# Learning Document Structure for Retrieval and Classification

Jayant Kumar  Peng Ye  David Doermann
*University of Maryland, College Park, USA*
{*jayant, pengye, doermann*}*@umiacs.umd.edu*

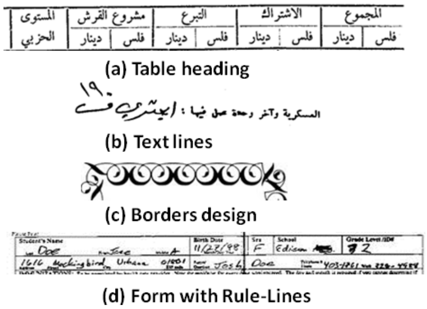
## Abstract

*In this paper, we present a method for the retrieval of document images with chosen layout characteristics. The proposed method is based on statistics of patch-codewords over different regions of image. We begin with a set of wanted and a random set of unwanted images representative of a large heterogeneous collection. We then use raw-image patches extracted from the unlabeled images to learn a codebook. To model the spatial relationships between patches, the image is recursively partitioned horizontally and vertically, and a histogram of patch-codewords is computed in each partition. The resulting set of features give a high precision and recall for the retrieval of hand-drawn and machine-print table-documents, and unconstrained mixed form-type documents, when trained using a random forest classifier. We compare our method to the spatial-pyramid method, and show that the proposed approach for learning layout characteristics is competitive for document images.*

## 1. Introduction

The problem of retrieving structurally similar document images from a large heterogenous collection given a few *relevant* images, has been of interest for many years [6, 10]. Although methods have been developed for layout-specific or content-specific document image retrieval [6, 10], a general approach which can detect salient structures and co-occurrence relationships among different regions of a document image automatically, is still being researched [6]. Content-based approaches are highly dependent and sensitive to the quality of optical character recognition (OCR) or component labeling classifiers. In cases of handwritten documents, these approaches may not be applicable since OCR for unconstrained handwritten documents is still a difficult problem.

Structural-based similarity features have been shown to enhance the capabilities of content-based matching, and often provide an effective way to reduce the set of candidate documents for matching. However, mining layout structure (e.g., location and extent of components, spatial relationships among the components) in unconstrained and noisy documents has been difficult due to variation in content, translation, rotation and scale of components. Moreover, the meaningful information in a document's layout is often hidden in the global structure of the document page.


المستوى الحزبي
مشروع القرش
التبرع
الاشتراك
المجموع
دينار
فلس

(a) Table heading

(b) Text lines

(c) Borders design

(d) Form with Rule-Lines

**Figure 1. Examples of document objects**

Another challenge in formulating the retrieval problem in this way is imbalance in training data. Often, the number of relevant documents provided for retrieval is much lower than the number of irrelevant documents causing an *imbalance* in the training data. Many learning methods suffer from *imbalance* data problem [9].

In this work, we propose to use statistics of raw-image-patches in different partitions of a document image to capture the structural relationships. It is often not clear what features are best suited for monochromatic document images. To address this issue, we explore unsupervised feature learning and use raw-image patches to construct a codebook representative of basic structural elements in document images. The structure and layout of document objects are often restricted to a specific direction to improve the readability. For example, text-lines are often written from right-to-left or vice versa, tables have horizontal and vertical lines aligned with borders, border-designs run across both horizontal

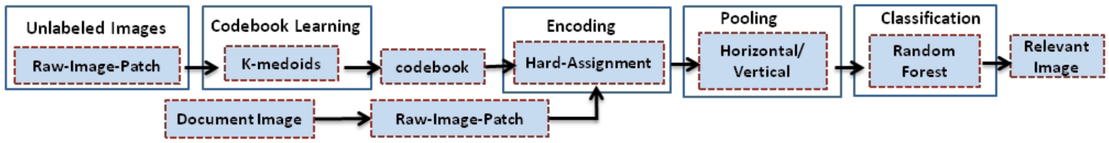


Figure 2. Block diagram of proposed document structure learning framework

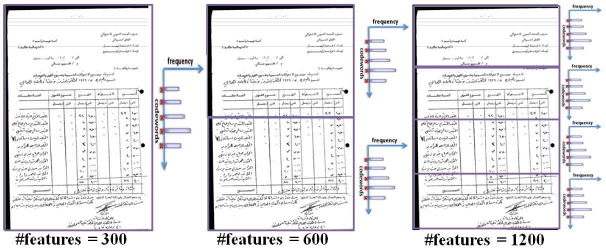


Figure 3. Pooling in horizontal direction.

and vertical directions (Figure 1). To capture spatial relationships, we recursively divide the document image horizontally and vertically, and compute histograms of *learned* codewords in these regions. We show that this strategy of modeling spatial relationship gives a high retrieval accuracy using random forest (RF), even when the data is imbalanced. We compare our strategy with the spatial-pyramid method [4] and show that the proposed method gives superior performance on two document retrieval tasks.

## 2. Related Work

The *Bag-of-words* approaches have recently shown impressive results on many computer vision tasks such as image categorization [7, 8], scene understanding [4], and document image classification [3]. However, these methods disregard the spatial relationship between codewords, and only count the occurrences of each codeword in an image. This leads to a limited descriptive ability of these methods and performance degrades in presence of noise and background clutter, variation of layout and content in images. Subsequently, methods which model the spatial relationships between visual codewords have been proposed. One of the popular methods proposes the creation of *spatial-pyramid* features by partitioning the image into increasingly finer grids and computing the weighted histogram of features in each region for scene classification [4]. Selecting the optimal feature pooling strategy and efficient ways to learn these local statistics are important research problems, and a number of methods have been proposed [7, 8].

Our approach differs from previous approaches in several ways: (1) we apply unsupervised feature learning to obtain a dictionary of representative structural elements of document objects, (2) we use a horizontal-vertical partitioning scheme for learning spatial relationships, and (3) we compare RF with Support vector machines (SVM) and show that it is competitive for this problem even when the data is imbalanced.

## 3. Proposed Method

A block-diagram of proposed method is shown in Figure 2. We describe each of these components in detail in the following sub-sections.

### 3.1 Image-patch based Codebook

We first randomly extract patches of dimension $B \times B$ from a set of representative document images. Using the K-medoids method, we obtain a set of exemplary codewords which represent the basic structural elements present in document image collection.

### 3.2 Feature Encoding and Pooling

We now describe the main contribution of this paper, *horizontal-vertical pooling*. We uniformly and randomly extract patches from images, and for each patch we find the nearest (L1-norm) codeword in the codebook. The image is recursively partitioned into horizontal and vertical halves (as shown in Figure 3 for the vertical direction). For each partition we compute a normalized histogram of codewords to capture the spatial statistics of codewords in that region. The weighted histograms obtained from all regions are used for training a RF Classifier. The number of features (N) using our approach is:

$$N \quad = \quad 2\sum_{l=0}^{L}\sum_{k=1}^{2^l}|C| \tag{1}$$

where $|C|$ is the number of codewords and L is the number of levels of partitioning. The constant 2 accounts for both the horizontal and vertical partitioning. If $(h, w)$ represents the dimension of documents, and $h \geq 1.2 \times w$, one additional partitioning is performed for $h$. Note that the *spatial-pyramid* (SP) approach partitions the image into four parts irrespective of its dimensions, while in our case the partitioning is performed at one more level in the dimension with higher value. Since the features per level in the SP method grows faster ($O(4^l)$) than the proposed method, even

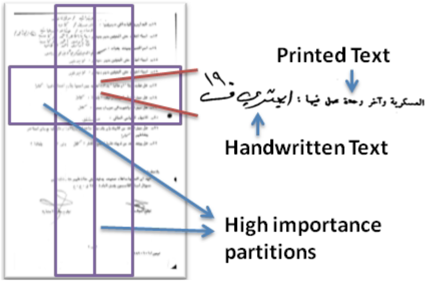


**Figure 4. Samples of mixed-forms**

with one additional level of partitioning we have the same number of features (6300). Additionally, using *importance plots* (explained in Section 3.3) from the RF we obtain a list of partitions crucial for classification, and do not compute features for unimportant regions. While the SP method normalizes histograms with the total weight of all features, in our method it is not required since we extract equal number of patches from each image.

### 3.3 Random Forest Classifier

In this work, we use the random forest (RF) classifier [1] for document image classification. The RF is an ensemble-based learning algorithm which constructs a set of tree-based classifiers, and then classifies new data points by taking a vote of the predictions of each classifier. There are multiple reasons to select the RF over other classifiers for this problem. The RF has been shown to work well when many features (on the order of thousands) are available. It does not over-fit with the increase in number of features and increases diversity among the classifiers by resampling the data, and by changing the feature sets over the different classifiers (trees). Random selection of features to split each node makes it more robust to noisy data. The *variable importance* plots obtained from the training stage can be used to analyze which regions and features are important for classification. The importance of a variable is estimated by looking at how the prediction accuracy decreases when out-of-bag (OOB) data for that variable is permuted while all others are left unchanged [1]. Using the *importance* estimates of variables in a particular region, we skip *unimportant* partitions, and do not compute features over those regions. This results in computational efficiency, and in some cases, better performance.

## 4. Experimental Results

**Datasets and Evaluation:** We experimented with two datasets: (a) a collection of document images containing hand-drawn and printed tables as relevant documents (Figure 3), and (b) a set of documents containing *mixed-form type* Arabic document images as relevant documents (Figure 4). Both datasets were created from a large collection of field data of Arabic documents. The difficulty for retrieval arises from high variation in structures and sizes of tables and form-layouts. We manually annotated a set of 618 images to obtain 216 table images. From another set of 793 images, we obtained 402 mixed-form images. Irrelevant images typically consisted of handwritten documents with signatures, logos, printed Arabic text and face images. We extracted 40,000 patches from another set of 30 images to obtain 300 codewords. In our experiments, we did not observe any significant increase in performance beyond $|C| = 300$.

**Table 1. Classification accuracy(%)**

| | SP-SVM | SP-RF | HVP-SVM | HVP-RF |
|---|---|---|---|---|
| Table | 92.2 | 95.4 | 93.2 | **97.4** |
| Forms | 96.6 | 97.8 | 97.6 | **98.9** |

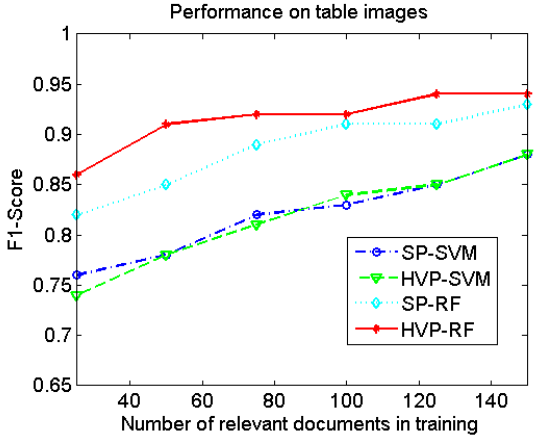


**Figure 5. F-1 scores for table images**

We evaluated and compared our pooling method against the spatial-pyramid method of Lazebnik *et. al* [4]. To test the effectiveness of RF classifier on unbalanced data, we compared our approach against SVM available with LibSVM package [2]. For evaluation we used the standard precision and recall measures to compute F1-measures for each method. Five-fold cross validation was used to obtain the median accuracies in all cases.

**Results and Discussions:** Table 1 shows the classification accuracy of spatial-pyramid (SP) and Horizontal-Vertical pooling (HVP) with SVM and RF. We uniformly and randomly extracted 5000 patches of size 100 × 100 from each image. The weights used for features at different levels were 0.125(l=0), 0.125(l=1),

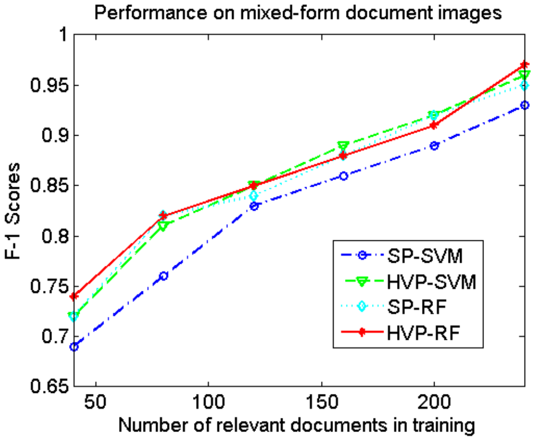


**Figure 6. F-1 scores for mixed-form images**

0.25(l=2), 0.5(l=3). Additional level (l=3) of partitioning was performed only in vertical direction. For RF, there are two main parameters:(1) number of trees (*nTree*), and (2) number of attributes selected for each tree(*mTry*). We used the standard values of $\lfloor \sqrt[2]{N} \rfloor$, where N is the number of features for *mTry* and *nTree* = 1000, suggested in previous work [5]. As suggested in [4], the weights used for SP were $\frac{1}{2^{L-l}}$ for $l = 0, 1, 2$. As shown, the proposed method outperforms the other three. Figure 4 shows the partitions with high-importance variable values. As illustrated, the structural pattern *handwritten and machine-print text-line side-by-side* which is crucial for classification is more probable in these regions.

We also evaluated performance of the proposed method on imbalanced data. For first dataset, the number of relevant documents in training data was varied from 25 to 150 in steps of 25. A fixed set of 250 irrelevant images was used to create the imbalance (from 1:10 to 6:10). Another set of 132 documents containing 66 relevant documents and 66 irrelevant images was used for testing. For *mixed-forms*, we used a total of 240 relevant documents (40 to 240 in steps of 40) and 320 irrelevant images for training. Figure 5 and Figure 6 show the F1-scores with increasing number of relevant documents in training data.

As seen, the proposed method achieves a high F1-score for retrieval with as few as 50 relevant documents. We observe that SP-RF combination performs competitively on the two datasets, and the additional gain achieved by the proposed method is primarily due to a different strategy for feature pooling and overlapping partitions. The best performance for SVM was achieved using *linear* kernel and shown in plots. This is due to the high dimensionality of feature space. RF on the other hand achieved high performance even when the data was highly imbalanced.

## 5. Conclusion

We presented a bag-of-words model based method for document image retrieval using a feature pooling strategy which aids in capturing spatial relationship between different image regions. Results on samples from real-world field data demonstrates that pooling strategy plays an important role in image classification. We compared RF with SVM and showed that RF is more suitable when the number of features is large.

### 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] L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[3] J. Kumar, R. Prasad, H. Cao, W. Abd-Almageed, D. Doermann, and P. Natarajan. Shape Codebook based Handwritten and Machine Printed Text Zone Extraction. In *Document Recognition and Retrieval*, pages 7874:1–8, 2011.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, pages 2169 – 2178, 2006.

[5] A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[6] C. Shin and D. Doermann. Document image retrieval based on layout structural similarity. In *Intl. Conf. on Image Processing, Computer Vision and Pattern Recognition*, pages 606 – 612, 2006.

[7] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, IEEE Conf. on*, pages 1794 –1801, 2009.

[8] Y. Yang and S. Newsam. Spatial pyramid co-occurrence for image classification. In *Computer Vision (ICCV), International Conference on*, pages 1465 –1472, 2011.

[9] Z. Zheng, X. Wu, and R. Srihari. Feature selection for text categorization on imbalanced data. *SIGKDD Explor. Newsl.*, 6(1):80–89, June 2004.

[10] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. Signature detection and matching for document image retrieval. *PAMI, IEEE Tran. on*, 31(11):2015 –2031, 2009.