2009 10th International Conference on Document Analysis and Recognition

Slit Style HOG Feature for Document Image Word Spotting

Kengo Terasawa[1,2] and Yuzuru Tanaka[3]

[1] *Department of Media Architecture, Future University-Hakodate, Japan*
[2] *PRESTO, Japan Science and Technology Agency, Japan*
[3] *Meme Media Laboratory, Hokkaido University, Japan*
kterasaw@fun.ac.jp, tanaka@meme.hokudai.ac.jp

Abstract

This paper presents a word spotting method based on line-segmentation, sliding window, continuous dynamic programming, and slit style HOG feature. Our method is applicable regardless of what language is written in the manuscript because it does not require any language-dependent preprocess. The slit style HOG feature is a gradient-distribution-based feature with overlapping normalization and redundant expression, and the use of this feature improved the performance of the word spotting. We compared our method with some previously developed word spotting methods, and confirmed that our method outperforms them in both English and Japanese manuscripts.

1. Introduction

Word spotting is the task of retrieving a text region that has a similar appearance to a query image specified by the user. It is of interest because it could treat the documents to which OCR (Optical Character Recognition) does not work well. Since it is purely appearance based method, word spotting has the advantage that it does not require prior learning. This advantage is important especially when we cannot provide enough corpus in advance, as is often the case with historical documents.

There exist a number of variations in word spotting methods. One such variation is to what extent the image should be segmented in preprocess. Word-segmentation-based methods such as [1, 2] assume that each word in the document images is separately clipped in the preprocess. Line-segmentation-based methods such as [3, 4, 5] assume that each line in the document images is separated in the preprocess, but do not assume word-segmentation. There also exist methods that do not require any prior segmentation in the preprocess such as [6].

Both in line-segmentation-based methods and word-segmentation-based methods, a sliding window (Fig. 1) is a widely used technique. In a sliding window technique, a window moves over the image along the writing direction and clips the sub-images. Each clipped sub-image is then converted to a feature vector (descriptor) by some kind of feature extraction method. In this way the document image is converted into the sequence of feature vectors. Thus the word spotting problem is converted to the problem of finding a sequence or subsequence of a vector that is similar to a query sequence. The remaining problem is how to extract feature vectors and how to find a similar sequences.

Many methods for feature extraction have been proposed. Rath and Manmatha [1] used projection, the position of upper/lower bound and the number of ink-paper transition as their descriptor. Wienecke *et al.* [7] proposed center of mass and angle at upper/lower bound as additional profiles. Kane *et al.* [8] counted ascender and descender for word profile. Since these methods are developed mainly for Latin manuscripts, some of their methods, such as baseline-estimation, slant-correction, ascender or descender detection, are not available for non-Latin manuscripts.

There also exist methods that are available regardless of what language is written in the manuscript. One of the authors has proposed the features based on eigenspace projection [4]. Recently, inspired by the success of the SIFT [9], the features based on oriented gradient were independently proposed by Rodríguez-Perronnin [2] and us [5].

As the ways of finding similar sequences, both DTW and HMM are often used. The advantage of HMM is that it can be also used in word recognition when appropriate prior training is possible. The advantage of DTW is that it does not require prior training.

This paper presents a word spotting method based on line-segmentation, sliding window, DTW-based matching, and slit style HOG feature. We compared our method with some previously developed word spotting methods, and confirmed that our method outperforms them in both English and Japanese manuscripts.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.118

2. Outline of the proposed method

In this section we present the outline of our word spotting method.

2.1. Line-Oriented Approach

Our method assumes that line-segmentation information is available but word-segmentation information is not. One advantage of line-segmentation-based method is that it can retrieve a hyphenated word that spans two lines. Another reason why we adopted such assumption was that we were intended to treat both Latin and non-Latin manuscripts. Word-segmentation is intrinsically impossible in Japanese manuscript because they do not put clear spacing between words in Japanese writing style. On the other hand, line-segmentation is possible for most languages and the process to obtain it is not so difficult in many cases. In fact, simply separating by the trough of the pixel projection along the writing direction was enough for our materials. For the case when the quality of such separation is not good enough, our algorithm can accept the manual correction, if needed.

After line separation, the background removal is performed to each separated image, just as we did in [4]. The background removal is more suitable for our method rather than the binarizing method because grayscale information is important for gradient-based feature. After that, we re-aligned the position of each column image using the center of mass in the relatively long window to remove the perturbation of handwriting. This process is similar to baseline estimation and correction used in Latin manuscript, but our method is also effective to non-Latin manuscript.

After preprocessing described above, we then apply the sliding window for each separated line. Here the window is a narrow rectangle that slides along the writing direction (Fig. 1). For each sub-image clipped by the sliding window, a feature vector is calculated (Our method for the feature calculation will be described in Sec. 3). In this way each document line is converted to the sequence of the feature vectors. By concatenating such sequences in the writing order, we can obtain the long sequence of the feature vectors. Thus, the word spotting problem is converted to the problem of retrieving similar sequence from the long sequence.

2.2. Continuous Dynamic Programming

For finding similar sequences, we used DTW (dynamic time warping)-based method. Since DTW does not require prior training, it is a better choice rather than HMM when we cannot obtain a large corpus in advance.

DTW is a method for evaluating similarities between two sequences of vectors with non-linear sequence alignment permitted. In our case where each vector in the long sequence is the candidate of the beginning point of the relevant sequence (remember that we do not assume word-segmentation), the CDP (continuous dynamic programming) [10] is a good choice for efficient calculation. CDP computes similarities between the pattern sequence and all possible subsequences of the full text at the same time with the constraint of local path as displayed in Fig. 2. The computational cost of CDP is bounded by $O(mn)$ where m is the length of the full text sequence and n is the length of the query sequence.

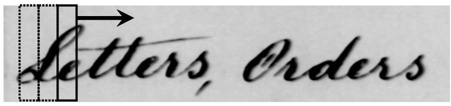

Figure 1. Sliding window

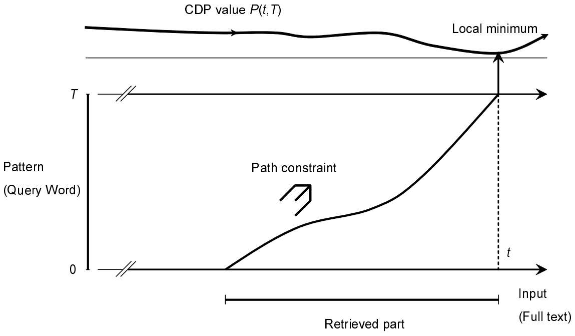

Figure 2. Continuous Dynamic Programming

3. Slit Style HOG Feature

This section presents the description of the slit style HOG (SSHOG) feature. SSHOG is a specifically modified variant of HOG [11] to be suitable for sliding-window-based word spotting.

3.1. HOG Feature

The HOG (Histogram of Oriented Gradients) [11] was originally designed for the task of human detection. Similar to SIFT [9], HOG computes a histogram of gradient orientations in a certain local region. One of the main differences between SIFT and HOG is that HOG normalizes such histograms in overlapping local blocks and makes a redundant expression. Another difference is that SIFT describes the scale- and orientation-normalized image patch around the specially detected keypoint, while HOG is computed in a rigid rectangular window without scale/orientation normalization.

In the HOG calculation, first we have to divide the image window into smaller rectangular regions called *cells*. Suppose we divide the image into $H{\times}W$ cells. Second, we must decide the number of the bins into which the weighted votes of the gradient vectors should be accumulated. In HOG, the orientation bins are evenly spaced over 0-180° ("unsigned gradient") or 0-360° ("signed gradient"). Note that SIFT always uses 8 bins over 0-360°. Let π denote the number of the bins over orientation. HOG accumulates the weighted vote according to the position and orientation of each pixel's gradients in three-dimensional (two for location and one for orientation) bins. To make the descriptor robust to small deformation, tri-linear interpolation should be applied. Thus we could obtain the histogram with $HW\pi$ bins.

HOG does not use this histogram as-is as a descriptor. Instead, HOG uses "Block Normalization." A block is defined as a group of $h \times w$ cells. The block slides inside the window image, that means $(H-h+1)\times(W-w+1)$ unique blocks exist. The HOG descriptor is a concatenation of the normalized block descriptors. Block descriptors are $hw\pi$ dimensional vectors each of which is a concatenation of the histogram components of the cells. Consequently, HOG descriptor has $(H-h+1)(W-w+1)hw\pi$ dimensionalities. As easily understood, this is a redundant expression in a sense that $HW\pi$ components in the original histogram composes a vector with $(H-h+1)(W-w+1)hw\pi$ dimensions. This redundancy is the salient characteristics of the HOG feature.

3.2. Slit Style HOG Feature

We made some modification to the HOG feature in order to make it more suitable for our word spotting task. Different from the human detection window used in [11], our window image (called "slit image") is a narrow rectangle as in Fig. 1. Therefore, we restricted the width of the block to be the same as the width of the slit. The horizontal overlapping of the original HOG could be well realized by the sliding window and sequential representation of the vectors.

Figure 3 represents the relationship between slit, blocks and cells in our method. Suppose that there is a slit image denoted as S_1. The slit image is divided into $H \times W$ cells as $h_{11}, \cdots, h_{24}$ in the figure. In this case, $H = 4$ and $W = 2$. Then we define the block as $h{\times}W$ group of cells. The width is limited to W; which is the difference to the original HOG. In the figure, we set the block size as 2×2. In this case we have 3 blocks as b_{11}, b_{12}, b_{13} with each block composed of 4 cells. Consequently, the dimensionality of our slit style HOG feature becomes $3 \times 4 \times \pi$ for this case.

Unlike the reference [11] used the unsigned gradient, we used the signed gradient for the orientation binning. The reason why unsigned gradient showed better result in human detection is estimated that the clothes of the human are sometimes brighter and sometimes darker than the background. The detection system cannot determine which model should be used in advance. On the other hand in our case, we cannot imagine the manuscripts in which characters darker than the background and characters brighter than the background are mixed. Therefore, a signed gradient should be appropriate rather than a unsigned gradient. We examined this fact by the experiment, and the result was as we had expected.

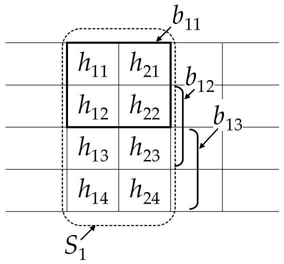

Figure 3. Block normalization for SSHOG

3.3. Optimal parameter estimation

In the algorithm described above, some parameters are remained undecided. Namely, the number of the bins for orientation, the size of the slit image, the number of the cells per slit and the number of the cells per block.

For estimating the optimal parameters, we executed preliminary experiments with two materials. One material was a Japanese manuscript: the scanned images of "Akoku Raishiki (The diary of Matsumae Kageyu)," which is exactly the same material as we have used in our previous study [5]. The other material was an English manuscript: George Washington collection (GW20) provided by Rath [12]. The sample page and the keywords used in the experiments are displayed in Fig. 4. Note that we can apply the same method to both vertical and horizontal manuscripts by just swapping x and y coordinate.

We calculated mean average-precision scores for all query words. After that, we calculated mean scores over all queries. For this experiment we used Recpret tool [13], which is a publicly available tool for evaluating the performance of the word spotting method.

3.3.1 Number of bins for orientation

The mean average-precisions with respect to the number of bins for orientation are plotted in Fig. 5 (a). As observed in the figure, the result with 8-bins for orientation was not the best one; this was the opposite to the result reported in [2]. In our observation, 12-bins and 16-bins produced the

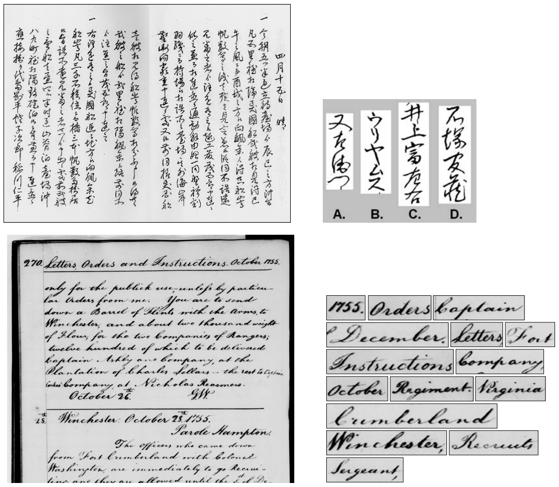

Figure 4. Experimental materials

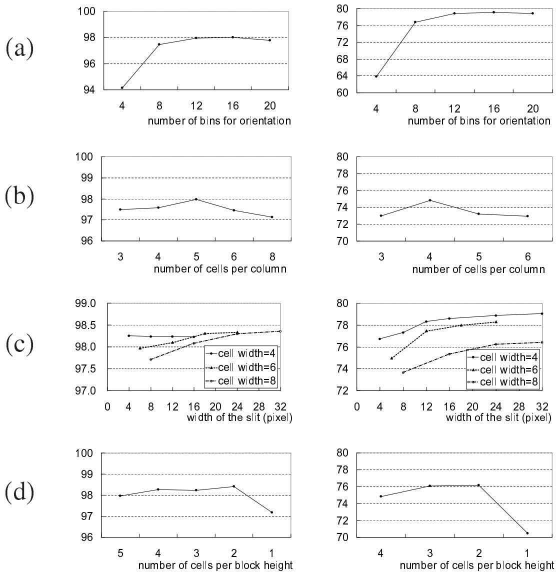

Figure 5. Optimal parameter estimation. Left column is a mean average-precision for Japanese manuscript. Right column is that for English manuscript

competitive result, 20-bins was slightly worse, and 8-bins was significantly worse.

3.3.2 Height of the cell

The effect of the number of cells in height is plotted in Fig. 5 (b). As observed in the figure, 4 was the best for English manuscript, and 5 was the best for Japanese manuscript. This result seems reasonable because Japanese characters tend to have higher complexity compared with English characters. In both manuscripts, the height per line was about 80 pixels including margins. Therefore, the estimated optimal heights of the cells were 20 pixels for English manuscript and 16 pixels for Japanese manuscript.

3.3.3 Width of the slit image and width of the cell

Figure 5 (c) depicts the effect of width of the slit image and width of the cell. In this experiment we could not find the clear optimum. As far as we have tested, to give a wider width for the slit image tended to improve the result, and to give a narrower width for the cell also tended to improve the result. Both of these parameter settings caused the increase of the dimensionality of the feature vector and made the computational cost expensive.

In the experiment in the following section, we decided to use parameters width_per_slit=16 and width_per_cell=4 as a compromise. These settings imply that the number of the cells for horizontal direction is four.

3.3.4 Number of cells per block height

Figure 5 (d) depicts the effect of the number of cells per block height. The leftmost point in the figure represents the case where number of cells per slit and number of the cells per block is the same, i.e., the case without redundant expression. As seen in the figure, if we set this parameter as two, the result was the best in both Japanese and English manuscripts. This result implies that the redundant expression again showed its effectiveness, as HOG showed in human detection task.

3.3.5 Other remarks

In [11] Dalal and Triggs declared that applying Gaussian blurring resulted in worsening the performance for human detection. On the other hand in our experiment, Gaussian blurring improved the result. Therefore we adapted Gaussian blurring in the experiment in the following section.

4. Experiment

With the optimized parameters obtained in the previous section, we compared the performance of our method with the performance of previously proposed methods. Experimental materials (both a set of images and a set of keywords to be retrieved) were just the same as the preliminary experiment described in Sec.3.3.

The result for Japanese manuscript is shown in Table 1. As observed in the table, our novel feature SSHOG outperformed our older feature GDF [5] in all query words.

Table 1. Mean average-precision for some keywords of Japanese manuscript

Keyword	freq.	GDF	SSHOG
A. Matazaemon	166	97.24	97.61
B. Uriyamusu	73	97.80	99.88
C. InoueTomizou	25	99.15	99.34
D. IshizukaKanzou	25	98.20	99.31

Table 2. Mean average-precision (mAP) and mean R-precision (mRP) for GW20 dataset

Method	Segmen-	Performance	
	tation	mAP(%)	mRP(%)
Terasawa (proposal)	line	79.14	74.94
Rath	word	73.95	
Leydier	none		60

The result for English manuscript is shown in Table 2. Here we have to mention that the set of keywords used in the experiment were not exactly the same between compared methods. Rath *et al.* [1] did not provide the list of the words used in their experiment. In [6], Leydier *et al.* specified the keywords they used in the experiment, and we used exactly the same keywords as they used, however, the details were slightly different; we included the hyphenated word in the ground-truth, while Leydier did not. The result displayed in the Table 2 indicates that our method again outperformed the previously proposed methods.

5. Conclusion and Future Work

In this paper, we proposed a word spotting method based on line-segmentation, sliding window, continuous dynamic programming, and slit style HOG feature. We have confirmed that the optimal number of bins for orientation is 12 or 16, rather than 8. We also confirmed that the overlapping normalization and redundant expression of the feature improved the performance of the word spotting. Our method outperformed the previously proposed methods in both Japanese and English manuscripts.

The drawback of the SSHOG feature is its high dimensionality that increases the computational cost. Our future work will focus on applying our Locality-Sensitive Pseudo-Code (LSPC) technique [14] for the SSHOG to compensate for the drawback of its high dimensionality. To verify the robustness of the optimal parameters will also be included in our future work.

References

[1] T. M. Rath and R. Manmatha, "Word Spotting for Historical Documents," Int. J. Document Analysis and Recognition, vol. 9, no. 2, pp. 139–152, 2007.

[2] J. A. Rodríguez and F. Perronnin, "Local gradient histogram features for word spotting in unconstrained handwritten documents," Proc. ICFHR2008, 2008.

[3] A. Kołcz, J. Alspector, M. Augusteijn, R. Carlson, and G. Viorel Popescu, "A Line-Oriented Approach to Word Spotting in Handwritten Documents," Pattern Analysis and Applications, vol. 3, no. 2, pp. 153–168, 2000.

[4] K. Terasawa, T. Nagasaki, and T. Kawashima, "Eigenspace Method for Text Retrieval in Historical Document Images," Proc. ICDAR2005, vol.1, pp.437–441, 2005.

[5] K. Terasawa, T. Nagasaki, T. Kawashima, "Improved Handwritten Text Retrieval Using Gradient Distribution Features," Proc. Meeting on Image Recognition and Understanding, MIRU2006, pp.1325–1330, 2006. (written in Japanese)

[6] Y. Leydier, F. Lebourgeois, and H. Emptoz, "Text Search for medieval manuscript images," Pattern Recognition, vol. 40, no. 12, pp. 3552–3567, 2007.

[7] M. Wienecke, G. A. Fink, and G. Sagerer, "Toward automatic video-based whiteboard reading," Int. J. Document Analysis and Recognition, vol. 7, no. 2–3, pp.188–200, 2005.

[8] S. Kane, A. Lehman, and E. Partridge, "Indexing George Washington's handwritten manuscripts," Technical Report MM-34, Center for Intelligent Information Retrieval, University of Massachusetts Amherst, 2001.

[9] D. G. Lowe, "Object recognition from local scale-invariant features," Proc. ICCV'99, vol. 2, pp. 1150–1157, 1999.

[10] R. Oka, "Spotting Method for Classification of Real World Data," The Computer Journal, vol. 41, no. 8, pp. 559–565, 1998.

[11] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," Proc. CVPR2005, pp. 886–893, 2005.

[12] V. Lavrenko, T. M. Rath and R. Manmatha, "Holistic Word Recognition for Handwritten Historical Documents," Proc. DIAL2004, pp. 278–287. 2004.

[13] K. Terasawa, H. Imura, and Y. Tanaka, "Automatic evaluation framework for Word Spotting," Proc. ICDAR2009, In press.

[14] K. Terasawa and Y. Tanaka, "Locality Sensitive Pseudo-Code for Document Images," Proc. ICDAR2007, vol. 1, pp. 73–77, 2007.