Handwritten and Machine Printed Text Separation in Document Images using the Bag of Visual Words Paradigm

Konstantinos Zagoris[1,2], Ioannis Pratikakis[2], Apostolos Antonacopoulos[1],
Basilis Gatos[3], Nikos Papamarkos[2]
[1]*Pattern Recognition and Image Analysis (PRImA) Research Lab*
School of Computing, Science and Engineering, University of Salford, Greater Manchester, UK
[2]*Department of Electrical and Computer Engineering*
Democritus University of Thrace, Xanthi, Greece
[3]*Institute of Informatics and Telecommunications, – National Center for Scientific Research "Demokritos" Athens, Greece*
kzagoris@ee.duth.gr, ipratika@ee.duth.gr, a.antonacopoulos@primaresearch.org, bgat@iit.demokritos.gr, papamark@ee.duth.gr

Abstract

In a number of types of documents, ranging from forms to archive documents and books with annotations, machine printed and handwritten text may be present in the same document image, giving rise to significant issues within a digitisation and recognition pipeline. It is therefore necessary to separate the two types of text before applying different recognition methodologies to each. In this paper, a new approach is proposed which strives towards identifying and separating handwritten from machine printed text using the Bag of Visual Words paradigm (BoVW). Initially, blocks of interest are detected in the document image. For each block, a descriptor is calculated based on the BoVW. The final characterization of the blocks as Handwritten, Machine Printed or Noise is made by a Support Vector Machine classifier. The promising performance of the proposed approach is shown by using a consistent evaluation methodology which couples meaningful measures along with a new dataset.

1. Introduction

There exist a rapidly growing number of digitization initiatives in libraries and archives, involving a variety of document types. Among several other obstacles, the presence of printed and handwritten text in the same document image gives rise to significant issues since each modality requires different treatment to recognize the corresponding characters [9, 11]. Furthermore, the automatic processing of application forms, bank checks, petitions, mail papers, etc. necessitates the discrimination of handwritten from machine-printed text.

Previously, Pal and Chaudhuri [21, 20] assert a method to separate the machine-printed and hand-written text lines for Bangla and Devnagari scripts, two popular scripts in south Asia.

Guo and Ma [10] segment the image document into blocks by generating initially the connected components and subsequently merge them based on a set of conditions. For each character inside the block, a projection profile is created and then quantized. Therefore, for each block a sequence of quantized values is computed. The classification of the aforementioned sequence as handwritten or machine-printed text is achieved by using Hidden Markov Models.

Fan et al. [7] propose a method to initially detect the orientation of a text block by analyzing the valleys in horizontal and vertical projection profiles. Then, the image character blocks are obtained by employing an X-Y cut algorithm. Lastly, the classification goal is achieved using the block layout variance as the feature that incorporates spatial information.

Zheng et al. [25] identify machined printed and handwriting text in noisy document images. They

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.207

calculate the connected components in a page and then merge them based on spatial proximity in order to form blocks. For the text identification (handwritten, machine-printed or noise) they initially extract several sets of features. For the block classification, the Fisher classifier is considered.

In this paper, we propose a new approach dealing with the problem of handwritten and machine-printed text separation using the Bag of Visual Words (BoVW) model and Scale-Invariant Feature Transform (SIFT) features. The paper is structured as follows: Section 2 details the proposed methodology, Section 3 discusses the corresponding evaluation and finally, at Section 4, conclusions are drawn.

2. The Proposed Methodology

2.1 Bag of Visual Words (BoVW) Model

The BoVW model is inspired by the Bag of Words (BoW) model employed in information retrieval in which a document is described by a set of words. Accordingly, the BoVW model comprises a set of "visual words" to describe the image content.

A "visual word" is expressed by a group of features that correspond to local image information which is identified by the image keypoints [23]. One of the most well-known local features is the Scale-Invariant Feature Transform (SIFT) [14], which is also employed by the proposed method. This is due to inherent SIFT's invariance to scale and rotation as well its robustness across considerable range of distortion, noise contamination and change in brightness.

These features are grouped in a number of clusters. A "visual word" is denoted as the vector which represents the center of each cluster while the set of the clusters defines a codebook which is analogous to a dictionary. In particular, each SIFT point belongs to a visual word which corresponds to the closest center of the cluster calculated by a distance function such as Euclidean, Manhattan, etc (see Figure 1: Visual Words Assignment). Finally, the image is represented by a vector which denotes the corresponding descriptor [13] and it reflects the frequency of each visual word that appears in the image. Figure 1 illustrates the BoVW paradigm.

There has been considerable work based on BoVW in a variety of subjects.

Sheng Xu et al. [24] use the BoVW model for object-based classification in land-use/cover mapping of high spatial resolution aerial photographs. They use a combination of spectral and texture features from which they create a visual vocabulary.

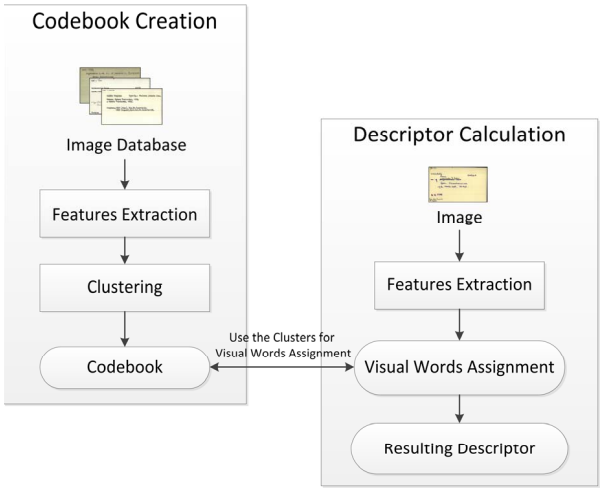


Figure 1. The BoVW paradigm.

Nilsback and Zisserman [19] introduced a flower classification technique by developing a bag of visual words model. They show that their work surpasses the baseline algorithms.

Deselaers et al. [6] presented an adult image detection and filtering method based on the BoVW classification model. They demonstrated that integrating standard skin color features into their system led to an improvement compared to the standard model.

It is worth noting, however, that to the best of the authors' knowledge there is no approach using the BoVW model to discriminate between handwritten and machine printed text in document images.

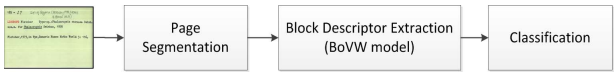


Figure 2. The main stages of the proposed method.

The incorporation of this model to the separation of machine printed and handwritten text is illustrated in Figure 2, which depicts the main stages of the proposed method. It is composed of three stages:

1. *Page Segmentation:* The objective of this stage is to detect blocks of interest in the document image. This is elaborated in Section 2.2.
2. *Block Descriptor Extraction:* In this stage, the descriptor is calculated based on the BoVW model. This procedure is described in Section 2.3.
3. *Classification*: The final stage, in which a machine learning system decides what type of text (if any) resides in the block based on its descriptor set. This technique is detailed in Section 2.4.

2.2. Page Segmentation

The main objective of this stage is to detect blocks of interest in the document image. Figure 3 shows the consecutive steps of the proposed technique. Initially, a locally adaptive binarisation method [8] is applied on the original image (Figure 4(b)) which and improves the quality of degraded documents enhancing the textual information without requiring any parameter tuning.

In the sequel, connected components (CCs) are identified in the image (Figure 4(c)) and the noisy elements are filtered out based on three characteristics of the CCs in the following:

- Bounding Box *Height* $H(CC)$ and *Width* $W(CC)$.

- The *elongation* $E(CC) = \dfrac{\min\{H(CC), W(CC)\}}{\max\{H(CC), W(CC)\}}$

- The *density* $D(CC) = \dfrac{Fn(CC)}{H(CC) \cdot W(CC)}$, which is the ratio of the number of foreground pixels $Fn(CC)$ to the total number of pixels in the bounding box.

After systematic experimentation, CCs are considered as noisy elements and are eliminated if $H(CC) < 2$ or $W(CC) < 2$ or $D(CC) < 0.05$ or $D(CC) > 0.9$ or $E(CC) < 0.08$ (Figure 4(d)). The values of the various parameters have been chosen with the goal being that CCs containing text are preserved.

Figure 3. The steps for page segmentation.

The next step involves merging of distinct CCs towards creating blocks of interest consistent with document words. It is not a requirement for the success of the proposed method but a tradeoff well suited to the problem. The block size must be large enough to contain SIFT points but at the same time not too large to give rise to ambiguities in the final descriptor. This task is accomplished by the Adaptive Run Length Smoothing Algorithm (ARLSA) [18] (Figure 4(e)) which is a modified version of the horizontal RLSA. This is a word segmentation method which resolves successfully challenges like text with various font sizes, high proximity text and not-text areas and warped of overlapping test lines.

The output of the Page Segmentation stage (Figure 4(f)) is a list of blocks in the document image. The next section details how each block is attributed to a descriptor.

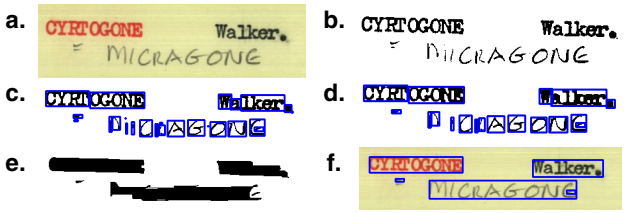


Figure 4. a. Original image; b. Binarised image; c. CCs before filtering; d. CCs after filtering; e. ARLSA output; f. Final result

2.3. Block Descriptor Extraction

This step involves the creation of the block descriptor by utilizing the BoVW model. First of all, the codebook which will accommodate all possible "visual words" present in all the blocks in the dataset. Figure 5 details the individual steps required to create the codebook.

Figure 5. The steps for codebook creation

After block detection and features extraction for each block a clustering is applied with a fixed number of clusters which also, defines the size of the codebook. Predicting the optimal codebook size is non-straightforward and dataset-dependent. Generally, it must accommodate the following rules:

- It must be small enough to ensure a low computational cost.
- It must be large enough to provide sufficiently high discrimination performance.

For the clustering stage the k-means algorithm is employed due to its simplicity and speed. At the end of the process, the centers of the output clusters are the visual words of the codebook.

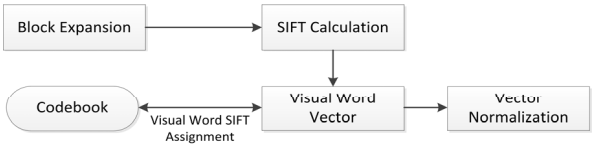
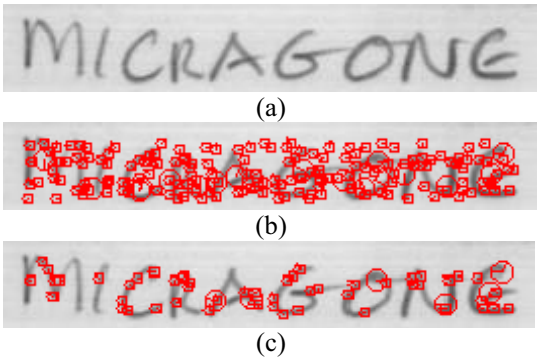


Figure 6. The creation of the block descriptor

After the codebook creation, the calculation of each block descriptor follows. Figure 6 illustrates the required steps. Initially, the dimensions of the block are expanded so that the foreground pixels touching the block borders do not interfere with the calculation of the SIFT features. The SIFTs are calculated on the greyscale version (Figure 7(b)) of the original document image and not on the binarised version of it.

Finally, those SIFTs whose position in the binary image does not match the foreground pixel are rejected (Figure 7(c)).



(a) (b) (c)

Figure 7. (a) An Example text block; (b) Initial SIFT keypoints; (c) Final SIFT keypoints

Each of the remaining local features is assigned 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 this particular block. For instance, consider a Codebook with 5 visual words and a block that contains 10 SIFTs which are assigned as follows: 2 SIFTs for the first visual word, 3 SIFTs for the second, 4 SIFTs for the third and 1 SIFT for the fifth. Then the Bag of Visual Words Vector is $BVWV = [2,3,4,0,1]$. Note that the dimension of the vector is equal to the number of visual words in the Codebook.

The last step involves the normalization of the vector by dividing to its norm and making it invariant to the number of the SIFTs inside the block.

2.4. Classification

In this final stage, a classifier decides if the visual word vector of the block contains handwritten or machine printed text or neither of the above (noise).

The proposed approach is based on the Support Vector Machines (SVMs) [3, 5]. The SVMs are based on statistical learning theory and have been applied to a large number of different classification problems. The SVMs are chosen based on their high performance and their ability that do not require large training sets.

The blocks resulting from the 'Page Segmentation' stage may contain three types of content: handwritten text, machine-printed text or noise. Therefore, the SVM must classify the block based on the Bag of Visual Words Vector in these three classes.

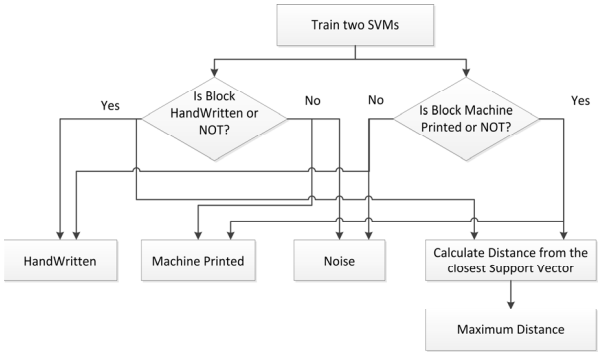


Figure 8. The Classification System algorithm

To achieve this, two SVMs are trained as follows: The first (SVM_1) deals with the handwritten text problem against all the other and the second (SVM_2) deals with the machine printed text problem against all the other. Figure 8 illustrates the Classification Scheme. There are four outcomes from the aforementioned SVMs.

- If the SVM_1 output is *TRUE* and SVM_2 is *FALSE* then the block contains handwritten text.
- If the SVM_1 output is *FALSE* and SVM_2 is *TRUE* then the block contains machine printed text.
- If the SVM_1 and SVM_2 output is *FALSE* then the block contains noise.
- If the SVM_1 and SVM_2 output is *TRUE* then the distance of the block descriptor with the closest Support Vector for each SVM_i is calculated. Finally, among those two distances the SVM_i that is related to the maximum distance defines the class of the block.

The above approach was chosen because the third class which corresponds to noise does not appear frequently (at least not in the chosen application domain). Therefore, if the common approaches are used (one-against-all, one-against-one) it may skew the results. Another advantage of the proposed approach is the training of only two SVMs instead of three SVMs. This reduces the computational cost and considerably increases the speed of the process. Figure 9 shows the output of the proposed method for an example document image.

3. Experimental Results and Discussion

For the evaluation of the proposed method, two datasets are used:

- 103 modified document images from the IAM Handwriting Database [17], which comprises forms that contain both handwritten and machine printed English text. In this dataset, the ground truth for the machine printed text was created by the authors.

- 100 representative images selected from the index cards in the UK Natural History Museum's card archive to the scientific names of world Lepidoptera [2]. These cards contain typewritten and handwritten text. Ground truth was created by the authors. This selection is denoted as PRImA-NHM

The ground truth files adhere to the Page Analysis and Ground-truth Elements (PAGE) format framework [22] which is an XML-based representation framework that records detailed information on various aspects of document images and their content. The ground truth files were created using the Aletheia tool [4], an advanced document layout and text ground-truthing system.

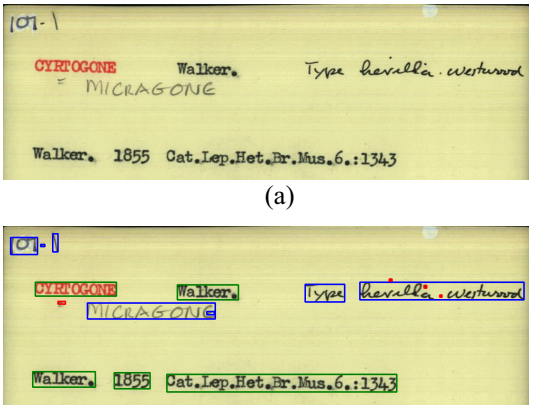
For the SVMs, we used a Radial Basis Function (RBF) kernel trained approximately on 10% samples of the entire content of each database. The datasets with the corresponding ground truth files are available freely (see http://datasets.primaresearch.org). The codebook is created by clustering the training samples in 150 "visual words". We use the default parameters for the Adaptive Run Length Smoothing Algorithm as they are provided from the authors of the original work except for the constant a which we decreased ($a = 1$) in order to adjust the merging of the connected components to produce the desirable size.

The evaluation of the complete proposed system is an aspect not as trivial as it might seem. For their experimentation most researchers use simple methods [16, 12] such as pixel-based or box-based recall, precision measures. Unfortunately, those evaluation strategies have several drawbacks. On the one hand, in box-based approaches the number of retrieved pixels does not correspond to proportional textual information and on the other hand the mapping between ground truth and detected objects in box-based approaches can produce arbitrarily in bounding box splits or merges among annotators and detectors. To overcome these problems, we employ the estimated character-based F-measure [1] technique. Table 1 shows the F-measure of the proposed method.

Furthermore, in order to demonstrate the effectiveness of the BoVW model, an experiment is conducted in which the Classification Stage output is always correct in the characterization of the blocks. Therefore, the error that originates from the Page Segmentation Stage is known and consequently the upper bound of the BoVW method is also known.

Moreover, to further evaluate the proposed method, the whole BoVW model is replaced with Gabor Filters [15]. The Page Segmentation Stage and the Classification Stage remain the same, but the block descriptor is calculated by the Gabor Filters.

As Table 1 shows, the proposed BoVW-based model exhibits better performance than the Gabor Filters-based one and it approaches the perfect outcome (in IAM database is approximately the same).



(a)

(b)

Figure 9. (a) The original image; (b) The output of the proposed method. Blocks in 'blue' contain handwritten text, blocks in 'green' contain machine printed text and blocks in 'red' contain noise.

Table 1. The F-measure of each method.

Dataset	IAM	PRImA-NHM
Upper Bound (Proposed Segmentation)	0.9887	0.7985
Proposed Method (Proposed Segmentation and BoVW)	**0.9886**	**0.7689**
Gabor Filters (Proposed Segmentation and Gabor Filters)	0.7921	0.5702

4. Conclusions

In this paper, a method based on the Bag of Visual Words paradigm was presented for the separation of the machine printed and handwritten text. It consists of three stages: The Page Segmentation stage which it detects blocks of interest on the document image, the Block Descriptor Extraction stage, which calculates the descriptors of the extracted blocks using the BoVW model and the Classification stage which characterizes the blocks as handwritten, machine printed or noise. Moreover, an evaluation dataset with ground truth is provided, created especially for this task. Experimental results using a consistent evaluation procedure have shown the significant promise of the proposed methodology.

References

[1] M. Anthimopoulos, B. Gatos, and I. Pratikakis. A two-stage scheme for text detection in video images. *Image and Vision Computing*, 28(9):1413–1426, 2010.

[2] G. Beccaloni, M. Scoble, L. Kitching, T. Simonsen, G. Robinson, B. Pitkin, and A. Hine. The global lepidoptera names index (lepindex). WWW electronic publication. http://www.nhm.ac.uk/entomology/lepindex [accessed 12 March 2012].

[3] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT*, pages 144–152, 1992.

[4] C. Clausner, S. Pletschacher, and A. Antonacopoulos. Aletheia-an advanced document layout and text ground-truthing system for production environments. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pages 48–52. IEEE, 2011.

[5] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–197, 1995.

[6] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-visual-words models for adult image classification and filtering. In *ICPR*, pages 1–4, 2008.

[7] K.-C. FAN, L.-S. WANG, and Y.-T. TU. Classification of machine-printed and handwritten texts using character block layout variance. *Pattern Recognition*, 31(9):1275 – 1284, 1998.

[8] B. Gatos, I. Pratikakis, and S. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39(3):317–327, 2006.

[9] V. Govindan and A. Shivaprasad. Character recognition – a review. *Pattern Recognition*, 23(7):671 – 683, 1990.

[10] J. K. Guo and M. Y. Ma. Separating handwritten material from machine printed text using hidden markov models. *Document Analysis and Recognition, International Conference on*, 0:0439, 2001.

[11] S. Impedovo, L. Ottaviano, and S. Occhinegro. Optical character recognition-a survey. *Issues*, 1(2):1–24, 1991.

[12] C. Jung, Q. Liu, and J. Kim. A stroke filter and its application to text localization. *Pattern Recognition Letters*, 30(2):114–122, 2009.

[13] M. Kogler and M. Lux. Bag of visual words revisited: an exploratory study on robust image retrieval exploiting fuzzy codebooks. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 3:1–3:6, New York, NY, USA, 2010. ACM.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[15] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In *ACM Multimedia*, pages 1085–1088, 2008.

[16] M. Lyu, J. Song, and M. Cai. A comprehensive method for multilingual video text detection, localization, and extraction. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(2):243–255, 2005.

[17] U. Marti and H. Bunke. The IAM-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 5(1):39–46, 2002.

[18] N. Nikolaou, M. Makridis, B. Gatos, N. Stamatopoulos, and N. Papamarkos. Segmentation of historical machine-printed documents using adaptive run length smoothing and skeleton segmentation paths. *Image and Vision Computing*, 28(4):590–604, 2010.

[19] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1447 – 1454, 2006.

[20] U. Pal and B. B. Chaudhuri. Automatic separation of machine-printed and hand-written text lines. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition*, ICDAR '99, pages 645–, Washington, DC, USA, 1999. IEEE Computer Society.

[21] U. Pal and B. B. Chaudhuri. Machine-printed and hand-written text lines identification. *Pattern Recognition Letters*, 22(3-4):431 – 441, 2001.

[22] S. Pletschacher and A. Antonacopoulos. The page (page analysis and ground-truth elements) format framework. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 257–260. IEEE, 2010.

[23] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Found. Trends. Comput. Graph. Vis.*, 3:177–280, July 2008.

[24] S. Xu, T. Fang, D. Li, and S. Wang. Object classification of aerial images with bag-of-visual words. *Geoscience and Remote Sensing Letters, IEEE*, 7(2):366 –370, april 2010.

[25] Y. Zheng, H. Li, and D. Doermann. Machine printed text and handwriting identification in noisy document images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(3):337 –353, march 2004.