Knowledge-Based Systems 90 (2015) 58–69

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Semi-supervised cluster-and-label with feature based re-clustering to reduce noise in Thai document images

N. Piroonsup, S. Sinthupinyo*

Department of Computer Engineering, Chulalongkorn University, 254 Phayathai Road, Pathumwan, Bangkok 10330 Thailand

ARTICLE INFO

Article history:
Received 6 March 2015
Revised 19 August 2015
Accepted 28 September 2015
Available online 8 October 2015

Keywords:
Noise reduction
Document enhancement
Semi-supervised classification
Cluster-and-label
Thai document

ABSTRACT

Noise components are a major cause of poor performance in document analysis. To reduce undesired components, most recent research works have applied an image processing technique. However, the effectiveness of these techniques is suitable only for a Latin script document but not a non-Latin script document. The characteristics of the non-Latin script document, such as Thai, are considerably more complicated than the Latin script document and include many levels of character alignment, no word or sentence separator, and variability in a character's size. When applying an image processing technique to a Thai document, we usually remove the characters that are relatively close to noise. Hence, in this paper, we propose a novel noise reduction method by applying a machine learning technique to classify and reduce noise in document images. The proposed method uses a semi-supervised cluster-and-label approach with an improved labeling method, namely, feature selected sub-cluster labeling. Feature selected sub-cluster labeling focuses on the clusters that are incorrectly labeled by conventional labeling methods. These clusters are re-clustered into small groups with a new feature set that is selected according to class labels. The experimental results show that this method can significantly improve the accuracy of labeling examples and the performance of classification. We compared the performance of noise reduction and character preservation between the proposed method and two related noise reduction approaches, i.e., a two-phased stroke-like pattern noise (SPN) removal and a commercial noise reduction software called ScanFix Xpress 6.0. The results show that semi-supervised noise reduction is significantly better than the compared methods of which an F-measure of character and noise is 86.01 and 97.82, respectively.

1. Introduction

Normally, the document analysis process works effectively on clean document images. The clean images, however, are rarely found in real-world situations. A real-world document image usually contains noise components that can dramatically decrease a performance of document analysis e.g., accuracy of optical character recognition (OCR). Noise in a document image stems from many sources (e.g., paper background, document aging, water drop and notation) with various properties. Recent research works on noise reduction usually employ one or more image processing techniques with specific noise properties [1]. These aim to reduce a specific type of noise, e.g., salt-and-pepper noise [2], line of writing [3], double-side writing interference [4], preprinted form [5], blob noise [6], and background [7,8].

Although the image processing technique is commonly applied to document image noise reduction, the effectiveness of this method is limited. First, it is only appropriated for a document image with a Latin based script (e.g., English, German and French) but not for a document with a non-Latin based script (e.g., Persian, Arabic and Thai) because the non-Latin script is ordinarily composed of many small characters whose size is similar to noise. The image processing technique typically specifies a component size threshold to separate character and noise components. If the size of character varies as in non-Latin script, some small characters may be removed when its size is less than the threshold, and a noise component will not be removed when its size is larger than the threshold. Second, many historical documents were digitized into black and white or binary images due to the limited storage space and technology during a digitization time. This black and white image contains restricted information (e.g., color depth, contrast) for image processing noise reduction methods. Re-digitizing these documents is impracticable because most of them were considerably more degraded over time, and many of them were destroyed. With the limitations above, an effective method is required to reduce noise in non-Latin script.

* Corresponding author. Tel.: +66 22186973; fax: +66 22186955.
E-mail addresses: nareeporn.p@student.chula.ac.th, naree.p@gmail.com (N. Piroonsup), sukree.s@chula.ac.th (S. Sinthupinyo).

http://dx.doi.org/10.1016/j.knosys.2015.09.033
0950-7051/© 2015 Elsevier B.V. All rights reserved.

In this paper, we propose a novel noise reduction method for reducing noise in a black and white Thai document image. A Thai document is a fascinating example of a non-Latin document because of its intricate properties, i.e., it consists of many small characters, many levels of character alignment and no word and sentence separator. In addition to using an image processing technique, we utilize a machine learning technique for reducing noise in the Thai document image. The benefit of applying a machine learning technique is that it need not specify noise criteria for distinction, but instead, a classifier is trained to distinguish noise from a character. An efficient classifier can discriminate even noise that its size is quite similar to a small character, as frequently presented in Thai script. However, to build an efficient classifier, we need a sufficient number of labeled examples. Although a large number of document images is available, the labeled components are scarce because the labeling process is effort intensive and is a time consuming task. To minimize the cost of human effort, we utilize the available unlabeled examples by applying a semi-supervised classification. The semi-supervised classification is a widely used method for the problem of limited labeled data, for example, a cross-lingual sentiment classification. Hajmohammadi [9] proposed a combination of semi-supervised and active learning methods to classify a sentiment document in the target language, which is not English and rarely finds the labeled data, by using the source language document, which is in English and the labeled sentiment documents are available.

In this work, we apply the semi-supervised cluster-and-label approach to create the classifier for noise reduction. First, all examples, with or without the class labeled, are clustered by their properties. Then, a majority class of labeled examples in each cluster is specified. The unlabeled examples in each cluster are then labeled as the majority class in theirs cluster. These recently labeled examples and the prior labeled examples are used to train the final classifier. Finally, the classifier classifies all components and removes the noise classified components from the document images.

When we applied the cluster-and-label procedure, another problem arose. The mislabeled examples were apparently found in a mixed-class cluster, or a cluster might contain various classes of labeled examples. Hence, we proposed a radical labeling method to improve accuracy of example labeling in the mixed-class cluster, namely, feature selected sub-cluster labeling. The idea is that if different-class examples are unintentionally grouped into the same clusters, sub-clustering will re-organize examples into a proper subgroup. Because the current feature set is ineffectual to separate examples, a new feature set for sub-clustering is then needed. A particular feature set is selected by a feature selection with information gain. This particular feature is then used to cluster examples in each mixed-class cluster into a satisfying sub-cluster. The unlabeled examples are then labeled with a class of labeled examples of their affiliated sub-cluster.

The performance of feature selected sub-cluster labeling is compared with a conventional majority vote labeling. The results show that the proposed labeling method improves the accuracy of labeling in the mixed-class cluster and provides an efficient classifier. The performance of semi-supervised noise reduction with feature selected sub-cluster labeling is compared with two related noise reduction methods, i.e., a two-phased stroke-like pattern noise (SPN) removal [10] and the commercial noise reduction software, namely, ScanFix Xpress 6.0 [11]. The results show that semi-supervised noise reduction is significantly better in noise reduction and character preservation than the compared methods. We further analyzed the reason for improvement by using our method and found that for the small characters that are used frequently in Thai documents, they were clustered in the mixed-class cluster and labeled incorrectly by the standard cluster-and-label method. However, our proposed method can improve the accuracy of the small character portion of the test set.

The structure of this paper is as follows. A review and discussion regarding state-of-the-art noise reduction methods in document images are presented in Section 2. The property of Thai script as opposed to Latin script is described in Section 3. An algorithm for semi-supervised cluster-and-label is described in Section 4. In Section 5, we present a methodology of the proposed semi-supervised noise reduction with feature selected sub-cluster labeling. The results and discussion are presented in Section 6. The last section provides a conclusion and future work.

This work differs from our previous work [12,13] in two aspects. First, in previous work, we considered a line of a connected component as an example with an aim to reduce noise that might attach to a character. However, we found that the line-level example might contain an insignificant amount of information to distinguish a character from noise. As a result, this work uses a connected component as an example and focuses on removing particular noise, whose size is similar to a small character. Second, in previous work, we applied a traditional majority vote labeling in semi-supervised cluster-and-label classification to reduce noise but, in this work, we proposed a novel feature selected sub-cluster labeling method for semi-supervised cluster-and-label.

2. Noise reduction in a document image

Noise in a document image, in this work, is defined as any foreground component in the document image except a printed character. Noises can stem from many sources either intentionally (such as a water mark, rubber stamp and a signature) or accidentally (such as a paper wrinkle, a water drop and a worm hole). To the best of our knowledge, these various noises have not been categorized into a distinct category. There was one study that proposed a noise category by a source of noise, e.g., physical noise caused by damage in a document paper and digitization noise caused by an error in a conversion process of a document paper to a digital image [14]. In this work, we categorize noise by its characteristics, i.e., a specific pattern noise that explicitly diverges from a character so-called "fixed-form noise" and fuzzy pattern noise that is possibly akin to some characters so-called "free-form noise".

Several studies tried to eliminate the fixed-form noise. Examples of fixed-form noise removal methods are a salt-and-pepper noise removed by applying the k-Filled algorithm [2]; a line of writing removed by considering a threshold of the character's length and width [3]; a pre-printed form removed by comparing it to a blank form [5]; a blob removed by considering the size and position of a large component [6]; and an interfering stroke in a double-sided handwriting document removed by considering contrast and the prior knowledge of alignment of handwritten English characters [4]. To apply any of these methods, a user must define criteria to identify noise from a character component and then apply some image processing techniques to reduce the selected component. However, these approaches seem dubious when applying them to free-form noise or noise whose attribute is quite similar to the character. For example, if line thickness of pre-printed forms is quite similar to that of the character, this noise will not be removed [5]. Although extensive studies have been performed on fixed-form noise removal, very few studies have given attention to free-from noise removal [6].

Another challenge is noise reduction on a binary document image. A binary document image is a document that is digitized into a black and white image due to the limited storage space and technology during the digitization time. Re-digitizing these documents could be impracticable because most of these documents were considerably more degraded over time, and many of them were destroyed. This binary image is typically historical documents that usually consist of considerable noise. So, performing noise reduction on this historical document image is essential. However, because the binary image contains restricted information (e.g., color depth, contrast) for the