

# Image classification: Classifying distributions of visual features

Prateek Sarkar

Perceptual Document Analysis  
Palo Alto Research Center, Palo Alto, California  
psarkar@parc.com

## Abstract

We classify an image by generating a list of salient visual features present in the luminance channel, and matching the resulting variable-length feature list to category-specific generative models for such features. To facilitate quick computation, we use thresholded Viola-Jones rectangular features, each represented by a five-dimensional descriptor. For each image category, a probability distribution for feature-lists is given by a latent conditional independence (LCI) model and classification is maximum likelihood. On the NIST tax forms database [3], where intra-category variations include variable scan-lightness, skew, noise, and machine-printed form-filling, our method improves performance over published results, while requiring very little training data, and without relying on an extensive set of handcrafted features.

## 1 Introduction

We address the problem of image classification, *i.e.*, analysing an image to assign one of several pre-defined category labels to it. In the context of document image analysis, image classification/categorization serves a role in indexing and retrieval, sorting and organization, as well as in tailoring automated analysis tasks to specific document types. This is particularly useful in high volume scan conversion operations where considerable effort is spent on visual or functional categorization of documents.

Most published document image classification systems rely on either OCR, so that documents can be categorized based on textual content, or on some form of layout analysis to produce a layout signature that can be compared to category-prototypes [4, 7]. Methods for categorization based on word-shape codes have also been reported. Shin et al. [9] propose an extensive set of features to characterize segmented documents by their appearance, and apply discriminative (decision tree) classifiers trained on thousands

of samples to achieve low error rates. In contrast we use a generative paradigm, and do not depend on document segmentation or handcrafted features to achieve lower error rates. The features are simple and robustness is achieved through redundancy because numerous features are generated. Diligenti et al. [2] report a very interesting machine learning approach based on *Hidden Tree Markov Models (HTMM)*, but do not report results on a public dataset.

We present our notation and statistical models in Section 2, briefly describe the Viola-Jones rectangular features in Section 3, and report experiments and results, along with illustrations in Section 4.

## 2 The Latent Conditional Independence model

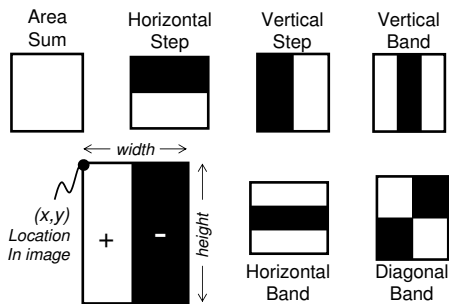
Each image to be classified is represented by a list of salient features  $[x_1, x_2, \dots, x_N]$ . Each element  $x_n$  is of the same type  $\mathcal{X}$ , *i.e.*,  $x_n \in \mathcal{X}$ . The specific type of features used in our experiments will be described in Section 3. The length,  $N$ , of this list may vary from image to image. The classification problem is to assign each feature-list to one of  $C$  pre-specified categories in the set  $\mathcal{C} = \{1, 2, \dots, C\}$ .

We classify by maximizing likelihood. The category  $c_{ML}$  of an observed image is computed as:

$$c_{ML} = \arg \max_{c \in \mathcal{C}} p(x_1, \dots, x_N | c) = \arg \max_{c \in \mathcal{C}} \prod_{n=1}^N p(x_n | c) \quad (1)$$

The essential distinction between different image categories is provided by the underlying category conditional feature-likelihood function  $p(x_n | c)$ . Each feature  $x_n$  is itself a  $D$ -dimensional vector of measurements or attributes  $[x_{n1}, x_{n2}, \dots, x_{nD}]$ . For each category, the likelihood function,  $p(x_n | c)$ , is modeled as a mixture of  $K$  components. Within each component, the feature dimensions are

**Figure 1. Viola-Jones rectangular features computed for document image classification.**



conditionally independent:

$$p(x_n|c) = \sum_{k=1}^K p_k(c) \prod_{d=1}^D p_k(x_{nd}|c) \quad (2)$$

This is a *Latent Conditional Independence (LCI)* model, or a mixture of *Naive Bayes* models. One advantage of such a model is that it can express joint distributions over a combination of discrete and continuous variables. Examples of prior use of these models can be found in [5, 8].

The atomic building blocks of our models are the densities (distributions) over continuous (discrete) measurements  $p_k(x_{nd}|c)$ . For simplicity we constrain these atoms to be Gaussian densities for continuous valued attributes, and multinomial probability mass functions when the attributes are discrete with a finite number of values. An Expectation-Maximization algorithm is used to train each model.

### 3 Viola-Jones features

For the classification task, we use a subset of the rectangular difference-of-intensity features described in the landmark paper on object recognition by Viola and Jones [11]. An input image (color, grayscale or black-white) is first transformed into a luminance image, and then into the integral image by accumulating luminance values over rows and columns.<sup>1</sup> The rectangular features types that we use are illustrated in Figure 1. The features are evaluated over a grid of locations in the image, and over several different scales. Any time the value satisfies preset threshold criteria, the feature *fires* and is recorded as a quintuple:  $\{\text{FeatureType}, x, y, \log\text{Width}, \log\text{Height}\}$ . *FeatureType* is a discrete value such as *AreaSumLow* or *DiagonalBandHigh*.  $x$  and  $y$  represent the coordinates at which the feature fired.  $\log\text{Width}$

<sup>1</sup>We report categorization experiments on black and white images. In other applications, color information may be crucial to the categorization task, and can be included.

and  $\log\text{Height}$  are the logarithms of the width and height of the rectangle within which the feature was measured. Each quintuple becomes a feature-entry  $x_n$  in the feature-list for an input image. The feature dimensionality,  $D$ , is thus five. Due to variability in the input image the length of the feature list is variable.

## 4 Experiments

### 4.1 NIST Tax Forms data sets

The NIST tax forms datasets (Special Databases SPDB2 and SPDB6) were prepared for the information extraction from forms, and are described in detail in [3]. Each of these datasets contain roughly 6000 images of tax forms filled by machine print (SPDB2) and handwriting (SPDB6), spread over 20 categories.<sup>2</sup> Apart from the form category, the samples varied in image lightness/darkness, amount of filled material, page skew, and minor size variations. Figure 2 shows thumbnails of example forms.

### 4.2 Model training

Ten images were picked at random from each category, and the list of Viola-Jones features that fired on these examples were concatenated to create a large feature list. This concatenated list is one way of capturing variability of features fired on examples of the category. Each such list was then used to train up an LCI model for that category with  $K \in \{25, 50, 100\}$  components.

Figure 3 shows diagrammatically a sampling from the long list of features that fired on two specific examples of different categories. Due to space limitations the figures are zoomed-in to a small sub-region of two different images (from categories *1040\_1* and *sch\_e\_2*), and only a fraction of fired features are illustrated to avoid visual clutter. We also limit the illustration to features of two different *FeatureTypes* (*HorizontalStep*), and (*VerticalStep*), two different  $\log\text{Width}$  and two  $\log\text{Height}$  values. Nevertheless, the difference between the two documents is evident in the distributions of these features.

Each example in the dataset (including the training docs) was then classified according to *approximate* maximum likelihood where, for computational efficiency, the summation in (2) is replaced by a max operation to obtain an approximation to the true likelihood for each class.

<sup>2</sup>These are not scans of actual forms – the filling up of forms was simulated. However, the forms do provide sufficient variability to test image classification techniques.

Figure 2. Examples of a few categories of forms shown as thumbnails.

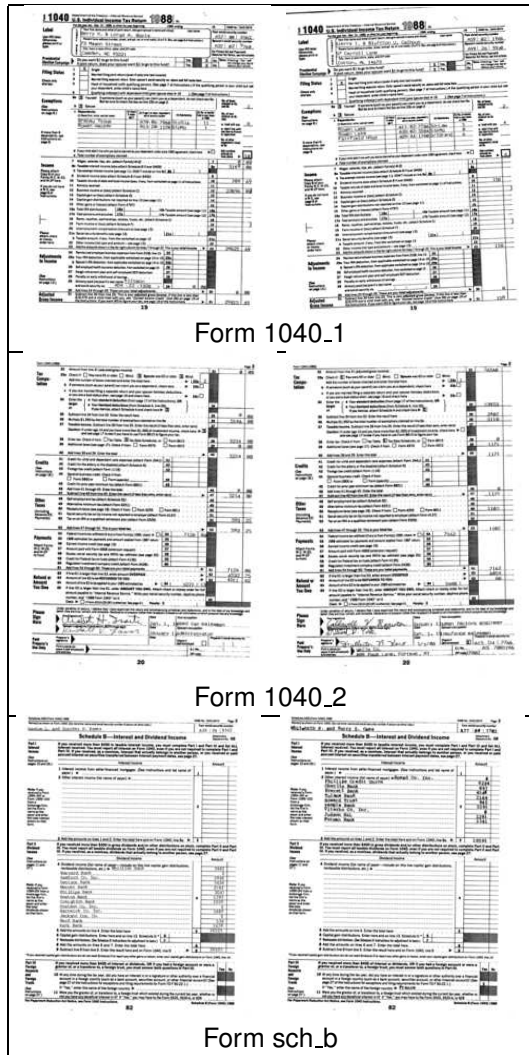
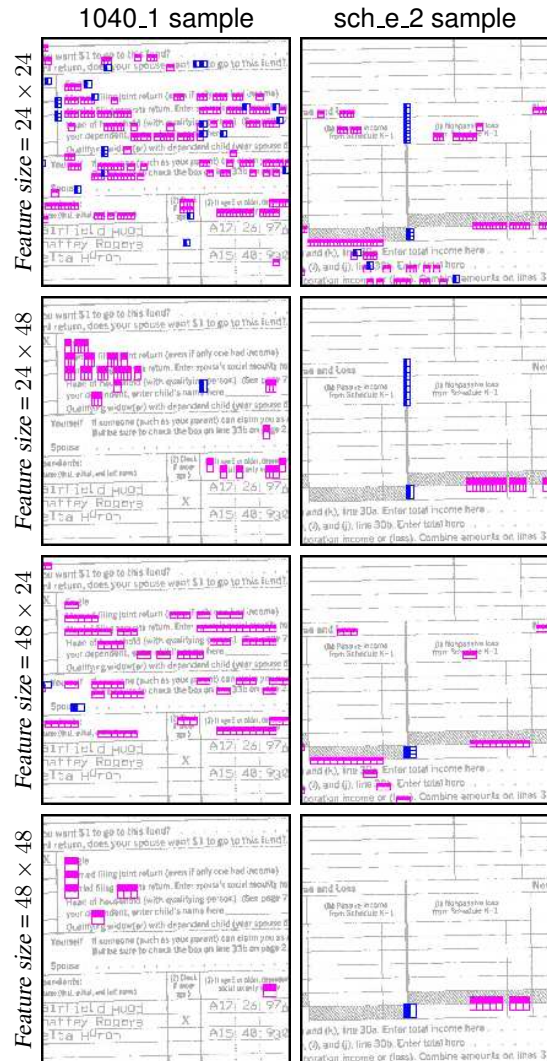


Figure 3. Illustration of Viola-Jones features that fired on a small region of two images.



### 4.3 Evaluation

As mentioned above, ten examples from each class (200 examples in all) were used for training our models, and all 5590 examples were used for testing.<sup>3</sup> The classification results are summarized in Table 1.

This method improves classification performance over that reported by Shin et al. [9] while using an order of magnitude less training data, and without relying on features handcrafted for Document Image Analysis.<sup>4</sup>

<sup>3</sup>The direct error rates are therefore optimistic, but only slightly so. The test error rate is at most 10/5390, rather than 10/5590

<sup>4</sup>The Viola-Jones features are simple and general purpose, and were proposed for detecting faces in photographs.

### 5 Discussion

The main contribution of this paper is to propose an OCR-free, segmentation-free approach to document image classification. Several prior approaches relied on segmentation of form images and analysis of the layout and appearance of the segments. Layout information, when available is certainly informative to the task. But automatic layout analysis itself can be a bottleneck – being error-prone in the face of variations observed in filled forms. For example, rotations disrupt (or introduce noise in) the identification of rectangular layout (such as XY-trees), and features such as horizontal and vertical lines. Small variations such as extraneous markings can cause big changes in results of

**Table 1. Performance of our best system (K=100) on the NIST Structured Forms Database SPDB2.**

Form	#Err	#Samples	Form	#Err	#Samples
1040_1	.	900	sch_c_1	.	198
1040_2	.	900	sch_c_2	.	109
2106_1	1	90	sch_d_1	.	223
2106_2	.	79	sch_d_2	.	242
2441	1	98	sch_e_1	.	341
4562_1	.	263	sch_e_2	.	358
4562_2	.	270	sch_f_1	.	86
6251	7	97	sch_f_2	.	59
sch_a	.	481	sch_se_1	1	132
sch_b	.	555	sch_se_2	.	109

10 errors in 5590 samples. Percentage Error = 0.18

connected component analysis. Our approach is to bypass layout analysis, and rely completely on local visual appearance features. Although each individual feature-firing cannot be consistently repeated on different examples of a category, robustness is achieved by (a) postulating broad equivalence classes of features (the mixture components) during training, and (b) examining populations of features on test-images, rather than focusing on single features.

The maximum-likelihood classification scheme presented here can also be looked upon as a minimum distance classification. The feature list derived from an image can be looked upon as a sampling or empirical distribution of features in a five-dimensional feature space. Each category is described by a category specific distribution in the same space. The input is assigned to the category whose distribution is closest, in the Kullback-Liebler sense, to the empirical distribution. This correspondence is well known in the text-categorization/retrieval community where observations are variable-length lists of words (*e.g.*, [6]).

The Viola-Jones features are simple and fast to compute, and have become popular in computer vision literature. Complete processing of a full 300 dpi image of a letter sized page takes only a few seconds with an unoptimized Java implementation. Working on subsampled images, along with efficient image traversal may provide an order of magnitude speedup without loss of performance.

The Viola-Jones features may, of course, be replaced or complemented in our model with other local features such as transition counts, interest points, texture features. Classification on the basis of such low-level features (rather than hand-crafted features), in principle, makes the technology applicable to a variety of document classification tasks.

One advantage of classification with generative models, is that category specific models can be built, and either selectively deployed or new categories added to the task with little or no retraining. Discriminative models such as deci-

sion trees or support vector machines require complete re-training with the addition of new classes.

Our approach is related to bags of key-points approaches in computer vision, *e.g.*, [10, 1], but spatial structure is modeled intrinsically by including the locations of features as part of the feature coordinates.

## 6 Conclusion

Document image classification is an important tool in automated document processing applications. We present a fast and accurate method for classifying form images based on visual appearance, and without the need for OCR or any document segmentation or even connected component analysis. The classification accuracy is the best reported so far on the NIST Structured Forms Database – a publicly available data set. The proposed approach promises to be broadly applicable to other image classification tasks owing to its simplicity, trainable model, and extensibility (*e.g.*, to other features sets).

## References

- [1] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.
- [2] M. Diligenti, P. Frasconi, and M. Gori. Hidden Tree Markov Models for document image classification. *IEEE Trans. PAMI*, 25(4):519–523, 2003.
- [3] D. L. Dimmick, M. D. Garris, and C. L. Wilson. Structured Forms Database. Technical Report Technical Report Special Database 2, SFRS, National Institute of Standards and Technology, December 1991.
- [4] F. Dubiel and A. Dengel. FormClas – a system for OCR free identification of forms. In J. Hull and S. L. Taylor, editors, *Document Analysis Systems II*, pages 189–208. World Scientific Publishing Co. Inc., Singapore, 1998.
- [5] M. J. Evans, Z. Gilula, and I. Guttman. Latent class analysis of two way contingency tables by Bayesian methods. *Biometrika*, 76(3):557–563, 1989.
- [6] T. Hofmann. Unsupervised learning by Probabilistic Latent Semantic Analysis. *Machine Learning Journal*, 42(1):177196, 2001.
- [7] J. Hu, R. Kashi, and G. Wilfong. Document image layout comparison and classification. In *Proc. 5th ICDAR*, pages 285–288, Bangalore, India, September 1999.
- [8] P. Sarkar, H. S. Baird, and J. Henderson. Triage of OCR output using ‘confidence’ scores. In *Proceedings of SPIE DR&R IX*, San Jose, California, USA, January 2002.
- [9] C. Shin, D. Doermann, and A. Rosenfeld. Classification of document pages using structure-based features. *IJDAR*, 3(4):232–247, 2001.
- [10] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *Proc. ICCV*, 2005.
- [11] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, pages 511–518, 2001.