

Automatic text block separation in document images

Abstract—Separation of printed text blocks from the non-text areas, containing signatures, handwritten text, logos and other such symbols, is a necessary first step for an OCR involving printed text recognition. In the present work, we compare the efficacy of some feature-classifier combinations to carry out this separation task. We have selected length-normalized horizontal projection profile (HPP) as the starting point of such a separation task. This is with the assumption that the printed text blocks contain lines of text which generate HPP's with some regularity. Such an assumption has been demonstrated to be valid. Our features are the HPP and its two transformed versions, namely, eigen profiles and Fisher profiles. Different well known classifiers, such as, SVM's and artificial neural networks have been considered and their efficacy has been compared against some standard classifiers like nearest neighbor classifier. A sequential floating feature selection technique has been adopted to enhance the efficiency of this separation task. The results show a lot of promise with errors being as low as 3.5%.

Keywords: *Eigen profiles, Fisher profiles, horizontal projection profile.*

I. INTRODUCTION

Processing of document images to convert the contained text into editable form is a primary research interest of many researchers around the world. Analysis of a document involves some of the following tasks: (i) to convert a text document to editable text for its re-usability, (ii) convert a text document image to editable text and extract important information from it, (iii) separate the text and non-text elements of the image and interpret the nature of the text image either by analysis of the text or non-text elements, (iv) analyze the document image, either in parts or as a whole to extract information to enable efficient archiving and retrieval. However, the major research has gone into transformation of a document image into editable text as such a transformation has lots of scope for its use.

Many official documents consist of a multiple set of objects such as text (printed and handwritten), signature, bar-codes, logos and seals. An ideally drafted document would contain these objects as non-overlapping blocks. In this situation, it is easier to separate each of these blocks into separate classes of objects, before further processing to interpret them. For example, if a segmented block from a text image is known to be a scenic natural image, we could probably employ a scene analysis technique to interpret the nature of the document. If the block has a face image, we may employ a face recognition algorithm to see the person and the context of the document. In many situations, this would aid the job of text image interpretation.

Djeziri *et. al.* [1] have proposed a scheme for extraction of signatures from bank cheques. They have assumed that such documents have a lot of background patterns which they state causes a lot of difficulty for extraction of signatures. Based on a topological criterion specific to handwritten lines, which

they call filiformity, they reported good accuracy in separating the signatures out of such patterned background documents.

Guo and Ma [2] have proposed a scheme which combines the statistical variations in projection profiles with hidden markov models (HMM's) to separate the handwritten material from the machine printed text. They hypothesize that machine printed text has a large number of regularities on the projection profile which is missing in handwritten annotations owing to the variations in style, author and environment. In a similar approach, Kavallieratou and Stamatatos [3] try to take advantage of the structural properties that help humans discriminate printed from handwritten text. In their opinion, the height of the printed characters is more or less stable within a text-line while the distribution of the height of handwritten characters is quite diverse. These remarks stand also for the height of the main body of the characters as well as the height of both ascenders and descenders. Thus, the ratio of ascenders' height to main body's height and the ratio of descenders' height to main body's height would be stable in printed text and variable in handwriting.

Fan *et. al.* [4] propose a scheme for classification of machine-printed and handwritten texts using character block layout variance. In their approach, the orientation of a text block is first divided into horizontal or vertical direction by analyzing the widths of valleys of vertical (X) and horizontal (Y) projection profiles of a text block image. Then, a reduced X-Y cut algorithm is utilized to obtain the base blocks from a text block image. Finally, the spatial feature, character block layout variance, is devised to achieve the classification goal. They claim that this technique could be applied to English or Chinese document images.

Neumann *et. al.* [5] present a comparative study involving the local and global shape descriptors for logo classification. They use the negative shape method which is based on local shape information and a wavelet-based method which uses global information. Finally, they use these results to develop a new adaptive weighting scheme which is based on the relative performances of the two methods. They claim that this scheme is robust to all kinds of degradations.

Hobby [6] proposes a scheme to find signatures, text and graphic objects in document images by using shape and layout information. Pal and Chaudhuri [7] have used horizontal projection profiles for separating the printed and hand-written lines in Bangla script. Sabari *et. al.* [8] use Gabor function based filter-bank to classify the text elements against all other kinds of clutter. They claim the technique to be working well on camera captured images as well.

Thus, researchers have taken various approaches to segment and analyze a document page image for its content objects. Most of such reported schemes, either deal with segmenting an image into blocks of text and non-text or assume a fair segmentation scheme and take each of the segmented blocks

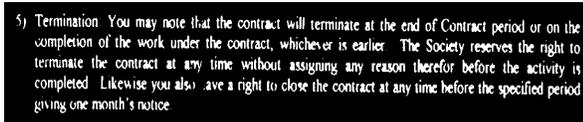


Fig. 1. Example of a typical text block.

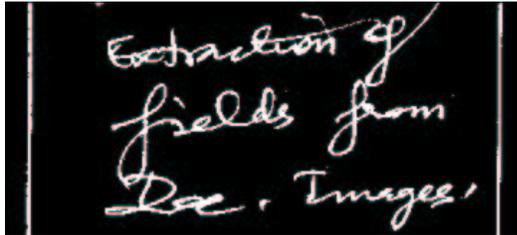


Fig. 2. Example of a typical non text block.

for its content analysis. Most of the latter case scenarios, segmentation of the image is accomplished manually. In the present work, we demonstrate the efficacy of various feature classifier combinations for separating text blocks from non-text blocks. Our non-text blocks consist of signatures, handwritten text, logos and other such objects. We have considered a block containing text and also some non-text elements as a text block. This is because we don't want to lose any printed text material. Typical examples of text and non-text blocks are shown in figures 1 and 2.

II. SYSTEM DESCRIPTION

We employ a transform based scheme for separation of text blocks from non-text blocks. However, in a case where both text and non-text elements are present, we consider this block as a text block. In the following sub-sections, we briefly describe the various features and classifiers we intend to employ for accomplishing the separation task and the results obtained by various combinations of them.

A. Building Block Segmentation

We start the process with the assumption that text blocks contain a few lines of text. When we consider the horizontal projection profile of such blocks, we see a repetitive pattern as demonstrated in the lower graph of Figure 3. In most cases of non-text blocks, this pattern is more of random in nature (see the upper graph in figure 3). Thus, we proceed to classify the text blocks from the non-text elements.

Our dataset consists of 100 document images, scanned at 200 dpi and stored in 1-bit depth, black and white text format. These documents contain signatures, logos and other such things along with free-flowing text paragraphs. Since our starting point is a set of document page images, containing both text and non-text elements, we divided the page image into smaller segments. These segments are typical paragraphs

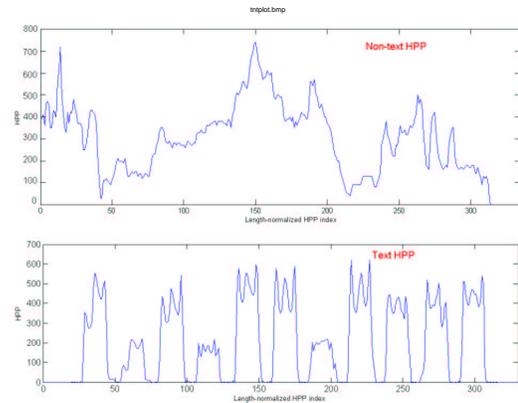


Fig. 3. Figure demonstrating the difference in nature of the HPP's of the text and non-text blocks.

and we call them building blocks, since they make up the page. This task is accomplished as mentioned below (refer figure 4):

- Run-length smooth the page document image both along the rows and columns. The parameter for this is selected such that the intra- and inter-character gaps, up to a paragraph level, are filled.
- Sometimes two building blocks join because of the presence of straight lines or other noisy elements. Mostly, the bridge between these building blocks are thin structures. We apply a morphological opening operation to separate these joints.
- Connected Component Analysis (CCA) segregates these building blocks into separate entities. Each segmented block is considered as an input for the text-separation scheme.

As mentioned above, any building block which contains a small amount of printed text in it is considered as a text-block.

B. Feature Vector Extraction

For separating the building blocks into text and non-text blocks, we extract a feature vector from each block. In this work, we have compared 3 different features that we have employed to achieve a goal. The features are (i) horizontal projection profile vector, (ii) eigen profile vector, and (iii) Fisher profile vector. We provide a brief description of each of these features in the following paragraphs.

1) *Horizontal projection profile (HPP)*: HPP of an image is a vector where each vector element contains the sum of the pixel values in the corresponding row.

$$HPP(i) = \sum_{j=0}^{No.of\ columns} I(i, j) \quad (1)$$

where $I(i, j)$ is the image. Since each building block has different number of rows, it generates HPP's of different dimension. We need to normalize the length of this feature vector to a pre-defined standard size, so that it could easily be used for classification. We have used bi-linear interpolation

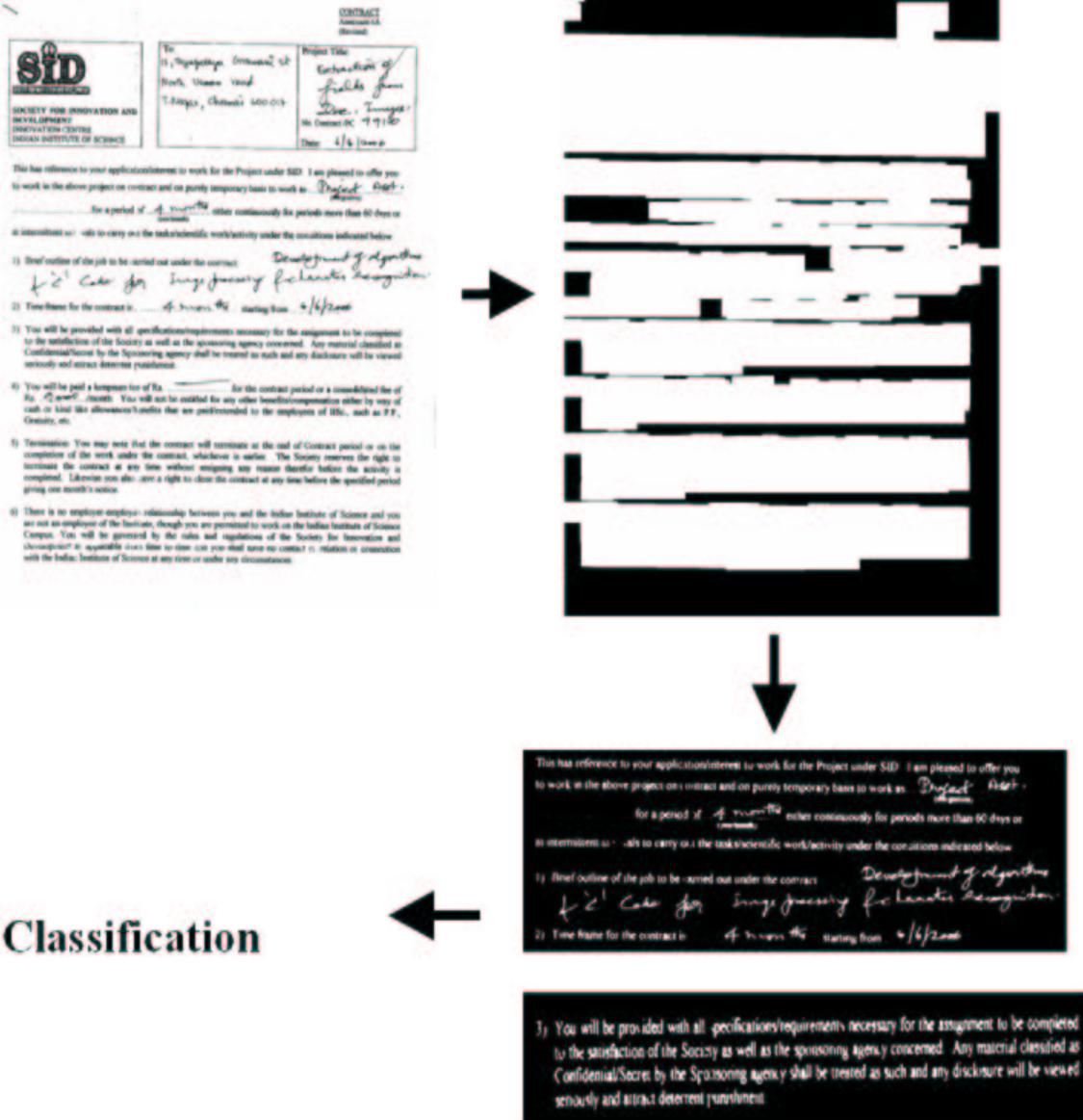


Fig. 4. Figure depicting the process of segmentation of the document image into blocks. First the document image is run-length smoothed. Then morphological operators are applied. From this image, the blocks are segmented using Connected Component Analysis. Finally these blocks are classified.

to get the HPP's to a standard length. The length of our normalized HPP is 330.

2) *Eigen Profile (EP)*: Turk and Pentland [9] have used K-L transform as a means of creating a linear subspace for face recognition. Here, they consider training samples of face images, each of the same dimension, $M \times N$. The rows of each face image are appended to make a MN dimensional vector. They consider each face as a point in this MN dimensional space. A linear subspace is created from this set of example images by taking the eigen vectors of the covariance matrix of the training set. This linear subspace is represented by these eigen vectors and they call such vectors as eigen faces.

We consider a training set consisting of equal number of text and non-text blocks. The length-normalized HPP's from this set are used to create a subspace, as mentioned above. We call these eigen vectors, representing the linear space, as eigen profiles. Each length-normalized HPP is then projected into this space and gets represented by a lower dimensional coefficient vector. We use this coefficient vector as our feature vector. For our experiments, we have taken 100 coefficients to represent each building block.

3) *Fisher Profile (FP)*: Belhumeur *et. al.* [10] proposed an extension of the work earlier proposed by Turk and Pentland [9]. They observed that the failure of the eigen faces, in case of lighting and pose variation, to yield superior quality discrimination is due to the large intra-class variability. The eigen faces, the eigen vectors of the covariance matrix, don't really discriminate between the intra- and inter-class variances. So, they proposed a generalized eigen vector solution to overcome this problem. They named these generalized eigen vectors, which are the eigen vectors pointing to the directions of large inter-class variance, as Fisher faces. The Fisher faces were also proven to yield better result in cases of large lighting, expression and pose variation. They yield no worse result if such variations are not there, since in such cases, the Fisher faces are none other than the eigen faces.

We have employed the scheme proposed by Belhumeur *et. al.* on our length-normalized HPP's and extract the generalized eigen vectors on our training set. We call them as the fisher profiles. Here, also we consider the feature vector containing 100 projected coefficients, as in the case of eigen profiles.

C. Classification

We have used the following classifiers to accomplish the separation task. They are: (i) nearest neighbor (NN) classifier, (ii) linear discriminant function (LDF), (iii) support vector machines (SVM's) and (iv) artificial neural networks (ANN).

Each building block is considered as a point in the feature space. Our training set is a distribution of points in the feature space. Each test feature vector is used for evaluating its distance from all the points, in the feature space, using the Euclidean metric. Using the Nearest neighbor (NN) classifier principle, we assign the class value of the training vector with minimum distance from the test vector, to the test building block.

We assign a class value of 1 to the text blocks and -1 to the non-text blocks. A linear discriminant function (LDF)

is a hyper-plane in the feature space which separates these two classes. A least-square solution approach is applied to evaluate this function from the training set. Each test vector is multiplied with this LDF to generate a class value. The test block is assigned a text-block label if this generated class value is found to be positive.

SVM's have been tested for their consistency in delivering classification results across various kinds of problems and datasets. We have tested two different topologies of SVM's, with RBF of degree 3 as the kernel function, to accomplish this text-block separation task. In the first case (SVM-I), the γ factor is considered to be 2^{-19} and the trade-off between training error & margin(C) is 1024. In the second case (SVM-II), the above parameter has been chosen as 2^{-18} and 512, respectively. These values have been chosen based on experimental evaluations over a wide range of values.

We also have explored the use of artificial neural networks (ANN's) as possible decision maker. Here again, we have explored the efficacy of two different topologies of ANN's. ANN-I is a linear neural network while ANN-II is a feed-forward 3 layer ANN with the hidden and the output layer containing 25 and 2 neurons, respectively. ANN-III is a Radial Basis network with spread being 1.5 for all three profiles.

D. Feature Selection

A feature vector consists of a number of features. These features could be statistical or syntactic features. We, generally, evaluate a number of features from our data without considering the classifiability of a feature. In a situation like this, a feature could have one of the following effects on the classifier:

- contributes positively to the classification process,
- contributes negatively, and
- doesn't do anything – sits idle.

If a feature contributes positively, it enhances the accuracy of classification. A feature with a negative impact, adversely affects the accuracy while a feature which sits idle neither enhances nor deteriorates the accuracy of the classification process. However, it is always a good practice to remove both the idle sitting and negatively impacting features from our feature set.

Pudil *et. al.* [11] have proposed a sequential floating feature selection scheme. This scheme has two approaches – (i) starting from a single feature, keep adding features till the optimal set is obtained, and (ii) starting from the whole set of features, go on removing the features till the optimal set is obtained. We have adopted the additive strategy from this scheme for our feature selection task.

III. EXPERIMENTAL RESULTS

In the first experiment, we tried to separate the text and non-text blocks using various combinations of features and classifiers. Here, we have used the HPP, EP and FP as the features. NN, LDC, ANN-I, ANN-II, ANN-III, SVM-I and SVM-II have been used as classifiers. It is, generally, observed that the length-normalized HPP's yield better result in the transform domain than otherwise. So, in the next experiment,

TABLE I

RECOGNITION ACCURACY FOR THE TEXT/NON-TEXT BLOCK SEPARATION FOR VARIOUS FEATURE-CLASSIFIER COMBINATIONS.

Method	HPP	EP	FP
LDF	61.5	68	69
NN	81	81.5	79.5
SVM-I	91	94	89.5
SVM-II	94	92	91
ANN-I	62	68	69
ANN-II	82	85	62
ANN-III	81.5	82	85

we tried to select those features from the EP's and FP's which yield us the maximum accuracy. Figure 5 demonstrates that adopting such a scheme rewarded us well. It can be noted that the accuracy of EP-NN combination rises to 96.5% after feature selection from 81.5% before. Besides, we achieve this improvement in accuracy with only 37 features in feature-reduced domain as against 100 features in the previous case. Examples of the classified and misclassified blocks are shown in figures 6(a), 6(b) and 6(c), 6(d) respectively. Similar studies on some other combinations also yield improvement in accuracy. The accuracies and the number of features for various feature-classifier combination are shown in tables II.

TABLE II

RECOGNITION ACCURACY WITH SELECTED FEATURES AND NUMBER OF FEATURES SELECTED IN THE FEATURE SELECTION PROCESS FOR VARIOUS FEATURE-CLASSIFIER COMBINATIONS.

Method	Accuracy		No of Features	
	EP	FP	EP	FP
LDF	91.5	84	10	85
NN	96.5	93	37	66

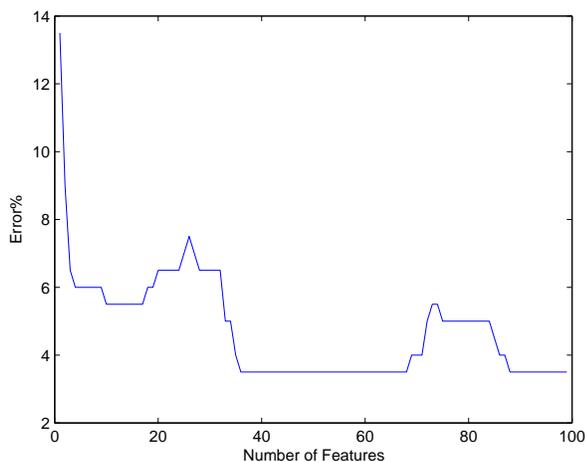


Fig. 5. Efficacy of sequential feature selection algorithm. The lowest error is achieved with 37 features only, further addition either doesn't improve the accuracy or increases error.

IV. CONCLUSION & DISCUSSION

In this paper we present a comparative study of text/non-text separation using various feature-classifier combinations. Some of these reported combinations perform with accuracy more than 90%. Our assumption that the text blocks have a repetitive and cyclic profile is justified. Given that there is little information about the non-text elements present in these blocks, the achieved accuracy can be considered a success of this experiment. However, further studies could only reveal the optimal feature-classifier combination for this task.

The results have disproved our anticipation about the performance of Fisher profiles. We, definitely, had expected better results from FP than from EP. The reason may be that there is not much intra-class variability in the dataset.

Figure 5 demonstrates the efficacy of the employed feature selection technique. It could be noted from this figure that the optimal set contains only 37 features out of the set of 100 features. Any further addition to this optimal set maintains the accuracy at the same level for some time and then reduces the accuracy. Thus our feature set not only has sit-idles but also many negative-impacting features.

The feature selection algorithm has definitely helped in improving not only the accuracy of separation but also the computational complexity. The selected set of 37 as against 100 features for EP-NN combination gives us much better accuracy and a huge saving in computation time. This study if evaluated for other classifiers as well, may provide us with better results.

REFERENCES

- [1] Salim Djaziri, F. Nouboud, and R. Plamondon, "Extraction of signatures from check background based on a filiformity criterion," *IEEE Transactions on Image Processing*, vol. 7, no. 10, pp. 1425–1438, 1998.
- [2] Jinhong K. Guo and Matthew Y. Ma, "Separating handwritten material from machine printed text using hidden markov models," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2001, pp. 439–443.
- [3] Ergina Kavallieratou and Stathis Stamatatos, "Discrimination of machine-printed from handwritten text using simple structural characteristics," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, 2004, pp. 437–440.
- [4] Kuo-Chin Fan, Liang-Shen Wang, and Yin-Tien Tu, "Classification of machine-printed and handwritten texts using character block layout variance," *Pattern Recognition*, vol. 31, no. 9, pp. 1275–1284, 1998.
- [5] Jan Neumann, Hanan Samet, and Aya Soffer, "Integration of local and global shape analysis for logo classification," *Pattern Recognition Letters*, vol. 23, no. 12, pp. 1449–1457, 2002.
- [6] J D Hobby, "Using shape and layout information to find signatures, text, and graphics," *Computer Vision and Image Understanding*, vol. 80, pp. 88–110, 2000.
- [7] U. Pal and B. B. Chaudhuri, "Automatic separation of machine-printed and hand-written text lines.," in *Proceedings of the International Conference on Document Analysis and Recognition*, 1999, pp. 645–648.
- [8] S Sabari Raju, P B Pati, and A G Ramakrishnan, "Gabor filter based block energy analysis for text extraction from digital document images," in *Proc. of the First Int. Workshop on Document Image Analysis for Libraries (DIAL'04)*, 2004, pp. 233–243.
- [9] M Turk and A Pentland, "Eigen faces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.
- [10] Peter N. Belhumeur, J. P. Hespanha, and David J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE transaction on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [11] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, 1994.

4) You will be paid a lumpsum fee of Rs _____ for the contract period or a consolidated fee of Rs 9000/month. You will not be entitled for any other benefits/compensation either by way of cash or kind like allowances/benefits that are paid/extended to the employees of IISc., such as P.F., Gratuity, etc

(a)

Extraction of mumbear
from printed text (seals).

(b)

 SID <small>SOCIETY FOR INNOVATION AND DEVELOPMENT</small> <small>INNOVATION CENTRE</small> <small>INDIAN INSTITUTE OF SCIENCE</small>	To 13, Thyagaraya Gramani st North Usman road T-Nagar, Chennai 600 017	CONTRACT Annexure 6A (Revised). Project Title: Extraction of fields from Doc. Images. No Contract/AC 9910 Date 6/6/2006

(c)

13, Thyagaraya Gramani st
North Usman road
T-Nagar, Chennai 600 017

(d)

Fig. 6. (a)Example of a correctly classified text block. (b)Example of a correctly classified non-text block. (c)Example of a mis-classified text block (d)Example of a mis-classified non-text block.