OCR with Adaptive Dictionary

Chenyang Wang[1], Yanhong Xie[2], Kai Wang[1](✉), and Tao Li[1]

[1] College of Computer and Control Engineering, Nankai University, Tianjin, China
wangk@nankai.edu.cn
[2] Shenzhouhaotian Technology Co., Ltd., Tianjin, China

Abstract. It has been proven by previous works that OCR is beneficial from reducing dictionary size. In this paper, a framework is proposed for improving OCR performance with the adaptive dictionary, in which text categorization is utilized to construct dictionaries using web data and identify the category of the imaged documents. To facilitate comparison with other existing methods that focus on language identification, an implementation is presented to improve the OCR performance with language adaptive dictionaries. Experimental results demonstrate that the performance of OCR system is significantly improved by the reduced dictionary. Compared with other existing methods for language identification, the proposed method shows a better performance. Also, any other categorization methodology is expected to further reduce the dictionary size. For example, an imaged document with specific language can be further categorized into sport, law, entertainment, etc. by its content.

Keywords: Dictionary size · Text categorization · Language identification · OCR

1 Introduction

Dictionary driven OCR is popular for text recognition from both document image [1] and imagery [2]. In previous works, it has been proven that OCR is beneficial from reducing dictionary size, e.g. about 40 % of fail-to-recognized words are corrected in [3,4] with a reduced dictionary. Manual or automated categorization of images into predefined categories is a natural way to reduce the size of the dictionary by the text embedded in images. For example, an image can be categorized into English, France, German, etc. with the different language, and it also can be categorized into sport, law, entertainment, etc. with the different content. A reduced dictionary with the specified category is expected for an improving OCR performance.

This work is supported by the National Natural Science Foundation of China under Grant No. 61201424, 61301238, 61212005, the Fundamental Research Funds of Tianjin, China under Grant No. 14JCTPJC00501, 14JCTPJC00556, and the Natural Science Foundation of Tianjin, China under Grant No.12JCYBJC10100, 14ZCDZGX00831.

Y.-J. Zhang (Ed.): ICIG 2015, Part II, LNCS 9218, pp. 611–620, 2015.
DOI: 10.1007/978-3-319-21963-9_56

As an application of text categorization [5], automated script and language identification has been studied in past decades for improving the performance of OCR. A. L. Spitz [6] first classifies the script of a document image as being either Han- or Latin-based with the vertical position distribution of upward concavities. Then language identification for Han images is conducted by optical density distribution. And language identification for Latin images is conducted by the most frequently occurring word shapes characteristic proposed in [7]. Rotation invariant texture features is used by T. N. Tan [8] for automated script identification, whose effectiveness is verified by the experiments on six languages that include Chinese, English, Greek, Russian, Persian, and Malayalam. A novel word shape coding scheme is proposed by S. Lu et al. [9] for language identification, in which a document vector is constructed by using the high-frequency word shape codes and then it is used to identify the language of the imaged document. The scheme is extended for the identification of both script and language in [10]. A comprehensive survey is presented by D. Ghosh et al. [11] on the developments of script identification. I. H. Jang et al. [12] propose a texture feature based method that combines Gabor and MDLC features for script identification. The Orientation of the Local Binary Patterns (OLBP) is proposed by M. A. Ferrer et al. [13] for line-wise script identification. Previous works on script and language identification can be categorized into texture-based such as [8,12,13] and shape coding-based such as [6,7,9,10]. Texture-based methods are used for script identification, in which the texture features are extracted from the text patches and a classifier is applied to identify the script of the imaged documents. Shape coding-based methods are used for both script identification and language identification, in which word shapes are coded and the statistics of word shapes are utilized to identify the script or the language of the imaged documents.

Previous works have only focused on script and language identification. The methods are difficult to be extended for the imaged document categorization by other categorization methodology. And the performance of OCR with the reduced dictionaries is not concerned. In this paper, a framework is proposed for improving OCR performance with the adaptive dictionary, in which text categorization is utilized to construct dictionaries using web data and identify the category of the imaged documents. The proposed framework is suitable for the reduction of dictionary size by any categorization methodology. To facilitate comparison with other existing methods that focus on language identification, an implementation is presented to improve the OCR performance with language adaptive dictionaries. Experimental results demonstrate that the performance of OCR system is significantly improved by the reduced dictionary. Compared with other existing methods for language identification, the proposed method shows a better performance.

The rest of the paper is organized as follows. The framework is proposed in Sect. 2 for OCR with adaptive dictionary. An implementation is presented in Sect. 3 to improve the OCR performance with language adaptive dictionaries. Experiments are conducted in Sect. 4 to verify our work. Summary of the paper are shown in Sect. 5.

2 Framework

The proposed framework is shown in Fig. 1, which consists of dictionary learning and OCR.

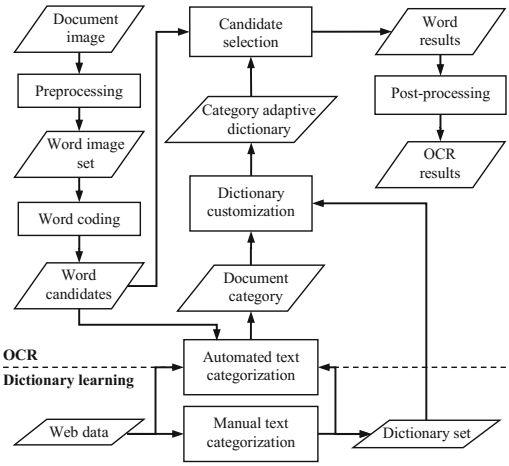


Fig. 1. The proposed framework

2.1 Dictionary Learning

In dictionary learning module, the data scratched from web is categorized by manual or automated text categorization. Then each dictionary in the dictionary set is initialized by manual categorized data and updated by automated categorized data. The real-time web data makes the proposed framework be adaptive for new words and new phrases.

2.2 OCR

In OCR module, preprocessing, word coding, automated text categorization, dictionary customization, candidate selection and post-processing are sequentially conducted to obtain the OCR results, which are briefly introduced as follows.

Preprocessing. Word image set is generated by analyzing the document image. The method can be top-down or bottom-up. Also, image processing such as binarization, noise removal, deskewing and auto-orientation should be applied on the document image for better OCR results.

Word Coding. Each word is coded by its shape. And then all candidate results for the word are generated, which always consist of some wrong results and no or a correct result.

Automated Text Categorization. All candidates are used for categorizing the imaged document to a specific category. To reduce the negative influence of noises for categorization, some candidates should be dropped if they are not matched with the words or the phrases in the dictionaries.

Dictionary Customization. The dictionary that corresponds to the detected document category is picked up from the dictionary set for the subsequent processing.

Candidate Selection. Each candidate is compared with the words in the category adaptive dictionary. And the mis-matched candidates are dropped. The phrases matching should be applied to select the best candidate if the candidate remained for one word is not unique.

Post-processing. The document layout is restored and the processing results are exported with user-specified format.

Compared with previous works, the advantages of the proposed framework are as follows.

- The proposed framework is suitable for the reduction of dictionary size by any categorization methodology instead of just by language.
- The real-time web data makes the proposed framework be adaptive for new words and new phrases.

3 Implementation

To facilitate comparison with other existing methods that focus on language identification, an implementation is presented to improve the OCR performance with language adaptive dictionaries. Only the methods for word coding and automated text categorization are introduced here. Other steps are just the same with that in a general OCR framework, thus they are ignored in this paper.

3.1 Word Coding

Simple shape features have been used for word coding in [6,7,9,10], and the proposed methods have been verified by experiments to be effective for script and language identification. Particularly, the method proposed in [9,10] is free of character segmentation. However, a high repetition coding rate always exists in the methods. That is, numerous candidate results are often generated for a word image, so that the negative influence of noises is inevitable for text categorization. To reduce the noises, only high-frequency words are used and only script/language identification is considered in previous works. It is difficult to adapt the methods for text categorization instead of just for language identification.

Character segmentation has been well studied in past decades [14]. It is not difficult to find the possible segmentation positions, but it is difficult to determine the correct segmentation positions. Sliding window based segmentation method is often used for the character recognition in imagery [2], which is always time

consuming due to multi-scale scanning. In this paper, an individual characters based method is used for word coding, in which all possible segmentation positions are considered to find all possible candidates. To be efficient, contour based segmentation [15] is applied to obtain the over-segmentation results. And then each possibly correct segmentation or the combination of segmentations is represented by a vector that consists of the optical density [6] of n*n cells. Finally, the vector is matched with each pre-defined template. A character may be the result that the vector corresponds to as long as the match distance is small enough. As a result, a word image always corresponds to multiple candidates by different segmentations combination and different matched characters. For example, it is possible that a word "me" is coded as (m|rn)(e|c). That is, the candidates of the word include "me", "rne", "mc" and "rnc". Compared with the character recognition in OCR, the coding method used in this paper is simple and more candidate results are generated.

3.2 Automated Text Categorization

To reduce the negative influence of noisy candidates for text categorization, the candidates are dropped if they are not matched with any word in the dictionaries. For example, "rne", "mc" and "rnc" should be removed from the candidates of the word "me" as they are not correct spelled words. The same idea is also used in [6,7,9,10], and only relative high frequency codes are remained for script and language identification.

Correct spelled words are used for subsequent text categorization. In this paper, the number of words is counted for each specific category. And the category with the most words is set as the text categorization result. The method used here is very simple, and it is expected to improve the performance of text categorization by using a more complex one such as the methods in [5].

4 Experiments

4.1 Dataset

As is shown in Table 1, 12807 document images with eleven languages are collected for the verification of the proposed method, which are captured by scanning from magazines, books and printed documents. Some images used in experiments are given in Fig. 2(a).

To be compared with the methods in [9,10], the original images are corrupted by Gaussian noise (mean = 0, variance = 0.02), salt & pepper noise (noise ratio = 0.05) and low scanning resolution (150 dpi), respectively. The degraded images that correspond to Fig. 2(a) are given in Fig. 2(b) and (c). The images with low resolution look similar with the original ones, thus they are ignored here.

Table 1. An overview on document images used in experiments

Language	Abbreviation	The number of images
English	EN	2,500
French	FR	963
German	DE	1,151
Italian	IT	1,112
Swedish	SW	1,082
Spanish	ES	1,096
Portuguese	PT	1,400
Norwegian	NO	659
Dutch	NL	844
Polish	PL	1,000
Finnish	FI	1,000
Total		12,807

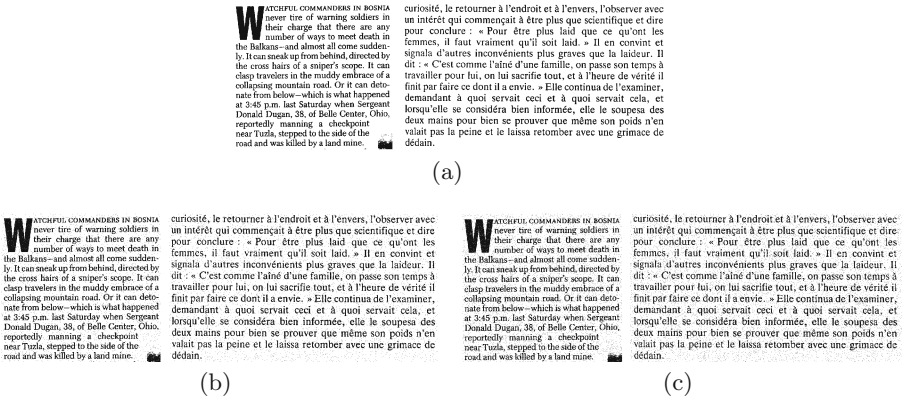


Fig. 2. Some images used in experiments. (a) Original images. (b) Images degraded by Gaussian noise (mean = 0, variance = 0.02). (c) Images degraded by salt & pepper noise (noise ratio = 0.05).

4.2 OCR Performance

To verify the performance of OCR with the reduced dictionary, two OCR systems are implemented, in which full dictionary and adaptive dictionary are respectively used for candidate selection. As is shown in Table 2, performance of OCR with full dictionary and that with adaptive dictionary are compared on the images with eleven languages. The total error rate is reduced from 1.96 % to 0.77 % by replacing the full dictionary with the adaptive dictionary, 60.91 % of mis-recognized characters being corrected.

Table 2. Performance of ocr with full dictionary and that with adaptive dictionary

Language	The number of total characters	OCR with full dictionary		OCR with adaptive dictionary		Reduction rate of errors by replacing full dictionary with adaptive dictionary (%)
		The number of mis-recognized characters	Error rate (%)	The number of mis-recognized characters	Error rate (%)	
English	4,404,992	78,148	1.77	67,695	1.54	13.38
French	3,715,867	94,334	2.54	61,284	1.65	35.04
German	3,848,309	64,128	1.67	33,092	0.86	48.40
Swedish	3,598,688	125,077	3.48	7,534	0.21	93.98
Spanish	3,877,759	46,649	1.20	31,308	0.81	32.89
Portuguese	4,380,493	85,771	1.96	11,872	0.27	86.16
Norwegian	1,963,128	50,168	2.56	5,990	0.31	88.06
Dutch	2,402,851	7,062	0.29	4,008	0.17	43.25
Polish	2,367,747	17,235	0.73	14,055	0.59	18.45
Finnish	2,236,520	96,879	4.33	9,340	0.42	90.36
Total	36,599,857	717,860	1.96	280,597	0.77	60.91

Table 3. Accuracy comparison for language identification (%)

Method	Salt and pepper noise	Gaussian noise	Low resolution	Three degradation combined
Spitz's [10]	83.53	86.90	78.77	72.22
Nobile's [10]	83.77	86.28	78.36	71.67
Suen's [10]	82.91	86.42	79.13	71.84
Lu's [10]	96.32	96.47	91.18	88.31
Ours	98.85	99.00	97.39	96.31

4.3 Language Identification

As is shown in Table 3 the proposed method outperforms other methods for language identification.

The detailed results between languages are given in Table 4, 5, 6 and 7. There are mainly two types of errors for language identification on images with salt and pepper noise or Gaussian noise. One is Portuguese being mis-identified as French, and another is English being mis-identified as Finnish. The samples with mis-identified language are analyzed and the mis-identified reasons are as follows.

Table 4. Identification results between languages on images with salt and pepper noise

True language	Detected language											Error
	EN	FR	DE	IT	SW	ES	PT	NO	NL	PL	FI	
EN	2433										67	67
FR	4	959										4
DE	3		1148									3
IT	15	1	7	1089								23
SW					1082							0
ES		1				1095						1
PT		45					1355					45
NO								659				0
NL									844			0
PL	2									998		2
FI	2										998	2

Table 5. Identification results between languages on images with gaussian noise

True language	Detected language											Error
	EN	FR	DE	IT	SW	ES	PT	NO	NL	PL	FI	
EN	2441										59	59
FR	2	961										2
DE			1151									0
IT	13	1	7	1091								21
SW					1082							0
ES		1				1095						1
PT		45					1355					45
NO								659				0
NL									844			0
PL										998		0
FI											998	0

– Some samples marked as Portuguese are the multi-lingual documents, e.g. the first two text lines in Fig. 3(a) are detected as Portuguese by Google Translate, while the last three text lines are detected as French.
– The text fonts in some English samples are special, e.g. Fig. 3(b).

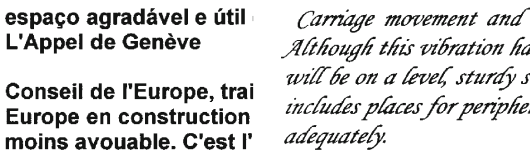
More errors are generated when the images are corrupted by low scanning resolution, the results being consistent with previous methods. To ensure the performance, an image resolution of at least 300DPI is suggested.

Table 6. Identification results between languages on images with low resolution

True language	Detected language											Error
	EN	FR	DE	IT	SW	ES	PT	NO	NL	PL	FI	
EN	2328		1			1					170	172
FR	2	961										2
DE	2		1149									2
IT	16	1	7	1088								24
SW					1082							0
ES	2	1				1093						3
PT		45					1355					45
NO								659				0
NL									844			0
PL	52									948		52
FI	33						1				966	34

Table 7. Identification results between languages on images with three degradation combined

True language	Detected language											Error
	EN	FR	DE	IT	SW	ES	PT	NO	NL	PL	FI	
EN	2255	2	1	1		1						245
FR	8	955										8
DE	2		1149									2
IT	25	1	7	1069								33
SW					1082							0
ES	3	1				1092						4
PT		45					1355					45
NO								659				0
NL									844			0
PL	79									921		79
FI	56										944	56



(a) (b)

Fig. 3. Samples with mis-identified language. (a) Portuguese is mis-identified as French. (b) English is mis-identified as Finnish.

5 Conclusion

A framework is proposed for improving OCR performance with the adaptive dictionary, which is suitable for the reduction of dictionary size by any categorization methodology instead of just by language. And the real-time web data makes the proposed framework be adaptive for new words and new phrases. One of our future works is to improve the OCR performance with the further reduced dictionary by other categorization methodology instead of just by language.

References

1. Nagy, G.: Twenty years of document image analysis in PAMI. IEEE Trans. Pattern Anal. Mach. Intell. **22**(1), 38–62 (2000)
2. Ye, Q., Doermann, D.: Text detection and recognition in imagery: a survey. IEEE Trans. Pattern Anal. Mach. Intell. **37**, 1480–1500 (2015)
3. Yao, C., Bai, X., Shi, B., liu, W.: Strokelets: a learned multi-scale representation for scene text recognition. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4042–4049 (2014)
4. Shi, C., Wang, C., Xiao, B., Zhang, Y., Gao, S., Zhang, Z.: Scene text recognition using part-based tree-structured character detection. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2961–2968 (2013)
5. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)
6. Spitz, A.L.: Determination of the script and language content of document images. IEEE Trans. Pattern Anal. Mach. Intell. **19**(3), 235–245 (1997)
7. Nakayama, T., Spitz, A.L.: European language determination from image. In: IEEE Conference on Document Analysis and Recognition (ICDAR), pp. 159–162 (1993)
8. Tan, T.N.: Rotation invariant texture features and their use in automatic script identification. IEEE Trans. Pattern Anal. Mach. Intell. **20**(7), 751–756 (1998)
9. Lu, S., Li, L., Tan, C.L.: Identification of Latin-based languages through character stroke categorization. In: 2007 International Conference on Document Analysis and Recognition (ICDAR), pp. 352–356 (2007)
10. Lu, S., Tan, C.L.: Script and language identification in noisy and degraded document images. IEEE Trans. Pattern Anal. Mach. Intell. **30**(1), 14–24 (2008)
11. Ghosh, D., Dube, T., Shivaprasad, A.P.: Script recognition review. IEEE Trans. Pattern Anal. Mach. Intell. **32**(12), 2142–2161 (2010)
12. Jang, I.H., Kim, N.C., Park, M.H.: Texture feature-based language identification using Gabor and MDLC features. In: 2011 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2011)
13. Ferrer, M.A., Morales, A., Pal, U.: LBP based line-wise script identification. In: 2013 IEEE Conference on Document Analysis and Recognition, pp. 369–373 (2013)
14. Casey, R.G., Lecolinet, E.: A survey of methods and strategies in character segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **18**(7), 690–706 (1996)
15. Kurniawan, F., Mohamad, D.: Performance comparison between contour-based and enhanced heuristic-based for character segmentation. In: 2009 Fifth International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), pp. 112–117 (2009)