Machine Printed and Handwritten Text Discrimination in Korean Document Images

Son Tung Trieu*, Guee Sang Lee**

Abstract

Nowadays, there are a lot of Korean documents, which often need to be identified in one of printed or handwritten text. Early methods for the identification use structural features, which can be simple and easy to apply to text of a specific font, but its performance depends on the font type and characteristics of the text. Recently, the bag-of-words model has been used for the identification, which can be invariant to changes in font size, distortions or modifications to the text. The method based on bag-of-words model includes three steps: word segmentation using connected component grouping, feature extraction, and finally classification using SVM(Support Vector Machine). In this paper, bag-of-words model based method is proposed using SURF(Speeded Up Robust Feature) for the identification of machine printed and handwritten text in Korean documents. The experiment shows that the proposed method outperforms methods based on structural features.

Keywords : recognition|handwritten|machine printed|bag-of-words model|document analysis

I. INTRODUCTION

Handwritten character recognition is a process of transforming handwritten text into machine excutable format. There are two types of recognition: on-line and off-line methods [1]. In the on-line system, two dimensional coordinates of successive points are represented as a function of time and the order of strokes made by the writer are also available.

Off-line handwriting recognition is an automatic conversion of text into an image into letter codes which are usable within computer and text-processing applications. Off-line recognition is more challenging because the shape of characters, large variation of character symbols, difference in handwriting style and document quality. In addition, a lot of applications including bank processing, mail sorting, postal address recognition require offline handwriting classification system. In that case, off-line handwriting recognition continues to be an active researching are towards exploring the newer techniques that would improve the classification accuracy.

Normally, the basic steps in off-line recognition system are word segmentation, feature extraction and classification. In the first step, the purpose is to detect blocks of interest in the document image. For feature extraction stage, the objective is to capture the important characteristics of the symbols. This is the most essential step of the recognition process. And the last stage, classification, a machine learning system decides which type of text in the word block based on its descriptor.

In the stage of feature extraction, mostly the structure of word blocks has been discussed recently. There are a few published research results in this issue. Y. Zheng and H. Li [2] extracted features based on overlap area of characters inside each block of word, the runlength histogram [3]. Also, L. F. Silva and A. Sanchez [4] calculated the density ratio, the mean and deviation of the width, height and area of each word block.

* This research was supported by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by MEST(NRF-2015R1D1A1A01060172).

* Student, Dept. of ECE, Chonnam National University
** Member, Dept. of ECE, Chonnam National University

Manuscript : 2016. 07. 28
Revised : 2016. 09. 28
Confirmation of Publication: 2016. 09. 29

Corresponding Author :Guee Sang Lee
e-mail : gslee@jnu.ac.kr

However, the performance of extracting structural features depends on the result of the word segmentation steps, especially the binarisation technique, and the word blocks which are used for extracting features must be binarised. In our methods, a technique based on Bag-of-Visual-Words (BoVW) model is presented. The feature extracted is Speeded Up Robust Features (SURF) which is inspired by the Scale-Invariant Feature Transfom (SIFT) descriptor [5]. Both SURF and SIFT features have been proved to be useful because of their invariance to scale, rotation of image, affine deformation, change of viewpoint, illumination changes and robustness. SIFT is the most accurate feature detector and descriptor. However, using a 128-element dimension of keypoint descriptor makes SIFT relatively slow to compute and match. On the other hand, SURF utilizes the local gradient histograms and integral images which speeds up the computation.

The rest of this paper is organized into four sections. Section II presents the comparison between SIFT and SURF features. Next, section III gives our proposed methods. Experimental results are provided in section IV. Finally, conclusions will be presented in Section V.

II. SIFT vs. SURF features

SIFT (Scale Invariant Feature Transform) algorithm proposed by Loew in 2004 [5] to solve the rotation of image, scaling, and affine deformation, change of viewpoint, noise, illumination changes, and also has robustness. The SIFT algorithm has four main stages: scale-space extrema detection, keypoint localization, orientation assignment, and finally, description generation. The first stage is to find the location and scales of keypoints using scale space extrema in the DoG (Difference-of-Gaussian) functions with different values of σ, the DoG function is convolved with the image in scale space partitioned by a constant factor k:

$$D(x,y,\sigma)=(G(x,y,k\sigma)-G(x,y,\sigma))\times I(x,y) \quad (1)$$

where, G is the Gaussian function and I is the image.

The Gaussian images are subtracted to produce a DoG, after that the Gaussian image is subsampled by factor 2 and applies the DoG function for sampled image. A pixel is compared with a 3x3 neighborhood to detect the local maxima and minima of $D(x,y,\sigma)$.

In the next stage, keypoint candidates are localized and refined by eliminating the keypoints where they discard the low-contrast points. In the orientation assignment stage, the collection of orientation of key point is based on local image gradient.

Finally, the description generation stage. The purpose of this stage is to compute the local image descriptor for each keypoint based on image gradient magnitude and orientation at each image sample point in a region centered at the keypoint [6]; these samples build a 3D histogram of gradient location and orientation which have 128-element dimension of keypoint descriptor with 4x4 array location grid and 8 orientation bins in each sample.

Fig. 1 shows the computation of the keypoint descriptor. First, the image gradient magnitudes and orientations are sampled around the location of the keypoint, and use the scale of the keypoint to select the level of Gaussian blur for the image. In order to gain the invariance of orientation, which is descriptor coordination, the gradient orientations are rotated relatively to the keypoint orientation.

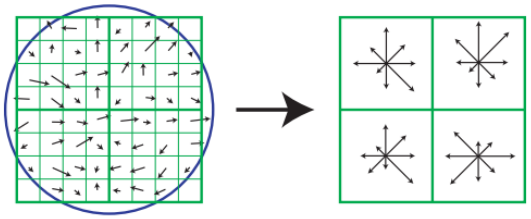

Fig. 1 SIFT descriptor generation. From left to right: Image gradients, and keypoint descriptor[5]

The keypoint descriptor is shown on the right side of Fig. 1. It allows significant shifts in gradient positions by creating the histogram of orientation over 4x4 sample regions. In Fig. 1, 8 directions for each orientation histogram are used with the length of each arrow corresponding to the magnitude of the histogram entry. A gradient sample on the left can both shift upto 4 sample positions and also contribute to the same histogram on the right. Therefore, 4x4 array location grid and 8 orientation bins in each sample which is 128-element dimension of keypoint descriptor are used.

SURF, however, is based on the theory of multi-scale space and the feature detector is determined by using Hessian matrix [7]. Since Hessian matrix has good performance and accuracy, given a point A(x,y) in the image I, the Hessian matrix $H(A,\sigma)$ at scale σ can be defined as

$$H(A,\sigma)=\begin{bmatrix} L_{xx}(A,\sigma) & L_{xy}(A,\sigma) \\ L_{yx}(A,\sigma) & L_{yy}(A,\sigma) \end{bmatrix} \quad (2)$$

where $L_{xx}(A,\sigma)$ is the convolution result of the second order derivative of Gaussian filter $\frac{\partial^2}{\partial x^2}g(\sigma)$ with in point A of the image I, similarly to $L_{xy}(A,\sigma)$ and $L_{yy}(A,\sigma)$.

SURF generates a stack without 2:1 down sampling for higher levels in the pyramid which results in images having the same resolution. Because of the use of integral images, SURF filters the stack using a box filter approximation of second-order Gaussian partial derivatives [8]. The convolution with box filters can be processed in parallel for different scales. Fig. 2 illustrates the Gaussian second-order patial derivatives in y-direction and xy-direction.

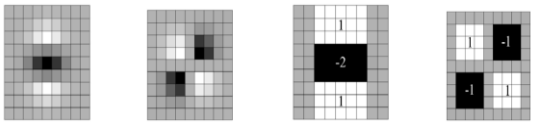

Fig. 2 The Gaussian second orders patial derivatives in y-direction and xy-direction [5]

III. PROPOSED METHOD

Our proposed identification system includes three stages: word segmentation using connected component grouping technique, feature extraction based on BoVW model, and classification using SVM. In the first stage, initially, a binarisation method is applied on the original image. Several methods are proposed based on the variety of thresholding techniques [9-15], Otsu's method is the most successful global thresholding method. Next, connected components (CCs) are identified and the noisy elements are filtered based on two properties of the CCs: bounding box height H(CC) and width W(CC) and the density $D(CC)=\frac{Fn(CC)}{H(CC).W(CC)}$ which is the ratio of the number of foreground pixels Fn(CC) to the total number of pixels in the bounding box.

CCs considered as noisy elements are eliminated if H(CC) < 2 or W(CC) < 2 or D(CC) < 0.05 or D(CC) > 0.9. The values of the various parameters have been decided to preserve the CCs containing text [16].

In order to localize each word, CCs in the same text line and with distance less than half of the average width and height of CCs or with overlapping pixels are united which are

forming words. The average width is calculated by:

$$AW=\frac{\Sigma_{i=0}^{k} W_i}{2k} \quad (3)$$

Fig. 3 Word segmentation procedure. From left to right: original image, binarized image, scanning for same block word, and word segmentation result.

$$AH=\frac{\Sigma_{i=0}^{k} H_i}{2k} \quad (4)$$

where I is the number of CCs, W_i and H_i are the width and height of all CCs in the image [4]. The result of this stage is illustrated in Fig. 3.

For the feature extraction stage, firstly, the codebook which contains all possible "visual words" present in all the blocks in the database as shown in Fig. 4. After that, a clustering algorithm is applied with a fixed number of clusters which defines the size of the codebook. It must be small enough to ensure low computation cost and large enough to provide accuracy performance. That is why k-means algorithm is satisfied because of its simplicity and processing speed. At the end of the process, the centers of the output clusters are the visual words of the codebook. In this paper, the number of centroids is 150.

When the creation of the codebook is finished, the calculation of each block descriptor will be started. Initially, the dimensions of the block are expanded so that the foreground pixels touching the block borders do not disturb with the calculation of the SURF features. Those SURFs whose position do not match the foreground pixel are discarded. Each of the remaining local features is assigned as a visual word from the codebook based on the minimum distance from the center of the corresponding cluster. Finally, a visual word vector is formed based on the appearance of each visual word of the codebook in each particular word block as shown in Fig. 5. The dimension of the vector is equal to the number of visual words in the

codebook. Normalization of the vector by dividing to its norm makes the number of the SURF features inside the block invariant.

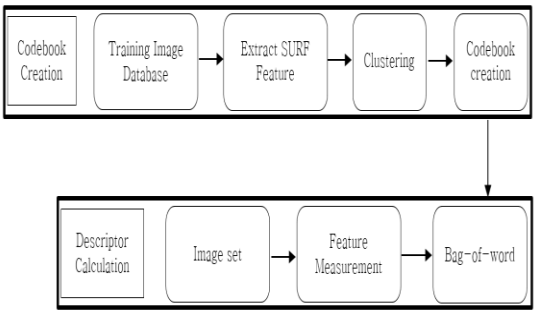

Fig. 4 Bag of Visual Word methodology

In the final stage, classification, a machine learning system decides which type of text in the block based on its descriptor. SVM is chosen based on its processing time and accuracy.

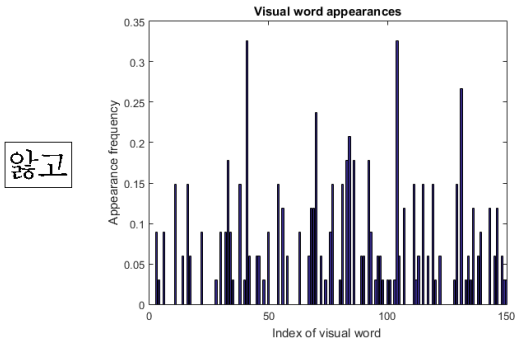

Fig. 5. Visual word vector. From left to right: original word block, and the visual word vector

IV. EXPERIMENTAL RESULTS

In the process of experimenting the proposed method, the dataset is used from Diotek company as shown in Fig. 6. Both methods using structure feature extraction and the proposed method are implemented in Matlab 2015 on Intel Core i7-4790 CPU at 3.6 GHz, 4 Gb RAM and Windows 7 system. The results are shown in Table 1. For evaluating the identification performance, 210 Korean document images including 3700 handwriting words and 27300 machine printed words are used. 1000 words of each kind are used for training, and the other for testing. The parameter accuracy is used to evaluate the results which is calculated by the following equation:

$$\text{Accuracy} = \frac{\text{number of right-classified word blocks}}{\text{total number of word blocks}} \quad (5)$$

Table 1. Experimental results based on structural features and bag-of-words model

	Structural features based method	Proposed method
Accuracy (%)	92.9	99.3
Processing time (sec)	370.21	750.74

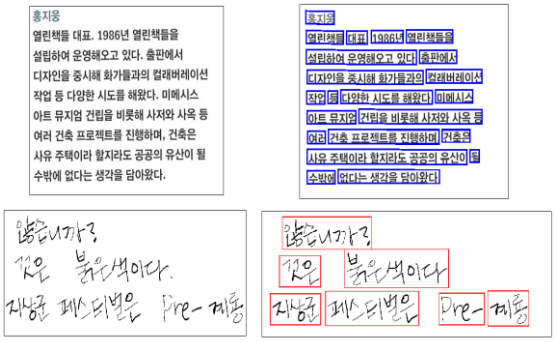

Fig. 6. The dataset and results. From right to left, top to bottom: Korean machine printed text and classification result; Korean handwritten text and classification result

V. CONCLUSION

This paper describes an off-line classification system based of Bag of Visual Words model using SURF feature. As you can see in the result part, the BoVW model shows a better results. However, the training and testing time is much slower than the structural features extraction. As a consequence, the future work is to improve the processing time of the BoVW model.

REFERENCES

[1] K. Dholakia, "A Survey on Handwritten Character Recognition Techniques for Various Indian Languages", IJCA, Vol. 115, April 2015.

[2] Y. Zheng, H. Li, D. Doermann, "Machine Printed Text and Handwriting Identification in Noisy Document Image", ICDAR, Sep. 2003.

[3] Y.Zheng, C.Liu, X.Ding, "Single Character Type Identification", in Proc. SPIE Conf. Document Recognition and Retrieval, pp.49-56, 2002.
[4] L.F. Silva, A. Sanchez, "Automatic Discrimination between Printed and Handwritten Text in Documents", ICDAR, 1999.
[5] D.Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", IJCV, 60(2):91-110, 2004.
[6] V.Vidyaharan and SubuSurendran, "Automatic Image Registration using SIFT-NCC", Special Issue of IJCA (0975-8887), pp.29-32, June 2012.
[7] H. Bay, A. Ess, T. Tuytelaars, F. V. Gool, "Speeded-Up Robust Features", JCVIU, Vol. 110, Issue 3, pp.346-359, June 2008,.
[8] L.Juan and O. Gwun, "A Comparison of SIFT, PCA-SIFT and SURF", IJIP, Vol. 3, Issue 4, Oct. 2009, pp.143-152.
[9] N. Otsu, "A threshold selection method from gray-level histogram", IEEE Trans. Syst. Man Cybern., Vol. 9, Issue 1, pp. 62-66, 1979,
[10] J. Bernsen, "Dynamic thresholding of gray level images", ICPR, pp. 1251-1255, 1986.
[11] G. Johannsen and J. Bille, "A threshold selection method using information measures", ICPR, pp. 140-143, 1982.
[12] N. J. Kapur, P. K. Sahoo, C. K. A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram", JCVPIP, Vol. 29, Issue 3, 273-285, 1985.
[13] J. Sauvola and M. Pietikainen, "Adaptive document image binarization", Pattern Recognition, Vol. 33, Issue 2, pp. 225-236, 2000.
[14] W.Niblack, "An introduction to digital image processing", pp. 115-116, Prentice Hall, Eaglewood Cliffs, 1986.
[15] J. Kittler and J. Illingworth, "Minimum error thresholding", Pattern Recognition, Vol. 19, Issue 1, pp. 41-47, 1986.
[16] K.Zagoris, I.Pratikakis, A. Antonacopoulos, B. Gatos, N. Napamarkos, "Distinction between Handwritten and Machine-printed Text Based on the Bag of Visual Words Model", Pattern Recognition Journal, ISSN 0031-3203, Vol.47, Issue 3, March 2014.

Authors

Son Tung Trieu

He received his B.E. degree in Microelectronics from Hanoi University of Science and Technology (HUST) in 2014. Since Sep 2015, he has been taking the M.S. course in Electronics and Computer Engineering at Chonnam National University, Korea. His research interests are mainly in the field of Image Processing, Computer Vision

Guee Sang Lee

He received his B.S. degree in Electrical Engineering and the M.S. degree in Computer Engineering from Seoul National University, Korea in 1980 and 1982, respectively. He received the Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering in Chonnam National University, Korea. His research interests are mainly in the field of Image Processing, Computer Vision and Video Technology.