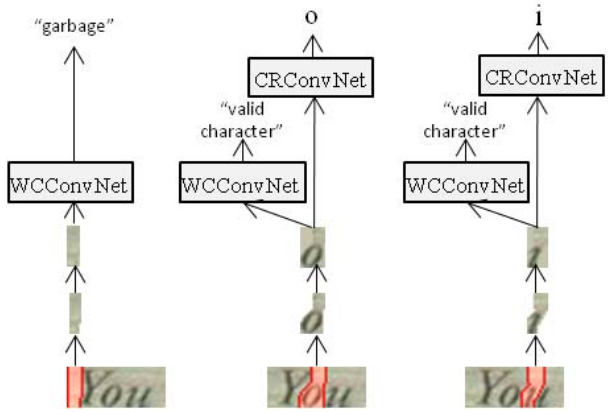

Figure 3. The classification of extracted windows when using a sliding window of width equal to $\frac{h}{2}$: from the left to the right, images correspond to windows placed at positions: $\frac{h}{4}$, $\frac{10h}{8}$, and $\frac{17h}{8}$.

III. Scene Text Recognition using a Graph Model and Linguistic Knowledge

Multi-scale window classification results can now be analyzed to recognize the full text present in a scene image. Since many windows at different scales are overlapping, a graph model is first built to represent spatial constraint between sliding windows. A best path search algorithm is then applied to determine the best sequence of characters. Some linguistic knowledge is also incorporated into the graph to remove recognition ambiguities.

A. Graph Model Construction

Aiming at determining the most probable sequence of characters that corresponds to a scene text, the different multi-scale windows throughout the image are first represented by means of a directed acyclic graph model (*cf.* Fig. 4). The vertices of the graph are the borders of the windows and thus encode spatial constraints between them. The first (resp. last) vertex corresponds to the left (resp. right) border

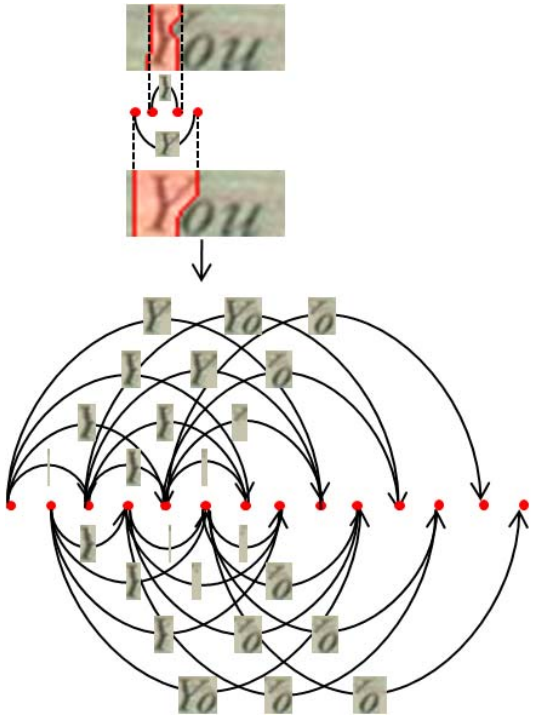

Figure 4. Graph model construction scheme: the upper figure illustrates the representation of window borders by vertices (*e.i.*, red dots) and the bottom one shows a part of the obtained graph.

of the first (resp. last) window position. Each directed edge, called arc, joining two vertices represents a window image. Considering the fact that four different window scales are used to scan the text image, each vertex v is thus connected to 4 successor vertices (*e.i.*, the right borders of the four different windows starting from v).

Since non-linear borders of windows are characterized by scores encoding their probabilities to correspond to right separations between characters (see subsection II-A), a weight, corresponding to the non-linear border score, is assigned to each vertex. Furthermore, the results of classification related to each window, namely the output of WCConvNet for non valid characters and the output vectors of CRConvNet for valid ones, are assigned to each arc. Fig. 4 shows one part of the graph built on a sample image representing each possible window. The classical Viterbi algorithm can then be applied on the graph to determine the most probable sequence of characters avoiding arcs that correspond to non valid ones.

B. Language Model Integration

As shown in subsection II-B, the recognizer, when applied at some scales and positions, can generate confusions between characters and thus introduce errors. To remove these ambiguities, we propose to incorporate some linguistic knowledge, able to take into consideration the lexical

context. In a previous work [8], a trigram model has been shown to be a probabilistic language model well adapted to the problem of word recognition. It allows to estimate the probability that a sequence of characters is observed in a given language assuming that a character depends only on its 2 predecessors.

In our application, a trigram model is first trained to learn joint probabilities of character sequences using a corpus of $11,800$ English words. Obtained probabilities are then integrated into the graph model in order to adjust the transition score between arcs according to the trigram model and predict the next character to be recognized in a sequence of characters. The optimal word is finally computed using the Viterbi algorithm which takes into account positions and scales of windows, their recognition results, and some language properties provided by the language model.

Next section evaluates the individual contribution of each step of our recognition scheme.

IV. Experimental Set up and Results

This section presents the evaluation of our OCR scheme and discusses obtained results.

A. Classifiers training and performance

In order to train the window classifier and the character recognizer networks, we used the training sets of the databases of isolated characters and scene text images ICDAR 2003[1]. The first set consists of $5,689$ images of isolated characters while the second set consists of $1,156$ text images that we used to generate $4,056$ images of non valid characters. 90% of these isolated character and generated interstice (*e.g.* non valid characters) images are used to train CRConvNet and WCConvNet with a classic online stochastic backpropagation algorithm. Table I provides recognition results obtained on the 10% remaining images.

Table I
Performance of window classification and character recognition

Network	Classification rate
CRConvNet	85.13%
WCConvNet	79.23%

B. Proposed OCR scheme evaluation

In order to compare our method to the state-of-the-art, our experiments have been carried out on the test set of the database of scene texts ICDAR 2003, initially created for a competition on scene word recognition [16]. The test set consists of $1,110$ scene text images which are of different sizes, present several kinds of distortions (non uniform illumination, occlusions, blur, etc.) and contain characters printed and painted in various fonts and colors (*cf.* Fig. 5).

[1]The database of scene texts ICDAR 2003 is available for download at http://algoval.essex.ac.uk/icdar/Datasets.html#Robust%20Word%20Recognition

Figure 5. Examples of scene text images from the ICDAR 2003 database.

Using the trained ConvNets, we evaluated the proposed OCR scheme on the test set. Fig. 6 presents an example of recognized word and illustrates the resulting best path within the graph model. Table II shows that our method achieves outstanding results of 70% of character recognition rate corresponding to about 47% of words correctly recognized. It also highlights the contribution of each processing step in the whole OCR scheme. These results confirm that multi-scale scans (MSS) are mandatory to cover different sizes of characters and to get satisfactory results (in the absence of this processing, the performance of the OCR notably decreases to 0.12% of word recognition rate). Non-linear borders (NLB) and the language model (LM) also result in an important improvement of the character recognition rate (the incorporation of each one of these processings increases the word recognition rate from about 26% to 47%).

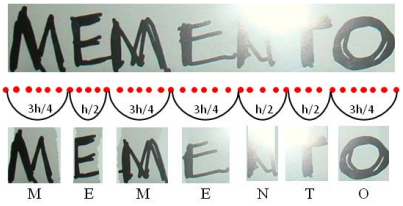

Figure 6. Example of recognized natural scene text.

Table II
Contributions of different processing steps incorporated in the proposed OCR scheme: RR stand for Recognition Rate.

Method	Character RR	Word RR
Complete scheme (MSS+NLB+LM)	**70.33%**	**46.72%**
MSS+NLB	54.32%	25.54%
MSS+LM	53.41%	27.18%
NLB+LM	9.12%	0.12%

Our character recognition results were also compared with those of three state-of-the-art methods [7], [8], [12] (*cf.* table III). In order to provide a meaningful comparison, three experiments were performed as described in the other works. Besides the experimentation of the full test set (Exp1), we

evaluate our OCR scheme on the 901 images selected in [7] (Exp2) and on the $1,065$ images selected in [12] correcting OCR output to be the word with the smallest edit distance in a dictionary built out of the ground-truth words which can be considered as a task of word spotting as presented in [12] (Exp3). These different tests show that our approach yields the best word accuracy. Our method also outperforms Wang et al.'s one [12] by about 7% even though their method uses hand-designed features to recognize characters. Our OCR scheme is finally compared to two commercial OCR engines, namely ABBYY FineReader and Tesseract, evaluated by Wang et al. [12]. Table III shows obtained results where our method achieves far better performance.

Table III
COMPARISON OF THE PROPOSED SCHEME TO STATE-OF-THE-ART AND COMMERCIAL OCR ENGINES

Method	**Word recognition rate**		
	Exp1	**Exp2**	**Exp3**
Proposed method	**46.72%**	**57.04%**	**66.19%**
Elagouni et al. [8]	34.81%	-	-
Saidane et al. [7]	-	54.13%	-
Wang et al. [12]	-	-	59.20%
ABBYY engine	-	-	42.80%
Tesseract engine	-	-	35.00%

V. CONCLUSION

In this paper, we have presented a novel approach for natural scene text recognition. Using a multi-scale scan scheme with non-linear borders, we have shown that the traditional character segmentation step can be avoided, by covering with the help of sliding windows the characters of a word at different positions and scales. These windows are classified by a robust automatic learning model, relying on a neural approach, that recognizes valid characters and identifies non valid ones. A graph model is proposed to represent spatial constraints between windows. A Viterbi algorithm on this graph enables to extract the most probable sequence of characters present in a scene image. We also demonstrated that the performance of our system can be improved by integrating some linguistic knowledge that takes into account the lexical context. The proposed scheme was evaluated on the database of scene texts ICDAR 2003 and compared to state-of-the-art methods and commercial OCR engines. Our method obtains outstanding performance (over than 70% of character recognition rate) and yields the best word recognition rate among concurrent approaches.

Even though obtained results are still not sufficient for practical applications, the proposed approach can be considered as a new promising direction in the field of OCR systems dedicated to scene text recognition.

REFERENCES

[1] M. Yokobayashi and T. Wakahara, "Segmentation and recognition of characters in scene images using selective binarization in color space and gat correlation," in *Proc. ICDAR*, 2005, pp. 167–171.

[2] K. Negishi, M. Iwamura, S. Omachi, and H. Aso, "Isolated character recognition by searching features in scene images," in *Proc. CBDAR*, 2005, pp. 140–147.

[3] Z. Saidane and C. Garcia, "Automatic scene text recognition using a convolutional neural network," in *Proc. CBDAR*, 2007, pp. 100–106.

[4] S. Lee, J. Seok, K. Min, and J. Kim, "Scene text extraction using image intensity and color information," in *Proc. CCPR*, 2009, pp. 1–5.

[5] R. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," *PAMI*, vol. 18, no. 7, pp. 690–706, 1996.

[6] C. Mancas-Thillou and B. Gosselin, "Character segmentation-by-recognition using log-gabor filters," in *Proc. ICPR*, vol. 2, 2006, pp. 901–904.

[7] Z. Saidane, C. Garcia, and J. Dugelay, "The image text recognition graph (iTRG)," in *Proc. ICME*, 2009, pp. 266–269.

[8] K. Elagouni, C. Garcia, and P. Sébillot, "A comprehensive neural-based approach for text recognition in videos using natural language processing," in *Proc. ICMR*, 2011.

[9] J. Weinman, E. Learned-Miller, and A. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *PAMI*, vol. 31, no. 10, pp. 1733–1746, 2009.

[10] T. Yamazoe, M. Etoh, T. Yoshimura, and K. Tsujino, "Hypothesis preservation approach to scene text recognition with weighted finite-state transducer," in *Proc. ICDAR*, 2011, pp. 359–363.

[11] Y. Kusachi, A. Suzuki, N. Ito, and K. Arakawa, "Kanji recognition in scene images without detection of text fields-robust against variation of viewpoint, contrast, and background texture," in *Proc. ICPR*, vol. 1, 2004, pp. 457–460.

[12] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. ECCV*, 2010, pp. 591–604.

[13] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, pp. 255–258, 1995.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[15] C. Garcia and M. Delakis, "Convolutional face finder: A neural architecture for fast and robust face detection," *PAMI*, vol. 26, no. 11, pp. 1408–1423, 2004.

[16] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "Icdar 2003 robust reading competitions," *IJDAR*, vol. 2, pp. 682–687, 2003.