2011 International Conference on Document Analysis and Recognition

# Document Image Classification and Labeling using Multiple Instance Learning

Jayant Kumar Jaishanker Pillai David Doermann
*Institute of Advanced Computer Studies*
*University of Maryland College Park, USA*
{*jayant, jsp, doermann* }*@umiacs.umd.edu*

***Abstract*—The labeling of large sets of images for training or testing analysis systems can be a very costly and time-consuming process. Multiple instance learning (MIL) is a generalization of traditional supervised learning which relaxes the need for exact labels on training instances. Instead, the labels are required only for a set of instances known as *bags*. In this paper, we apply MIL to the retrieval and localization of signatures and the retrieval of images containing machine-printed text, and show that a gain of 15-20% in performance can be achieved over the supervised learning with weak-labeling. We also compare our approach to supervised learning with fully annotated training data and report a competitive accuracy for MIL. Using our experiments on real-world datasets, we show that MIL is a good alternative when the training data has only document-level annotation.**

***Keywords*-Document Image Labeling, Signature Detection, Machine-print Documents**

## I. Introduction

Search and retrieval of relevant documents from a large collection of document images has been a problem of interest for many years. One approach for solving this problem is to use *supervised learning* to train a classifier for detecting the documents of interest, using the features extracted from the image. Since images may contain multiple, possibly diverse document objects like logos, signatures, figures and tables, obtaining a consistent, global descriptor is difficult. Hence, *supervised learning* based approaches often segment the image into different zones and attempt to classify each zone. This requires the training images to be fully annotated. For example, for the retrieval of signature documents, a classifier is trained to classify each zone as *signature* or *non-signature*. This requires a set of training images with the label for each signature zone. Similarly, a multi-class labeling requires all zones to be labeled. For large datasets containing thousands of document images, the number of segments can be significant. It is extremely costly and time-consuming to manually label all the segments. Moreover, in the future, if one employs another segmentation method, then re-labeling of segments may be required.

Multiple Instance Learning(MIL) has been proposed as an alternative to *supervised learning* when the complete knowledge of the labels is not available [1]. In contrast to *supervised learning*, in MIL, labels are required only for groups of instances called *bags*. A document is represented by a *bag* of instances, where each instance is a descriptor for one of the regions after segmentation. In the binary classification case, a bag is assigned a positive label if at least one of the instances in it is from the positive class. It is labeled negative only if all the instances in it are negative. Figure 1 illustrates example of a signature detection problem. In this case an image with a signature in it contributes a positive bag, and an image with no signatures contributes a negative bag. We refer to this as *weak-labeling*. The goal of MIL is to learn models using this weakly-labeled data to classify test bags and/or instances (Figure 2).

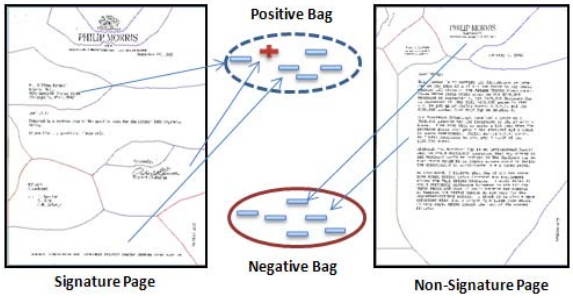


Figure 1. Segments of a Signature and a Non-Signature page from Tobacco-800 dataset using Voronoi++. Instances from a Signature page form a positive bag whereas instances from a Non-Signature page form a negative bag.

In this paper we apply MIL for two different document labeling problems, and show that MIL is a good alternative when the exact labels of training instances (zones) are not available. We first employ different MIL approaches for signature detection and compare it with Support Vector Machines(SVM) [8]. We report results on Tobacco-800 dataset [15] which has evolved as a standard dataset for signature detection and has been used in previous work [4], [5]. We show that MIL achieves almost the same accuracy for signature zone detection as obtained by *supervised learning* (using zone-level annotation). We then perform a similar comparison for the detection of machine-printed text on a set Arabic document images. We use shape-codebook based features to demonstrate that MIL is competitive with supervised learning. A gain of 15-20% over the *weakly-labeled supervised classification* is obtained for both signature and machine-print text detection.

The remainder of the paper is organized as follows. We

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.214

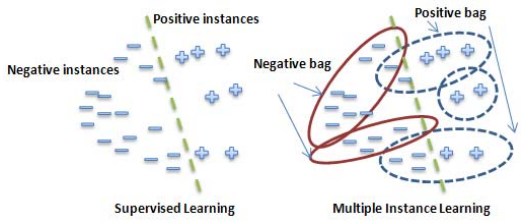


Figure 2. Comparison of supervised learning and multi-instance learning. The classifier is learned over bags instead of instances in MIL.

briefly review the selected MIL algorithms and applications in Section II. Section III discusses the labeling of signature and machine-print documents using the MIL framework. Section IV describes our experimental results and we conclude the paper in Section V.

## II. Previous Work

In the last few years, many algorithms have been proposed to learn models in a multiple instance setting. Some algorithms are specifically designed to learn the multiple-instance concepts whereas others are adapted from standard single-instance learning algorithms. A detailed survey of the various algorithms for MIL is given in [3]. In this section, we briefly review a few methods and applications.

The first work on MIL was done for drug activity prediction [1]. The problem was to determine if a given drug molecule will strongly bind to a target protein. Dietterich *et al.* showed that the MIL approach outperforms the *supervised learning* method which does not take into account the multiple-instance nature of the problem. Another interesting application of MIL is in content-based image retrieval and categorization [2]. Natural scene images usually contain multiple objects and classifying image based on a global description is often difficult. However, this naturally fits into the MIL setting where each image can be considered as a bag and segments in the image as instances. Andrews *et al.* [2] reformulated the SVM for MIL and presented impressive results for the detection of a tiger, an elephant and a fox in images. They introduced two approaches namely *mi-SVM* and *MI-SVM*, where *mi-SVM* is for instance level classification and *MI-SVM* is for bag level classification. In *mi-SVM*, the instance labels $y_i$ are considered hidden variables subject to constraints imposed by the bag labels $Y_I$. The goal is to maximize the instance margin jointly over the unknown instance labels and the kernel parameters. Hence the same formulation is used as SVM, but the minimization is done over the individual labels as well, subject to the constraint that in a positive bag, at least one instant should have a positive label and all instances should have negative labels in a negative bag. In MI-SVM, the parameters of the model are obtained by maximizing the bag margin which is defined as the margin of the most positive instance in a positive bag and the least negative instance in a negative bag.

Diverse density (DD), proposed by Maron and Lozano-Perez [6], is one of the best known frameworks for MIL. The purpose of this approach is to learn a concept that is close to at least one instance in each positive bag, but far from all instances in the negative bags. The hyperplane learned describes a region of instance space that is not only *dense* in instances from positive bags, but also *diverse* in that it describes every positive bag. MIDD maximizes the bag-level likelihood using the *noisy-or* model at training time. The label of a new bag is the class that receives maximum probability. Ray *et al.* [13] designed *Multiple Instance Logistic Regression* (MILR) to learn linear models in an MI setting, which is derived by generalizing DD framework.

## III. Document Image Labeling

### A. Signatures

Signatures provide a unique way of indexing a large set of forensic and business documents [5]. Searching documents containing signatures is pivotal for signature based indexing and retrieval. We pose the problem of detecting signatures in a large collection of document images as a MIL problem, where a positive bag consists of instances extracted from the zones of images containing signatures. Zones of document images which do not contain any signature contribute to negative bags. The classifier is trained to classify a given image into a positive or a negative bag. In each bag, the labels of instances are also obtained and a positive instance label corresponds to a signature zone.

We segment the training images into different regions using a *Voronoi* based segmentation [7]. Our assumption is that the signatures are segmented to one of the regions although exact segmentation is not required. Some documents may have multiple signatures in them. The segmentation boundary shown in Figure 1 is obtained using this method.

**Chain-code based Histogram Features**: We generate chain-code features at the zone level after a simple preprocessing based on size, aspect-ratio and mass of components to filter noise-components. The gradient at each pixel of the edge-detected image is computed using a Sobel operator (Figure 3(a)). We then use eight bins for the gradient direction and a 3x3 sampling region to obtain a histogram at each gradient direction. This gives us a total of 72 features for each component (Figure 3(b)). Using SimpleKMeans clustering method available in Weka [14], we cluster the extracted features from a set of representative zones and find an exemplar for each cluster (Figure 3(c)). This provides us with codewords, which we use to obtain histogram features for each zone. In a given zone, we find the chain-code of each connected-component and obtain the corresponding codeword closest to it based on the Kullback–Leibler divergence. We then compute the frequency of each

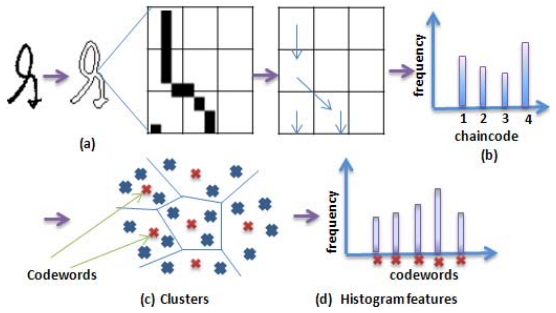


Figure 3. (a) Edge-map of a character and illustration of orientation (b) Histogram of orientation(c) Construction of a codebook (d) Histogram features for signature detection

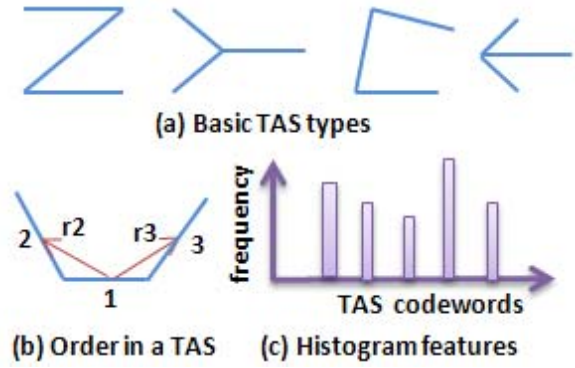


Figure 4. (a) Four basic TAS types (b) Ordering in a TAS (c) Histogram features using learned codewords

codeword to obtain a normalized histogram (Figure 3(d)). The main motivation behind using gradient features is that signatures are written in a free-flowing manner, unlike other content such as graphics or printed-text. Similar features were used by Chanda et al. [9] for sparse machine-print and handwritten text classification. We also use the mean and the variance of the width, height and area of components in the zone as additional features.

### *B. Machine-printed Text*

Since the pre-processing and character recognition techniques may be different for machine-print and handwritten text, it is often necessary to identify the two types of text before feeding them to their respective processing systems. We formulate the problem of detecting printed pages among a large set of handwritten and printed documents as multi-instance learning, where instances coming from the printed pages form positive bags and the remaining pages give rise to negative bags.

**Shape-Codebook Based Features**: As with signatures, we first extract zones present in the image using *Vornoi* based segmentation [7], then obtain a list of edges present in each zone using a Canny edge detector [12]. Within a specified tolerance, we find a similar list of line segments by fitting a line to each edge segment. Every triplet within each connected-component forms one of the four basic Three-Adjacent-Segment (TAS) types defined as shown in Figure 4(a). The first segment ($s_1$) is the one with the midpoint closest to the centroid. The second and third segments are ordered from left to right. An example ordering of a typical TAS is shown in Figure 4(b). The descriptor of a TAS is composed of 10 values given as follows:

$$(\frac{r_2^x}{N_d},\ \frac{r_2^y}{N_d},\ \frac{r_3^x}{N_d},\ \frac{r_3^y}{N_d},\ \theta_1,\ \theta_2,\ \theta_3,\ l_1,\ l_2,\ l_3)$$

where $\mathbf{r}_i = (r_i^x, r_i^y)$ denotes the vector from the midpoint of $s_1$ to the midpoint of $s_i$. $\theta_i$ and $l_i$ represent the orientation and length of the segment $s_i$. $N_d$ is the distance between the two farthest midpoints and used as a normalization factor. The descriptor is rotation and scale invariant. These local shape features were first proposed by Ferrari *et al.* for object detection [10] and have been shown effective for language identification and printed page detection [11].

We select a subset of printed and handwritten text zones for creating a shape codebook for both printed and handwritten Arabic. Using SimpleKMeans [14], we cluster the TAS features extracted from the zones and obtain exemplary codeword in each cluster. We also find the cluster radius which is defined as the maximum distance from the exemplary codeword to all the other TASs within the cluster. Finally, we compute a descriptor for each zone that provides statistics of the frequency of each TAS feature occurrence. We increment the number of occurrences of the codeword which is nearest to detected TAS feature and within corresponding cluster radius. We concatenate the two normalized histograms obtained using printed and handwritten codebooks to obtain a single feature vector for each zone (Figure 4(c)).

## IV. Experiments and Evaluation

**Datasets**: For signature experiments, we used a randomly selected subset of 600 images from the Tobacco-800 dataset available at [15]. This dataset has mainly multi-page business documents many of which contain signatures. For machine-print text detection we used a subset of 800 Arabic document images from the dataset used in [11]. Images in this dataset have primarily printed and handwritten Arabic, in addition to logos, figures, signatures and stamps. We partitioned the datasets into a training set (70%) and a test set (30%).

**Metrics**: For evaluation, we compute *precision* and *recall* values along with the *F1-score* to evaluate and compare MIL methods. If a detected zone has the same label as its ground-truth then it is counted as a *true positive* (TP), otherwise it is counted as a *false positive* (FP). *False negatives* (FN) are

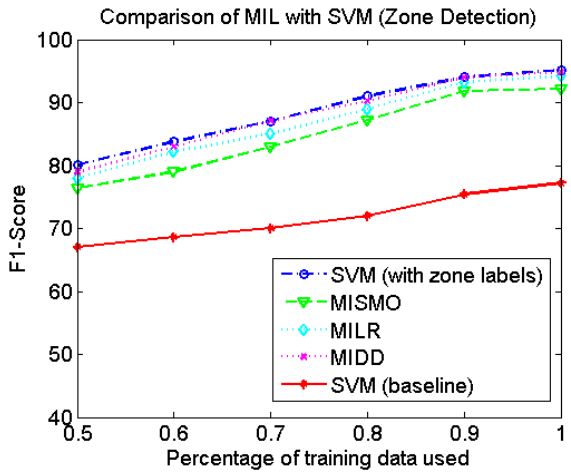


Figure 5. Plots of F1-scores for signature zone detection

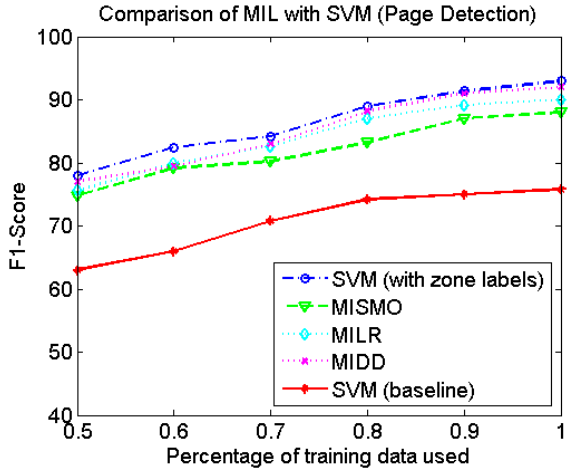


Figure 6. Plots of F1-scores for signature page detection

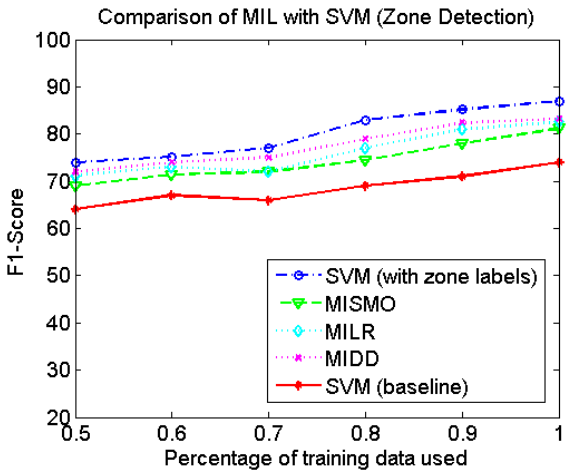


Figure 7. Plots of F1-score for machine-print zone detection

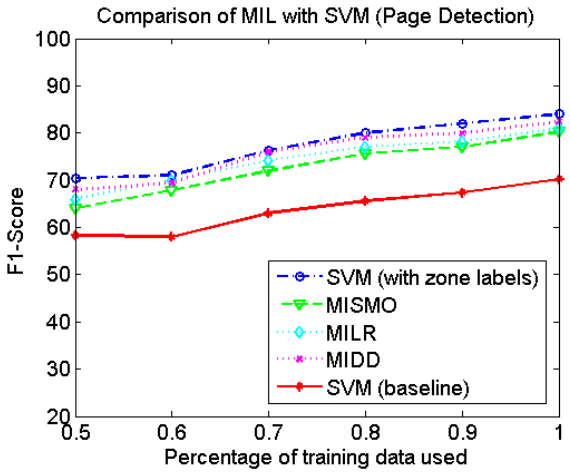


Figure 8. Plots of F1-score for machine-print page detection

those zones which are missed by the method. Using these counts we obtain the *precision*, *recall* and *F1-score* using the following equations:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{1}$$

$$F_1 score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2}$$

We used the Weka data mining toolbox [14] to compare various MIL algorithms with SVM. In all our experiments, we used the default polynomial kernel for SVM. Of the several MIL implementations available in the package, we used three methods: *MISMO*, *MIDD* and *MILR*. Figure 5 and Figure 6 show the plot of F1-scores of these methods for signature detection. For comparison, we trained two SVM classifiers for zone classification, one with the correct labels of zones and another with the label of its document(*signature* or *non-signature*). The first classifier uses the *correct* labels of zones in the training set to learn a model for zone-classification, whereas second classifier uses only the document-level annotation. During evaluation, if any of the instances in the test image was classified as *signature*, the image was considered as a signature page. If all the instances were classified as *non-signature*, the page was given a non-signature label. A similar protocol was used for obtaining the supervised learning performances of machine-print page detection. The plot of F1-scores of both SVM classifiers is shown for comparison.

MIDD gives the best performance among all of the MIL classifiers, with a precision of 96.4% and a recall of 94.2% for signature zone detection. As shown, the accuracy achieved by MIDD is very close to the accuracy of SVM using correct zone labels (96%). A similar plot for machine-

Table I
TIME (IN SECONDS) FOR MIL METHODS AND SVM (SMO)

| Methods | MISMO | MIDD | MILR | SVM |
|---|---|---|---|---|
| Training Time | 31.2 | 262.6 | 2.77 | 2.2 |
| Testing Time | 2.1 | 4.6 | 1.2 | 1.4 |

print documents is shown in Figure 7 and Figure 8. In this case, MIDD also achieved the best F1-score of 83.2%, while the accuracy obtained by SVM using zone labels was 86%. For machine-print page detection, we obtained a high precision (94.2%) and a low recall (74%) using MIDD, which is consistent with the supervised learning result reported on this dataset [11]. Upon inspection of error images, we found that the poor recall is due to those documents, where the printed-text content is limited. These images were missed by both MIL and supervised classifiers. The performance of SVM zone-classification with only document-level annotation is poor in all the cases. This is expected, as the classifier did not have exact labels for the zones in signature-pages. Compared to this *weakly-labeled supervised learning*, MIL achieves a gain of 15-20%. The higher accuracy of zone classification as compared to document classification is attributed to the MIL setting, in which for a document to be classified correctly the condition needs to be satisfied exactly. For example, for a non-signature page all the segments need to be classified as negative instance.

Table I shows the average time taken (in seconds) by each MIL method and SVM on the training and testing data. Although, MIL methods optimize over both bag and instance labels, the training time is comparable to supervised learning. MIDD, which gave the best performance for classification, has much higher training time compared to other methods, but the testing time is comparable. MILR shows a good performance for both time and detection accuracy.

## V. CONCLUSION

In this work, we demonstrated the advantages of using Multiple instance learning for two document image classification problems. Although the multiple-instance setting leads to a harder optimization problem, even simple approaches for solving it offer competitive results when compared to the *supervised learning*. To see the effectiveness of MIL, we considered problems of different nature and tried different features on real-world data sets. In both the applications, we found that MIL provides a gain of 15-20% over the *weakly-labeled supervised classification*. The results are competitive to those obtained from supervised learning with fully-annotated data. In the future, we would like to apply MIL for logo and stamp detection.


## ACKNOWLEDGMENT

The partial support of this research by DARPA through BBN/DARPA Award HR0011-08-C-0004 under subcontract 9500009235, the US Government through NSF Award IIS-0812111 is gratefully acknowledged.


## REFERENCES


[1] T. G. Dietterich, R. H. Lathrop and T. Lozano-Perez, *Solving the Multiple Instance Problem with Axis-parallel Rectangles*, Artificial Intelligence, vol. 89(1–2), pp. 31–71, 1997

[2] S. Andrews, I. Tsochantaridis and T. Hofmann, *Support Vector Machines for Multiple-instance Learning*, NIPS, 2003, pp. 561–568, MIT Press.

[3] Z. H. Zhou, *Multi-Instance Learning: A Survey*, Technical Report, AI Lab, Nanjing University, China, 2004.

[4] S. Srihari, S. Shetty, S. Chen, H. Srinivasan, C. Huang, G. Agam and O. Frieder, *Document Image Retrieval Using Signatures as Queries*, ICDIAL, pp. 198-203, 2006.

[5] G. Zhu, Y. Zheng, D. Doermann and S. Jaeger, *Multi-scale Structural Saliency for Signature Detection*, CVPR, pp. 1–8, 2007.

[6] O. Maron and T. Lozano-Perez, *A Framework For Multiple-Instance Learning*, NIPS, pp. 570–576, 1998, MIT Press.

[7] M. Agrawal and D. Doermann, *Voronoi++: A Dynamic Page Segmentation Approach based on Voronoi and Docstrum Features*, ICDAR, pp. 1011–1015, 2009.

[8] C. Cortes and V. N. Vapnik, *Support Vector Networks.* Machine Learning, **20**, pp. 273–297, 1995.

[9] S. Chanda, K. Franke and U. Pal, *Structural Handwritten and Machine Print Classification for Sparse Content and arbitrary oriented Document Fragments*, ACM Symposium on Applied Computing, pp. 18–22, 2010.

[10] V. Ferrari, L. Fevrier, F. J. and C. Schmid, *Groups of Adjacent Contour Segments for Object Detection*, IEEE Trans. on PAMI, 30, pp. 36–51, 2008.

[11] J. Kumar, R. Prasad, H. Cao, W. Abd-Almageed, D. Doermann and P. Natarajan, *Shape Codebook based Handwritten and Machine Printed Text Zone Extraction*, DRR, vol:7874(06), pp. 1–8, 2011

[12] J. Canny, *A Computational Approach to Edge detection*, IEEE Trans. on PAMI, 8(6), pp. 679–697, 1986.

[13] S. Ray and M. Craven, *Supervised versus Multiple Instance Learning: An Emprical Comparison*, ICML, pp. 697–704, 2005.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, *The WEKA data mining software: an update* SIGKDD Explor. Newsl., pp. 10–18, ACM volume 11, 2009.

[15] *Tobacco-800 Signatures and Logos Dataset*, Laboratory for Language and Media Processing (LAMP), University of Maryland, http://lamp.cfar.umd.edu, 2010