Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis
Department of Computer Science
Ryerson University, Toronto, Ontario
Email: {aharley, aufkes, kosta}@scs.ryerson.ca

***Abstract*—This paper presents a new state-of-the-art for document image classification and retrieval, using features learned by deep convolutional neural networks (CNNs). In object and scene analysis, deep neural nets are capable of learning a hierarchical chain of abstraction from pixel inputs to concise and descriptive representations. The current work explores this capacity in the realm of document analysis, and confirms that this representation strategy is superior to a variety of popular handcrafted alternatives. Extensive experiments show that (i) features extracted from CNNs are robust to compression, (ii) CNNs trained on non-document images transfer well to document analysis tasks, and (iii) enforcing region-specific feature-learning is unnecessary given sufficient training data. This work also makes available a new labelled subset of the IIT-CDIP collection, containing 400,000 document images across 16 categories.**

I. Introduction

Many document types have a distinct visual style. For example, "letter" documents are typically written in a standard format, which is recognizable even at scales where the text is unreadable. Motivated by this observation, this paper addresses the problem of document classification and retrieval, based on the visual structure and layout of document images.

Content-based analysis of document images has a number of applications. In digital libraries, documents are often stored as images before they are processed by an optical character recognition (OCR) system, which means image analysis is the only available tool for initial indexing and classification [18]. As a pre-processing stage, document image analysis can facilitate and improve OCR by providing information about each document's visual layout [8]. Furthermore, document information that is lost in OCR, such as typeface, graphics, and layout, is often stored and indexed using images or image descriptors. Therefore, image analysis is complementary to OCR at several stages of document analysis.

The challenge of document image analysis arises from the fact that within each document type, there exists a wide range of visual variability. This point is illustrated in the document images shown in Figure 1. Intra-class variability renders spatial layout analysis difficult, and template-based matching impossible [6]. Another issue is that documents of different categories often have substantial visual similarities. For instance, there exist advertisements that look like news articles, and questionnaires that look like forms, and so on. From the perspective of "visual styles", some erroneous retrievals in such circumstances may be justifiable, but in general the task of document image analysis is to classify and retrieve documents despite intra-class variability, and inter-class similarity.

Similar challenges appear in other fields, such as object recognition and scene classification. In those domains, the current state-of-the-art approach involves training a deep convolutional neural network (CNN) [16] to learn features for the task [20]. Inspired by the success of CNNs in other domains, this paper presents an extensive evaluation of CNNs for document classification and retrieval.

A. Related Work

In the past twenty years of document image analysis, research has oscillated between region-based analysis and whole image analysis, and simultaneously, between handcrafted features and machine-learned ones.

The power of region-based analysis of document images has been clearly demonstrated in the domain of rigidly structured documents, such as forms and business letters [5]. To some extent, the classification of perfectly rigid documents (*e.g.*, forms) can be reduced to the problem of template matching, and less-rigid document types (*e.g.*, letters) can similarly be classified by fitting the geometric configuration of the document's components to one of several template configurations, via geometric transformations [9]. However, for documents with more flexible structures, as considered herein, template-based approaches are inapplicable.

An alternative strategy is to treat document images holistically, and search for discriminative "landmark" features that may appear anywhere in the document. This strategy is sometimes called a "bag of visual words" approach, since it describes images with a histogram over an orderless vocabulary of features [7]. For example, finding a salutation in a document (potentially through OCR) is a good cue that the document is a letter, regardless of that feature's exact spatial position [22]. This approach has been successful in retrieving and classifying a broader range of documents than the template-based approaches, although the approach is less discriminating in the domain of rigid-template documents.

Recently, there have been attempts to bridge the gap between region-based and holistic analyses. By concatenating image features pooled at several region sizes, it is possible to build a descriptor that contains both global and local layout characteristics [15]. This technique, known as spatial pyramid matching, was initially developed for categorizing scenes, but it has been shown to apply well to documents also, especially if the pooling regions are designed with document categorization in mind [14]. For document retrieval, this type of representation represents the current state-of-the-art.

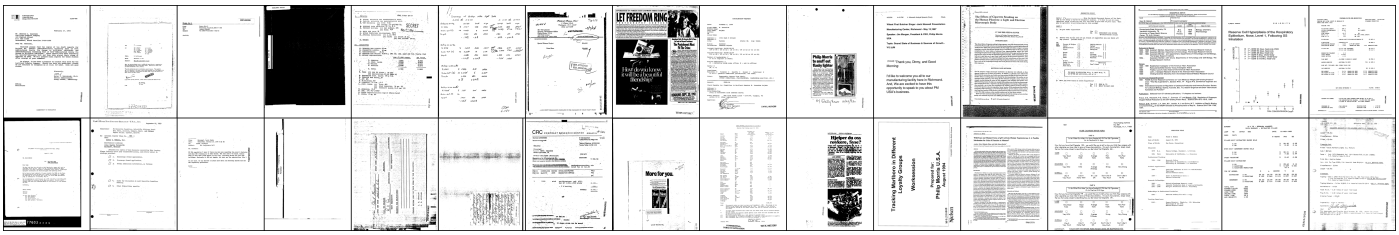

Fig. 1. Representative examples from each category of the RVL-CDIP dataset used for evaluation in this paper. For each category, two images are shown in a column. The document classes shown are (left-to-right) "letter", "memo", "email", "file folder", "form", "handwritten", "invoice", "advertisement", "budget", "news article", "presentation", "scientific publication", "questionnaire", "resume", "scientific report", and "specification".

At the same time, many researchers have replaced handcrafted features and representations with machine-learned variants [8]. Most recently, it was shown that every component of a document image analysis system, from feature-building to classification, can be learned by a convolutional neural network (CNN) [11]. In that work, the authors reported a substantial increase in classification accuracy compared to the previous best reported on the same dataset. Similar successes with CNNs have been reported in object recognition [12], and also in fine-grained object recognition [20].

The success of CNNs in fine-grained object recognition is especially relevant to document image analysis, since that field shares some significant challenges with document analysis, *e.g.*, (i) inter-class similarity, and (ii) a lack of labelled training data. In that field, it has been found that before training the CNN on the data of interest, it is best to pre-train the network on a larger related problem (*e.g.*, the ImageNet 2012 challenge [21]) to avoid overfitting. Additionally, in problems where region-specific information is important, it is potentially better to encode this information in multiple networks trained on the regions of interest than in a single network trained on the entire image [4]. This paper seeks to investigate whether these insights are relevant to document image analysis.

Finally, CNNs in other domains have recently been extended to the task of image retrieval. After a CNN is trained on classification, the layers of the network can be interpreted as forming a hierarchical chain of abstraction, where the lowest layers contain simple features, and the highest layers contain concise and descriptive representations [16]. Output extracted near the top of a CNN can therefore serve as a feature vector for any task, including retrieval [20]. The present work is the first to apply this idea toward document retrieval.

B. Contributions

In the light of previous work, this paper makes the following contributions. First, the paper presents experiments showing that features extracted from CNNs are superior to the state-of-the-art handcrafted alternatives. Second, experiments in feature compression show that the CNN features can be compressed to very short codes with negligible loss in performance. Third, this work demonstrates that CNNs trained on non-document images transfer well to document-related tasks. Fourth, the paper explores a strategy of embedding human knowledge of document structure into CNN architectures, by guiding an ensemble of CNNs toward learning region-specific features. Finally, this work makes available a new labelled subset of the Illinois Institute of Technology Complex Document Information Processing (IIT-CDIP) collection of tobacco litigation documents [17]. The new dataset, named the Ryerson Vision Lab Complex Document Information Processing (RVL-CDIP) dataset, contains 400,000 document images across 16 categories, and is available at http://scs.ryerson.ca/~aharley/rvl-cdip.

II. Technical Approach

In structured documents, the layout of text and graphics often reflects important information about genre. Therefore, documents of a category often share region-specific features. This paper attempts to learn these informative features by training either a single holistic CNN or an ensemble of region-based CNNs. Additionally, the paper explores two different initialization strategies: the first initializes the weights of the CNNs randomly; the second "transfers" weights learned in another visual classification task.

A. Holistic convolutional neural networks

Convolutional neural networks take a matrix of pixels as input, process this input through a stack of convolutional layers, then classify the output of those convolutional layers using two or three fully-connected layers [16]. The fully-connected layer activations in CNNs are generally not invariant to geometric transformations in the input. In applications such as object detection, this is an inconvenient property, and much work has been done to add spatial invariance, *e.g.*, by "jittering" the training data [16]. For document analysis, however, spatial specificity of features may be beneficial. By design, a holistic CNN trained on a dataset of well-aligned document images may be capable of learning region-specific features automatically.

Typically, CNNs are trained to perform a classification task, but a CNN trained on classification can be exploited to perform retrieval also. It has been shown that activation patterns near the top of a deep CNN provide discriminative feature vectors for a variety of tasks [20]. These feature vectors are high-dimensional (*e.g.*, 4096 dimensions), but their dimensionality can be reduced (*e.g.*, via principal component analysis) without significantly affecting their discriminative power [2].

B. Region-based guidance

Guiding CNNs to learn region-based features may aid fine-grained discrimination. For example, it is possible that a holistic CNN may learn that the "header" region of a document is important, but training a CNN to classify documents using only this region increases the likelihood that this feature will be learned. The idea of this approach is to devote one CNN

to each region of interest, and therefore force multiple CNNs to learn rich region dependent representations, from which features can be extracted and combined.

Any number of region-specific CNNs can be used in this approach. In this work, a total of five CNNs are used. Four of these are region-tuned, placed at the header, left body, right body, and footer of the document images. The fifth is a holistic CNN, trained on the entire images. The final region-based representation of document images is built by combining and compressing features extracted from each region-tuned CNN. The final descriptor is represented by the concatenation of region specific features: $[\phi_0, \phi_1, \ldots, \phi_n]$, where ϕ_0 represents the PCA-compressed feature vector extracted from the holistic CNN, and $\phi_1, \ldots, \phi_n$ represent the analogous vectors extracted from regions 1 through n.

C. Transfer learning

The goal of transfer learning is to facilitate learning on problems with insufficient training data, by taking advantage of shared structure in related problems [1]. In the context of CNNs, transfer learning can be implemented at the weight initialization step. The typical initialization strategy for CNNs is to set all weights to small random numbers [16]. An alternative strategy is to pre-train the network on a complementary task, which has more training data than the target task, effectively transferring the learned weights to the new problem. A popular choice for pre-training is the ImageNet 2012 challenge, as it contains over a million training examples of natural images, in 1000 object categories [21]. Features extracted from an ImageNet-trained network have been shown to be effective general-purpose features in a variety of other vision challenges, even without fine-tuning on the target problem [20]. This paper investigates whether the ImageNet features are general enough to be applied to documents, and furthermore whether weights transferred from such pre-training provide better results than random initialization for document-classifying CNNs.

III. Empirical Evaluation

A. Datasets

The performance of the proposed approach was evaluated on two versions of the IIT-CDIP test collection [17]. This collection contains high resolution images of scanned documents, collected from public records of lawsuits against American tobacco companies [23]. In total, the database has over seven million documents, hand-labelled with tags. Many documents in the dataset have erroneous tags.

The first version of the dataset, listed in the results as Small-CDIP, is a sample of 3482 images from the collection, selected and labelled in another work [13]. Each image in this dataset has one of ten labels; the most common label is "letter".

The second version of the dataset, listed in the results as RVL-CDIP, is a new random sample of 25,000 images from each of 16 categories in the IIT-CDIP collection, for a total of 400,000 labelled images. This sample was collected specifically for the present paper. The 16 categories are "letter", "memo", "email", "file folder", "form", "handwritten", "invoice", "advertisement", "budget", "news article", "presentation", "scientific publication", "questionnaire", "resume", "scientific report", and "specification". The selection of categories was guided by earlier work on document categorization [19], and also by the range of categories present in Small-CDIP. Another factor was the knowledge that CNNs do well with large datasets (*e.g.*, over a million images) [12], so the selection of categories was restricted to document types that were well represented in the dataset. A sample of the dataset is shown in Figure 1. The dataset can be downloaded at the following url: http://scs.ryerson.ca/~aharley/rvl-cdip.

Each dataset was split into three subsets for the purposes of experimentation. The Small-CDIP dataset was split as in the related work [13]: 800 images were used for training, 200 for validation, and the remainder for testing. Since this is a small dataset, 10 random splits in those proportions were created; reported results reflect the median performance from those splits. In the case of retrieval, the median was selected based on the mean average precision at the 10th retrieval (mAP@10). The RVL-CDIP dataset was split in proportions similar to those of ImageNet [21]: 320,000 images were used for training, 40,000 images for validation, and 40,000 images for testing. The validation sets were used to find plateaus in the CNN training process. All results are reported on the test sets.

B. Implementation details

The CNNs were implemented in Caffe [10]. All networks computed an N-way softmax at the top layer, where N is the number of categories being learned.

All but two of the CNNs used Caffe's reference ImageNet architecture, which is based on the work of Krizhevsky *et al.* [12]. This network has five convolutional layers, and three fully-connected layers, with pooling, ReLU, and drop-out employed at several stages in between. As input, the network takes images of size 227×227. The full architecture can be written as $227 \times 227 - 11 \times 11 \times 96 - 5 \times 5 \times 256 - 3 \times 3 \times 384 - 3 \times 3 \times 384 - 3 \times 3 \times 256 - 4096 - 4096 - N$. Features were extracted from these CNNs by taking the output of the first fully-connected layer, which has 4096 dimensions.

The first network with a different architecture is listed in the results as a "small" holistic CNN. This network uses hyperparameters established in another work on document image classification [11]: two convolutional layers and three fully-connected layers, with pooling, ReLU, and drop-out employed at several stages in between. The network takes as input images of size 150×150. The full architecture can be written as $150 \times 150 - 36 \times 36 \times 20 - 8 \times 50 - 1000 - 1000 - N$. As with the ImageNet networks, features were extracted from this network by taking the output of the first fully-connected layer, which in this case has 1000 dimensions.

The second network with a different architecture is the "Ensemble of CNNs" network, which uses vectors extracted from the region-based CNNs to perform retrieval and classification. For classification, the individual region-based vectors were compressed using principal component analysis (PCA) to 640 dimensions, and then concatenated into a feature vector of size 3200. Finally, a fully-connected network of size $3200 - 4096 - N$ was trained to classify these features. For retrieval, features were created by individually compressing each region's feature vector to 128 dimensions, and then concatenating, resulting in a vector with 640 dimensions.

To extract regions from the images, all images were first resized to 780×600. The header region included the top 256 rows of pixels in each image. Similarly, the footer region included the bottom 256 rows. The left body region was delineated by the intersection of the 400 central rows and the 300 left columns; the right body region was symmetrically defined. Every extracted region was resized to 227×227 before being used as input to a CNN.

Several state-of-the-art Bag of Words (BoW) approaches to document representation were also implemented. As in previous work [14], the words were k-means clustered SURF features [3]. These features were pooled in a spatial pyramid, as well as in various combinations of horizontal and vertical partitions [14]. In the results, we denote these horizontal-vertical partitioning schemes with HaVb, where a is the number of times the image was recursively split horizontally, and b is the number of times the image was recursively split vertically. For example, H0V3 has 15 bags: 1 for the original image, 2 for the first vertical split, 4 for the second vertical split, and 8 for the third. For the holistic bag of words, the resulting feature vector has 300 dimensions; H2V0 has 2100 dimensions; H0V3 has 4500 dimensions; H2V3 and the spatial pyramid approach both have 6300 dimensions. For classification of the BoW features, a random forest with 500 trees and $\sqrt{D}$ feature dimensions was trained, where D was the length of the feature vector of the complete bag of words.

Retrieval was performed using the Euclidean distance between the test set descriptors and every descriptor of the training set. The sorted distances were used to rank the images of the training data, and return a sorted list of documents for each test query. For all approaches with feature vectors larger than 128 dimensions, the vectors were first compressed to 128 dimensions using PCA before they were used for retrieval. This is consistent with the related work [20]; it not only enables fast retrieval, but also keeps the task within reasonable memory limits. The feature vectors were L2-normalized before and after PCA compression.

C. Classification results

Table I shows the classification accuracies of the various BoW approaches, along with the various CNN-based approaches, on both the Small-CDIP dataset and the RVL-CDIP dataset.

On Small-CDIP, the ensemble of region-based CNNs performed best, which demonstrates the strength of region-based analysis. The "small" holistic CNN performed similarly to the ImageNet-sized holistic CNN when both were initialized with random weights. Performance of the large network improved substantially when it was initialized with ImageNet weights. The BoW approaches performed similarly to the random-initialized CNNs. Between the BoW approaches, the spatial-pyramid-pooled BoW outperformed the rest by a small margin.

On RVL-CDIP, the holistic CNN fine-tuned from ImageNet performed better than any other approach, including the ensemble of CNNs. This suggests that given sufficient training data, the holistic CNN may be able to learn some amount of the information that the region-based analysis was expected to add. Between the random-initialized networks, the large holistic network performed better than the small network, likely due to the benefit of additional training data. As observed in Small-CDIP, fine-tuning from ImageNet improved results over random initialization, although by a smaller margin in this case. While the CNN approaches showed better performance on this larger dataset, BoW performance dropped by nearly 20%, suggesting that (i) the larger dataset presents a more difficult classification task (perhaps because it has more categories), and/or (ii) the additional training data does not help these approaches.

TABLE I. CLASSIFICATION ACCURACIES ON BOTH DATASETS.

Approach type	Region strategy	Small-CDIP	RVL-CDIP
Bag of Words	Holistic	.645	.446
	H0V3	.679	.483
	H2V0	.652	.461
	H2V3	.681	.493
	Spatial pyramid	.687	.491
Random init. CNN	Holistic (small)	.643	.851
	Holistic (ImageNet size)	.634	.878
ImageNet init. CNN	Header	.710	.849
	Left body	.667	.827
	Right body	.708	.795
	Footer	.622	.794
	Holistic	.756	**.898**
	Ensemble of regions	**.799**	.893

D. Retrieval results

Retrieval was measured using mean average precision (mAP). Average precision computes the average value of precision as a function of recall on some interval. Formally, the discrete version of this metric is given by

$$\text{AP} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{number of relevant documents}}, \tag{1}$$

where n is the number of retrieved documents, $P(k)$ is the precision of the first k retrievals, and rel(k) equals 1 if the kth retrieval is relevant and 0 otherwise. This metric is sensitive to ranking order, so the score is higher if relevant documents are retrieved before irrelevant documents. Mean average precision is simply the average precision summed over all queries, divided by the number of queries. Retrieved documents were determined to be "relevant" if they had the same class label as the query image.

Retrieval results on both datasets are summarized in Figure 2. In Small-CDIP, the CNN approaches seemed to have slightly better features for retrieval than the BoW approaches. Between the CNN approaches, the ensemble of CNNs performed better than any other approach, suggesting that the region-based training was beneficial. On the RVL-CDIP dataset, the CNN approaches showed a large improvement in accuracy, outperforming the BoW approaches by a wide margin. On this larger dataset, the holistic CNN outperformed the ensemble of CNNs. This suggests that with sufficient training data, region-based training may be unnecessary. Interestingly, the ImageNet-trained CNN outperformed the BoW approaches at all levels of retrieval, on both datasets. This supports the idea that features learned on object classification transfer well to document analysis.

An additional experiment was performed to measure the effect of PCA compression on mAP@10 performance in the RVL-CDIP dataset, the results of which are summarized in Figure 3. Remarkably, the CNN vectors showed almost no

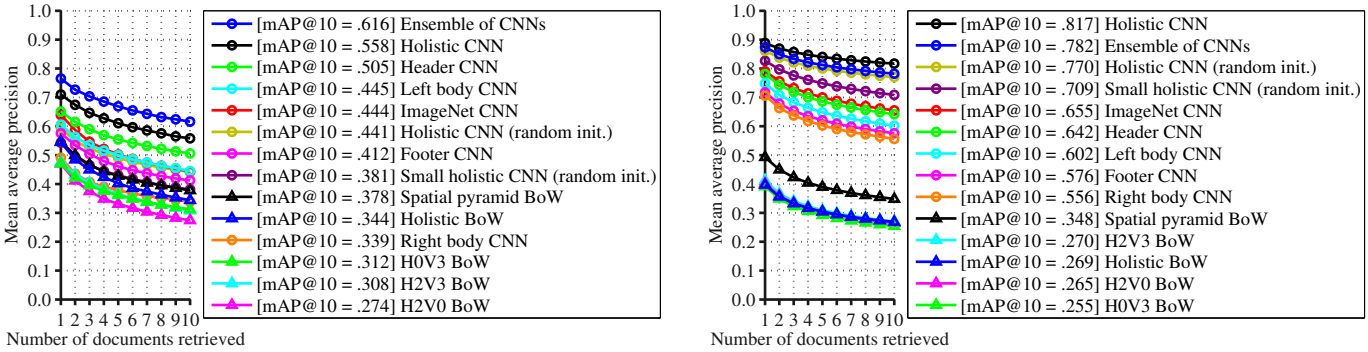

Fig. 2. Mean average precision at retrievals 1 through 10 for a variety of approaches on the Small-CDIP dataset (left) and the RVL-CDIP dataset (right). The mAP@10 results are listed in square brackets. Except where otherwise noted, all CNNs were pre-trained on ImageNet.

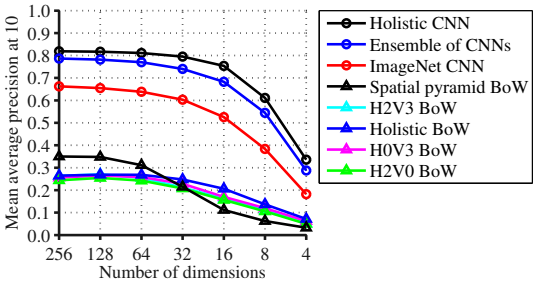

Fig. 3. The effect of PCA reduction on mean average precision at the 10th retrieval (mAP@10) on RVL-CDIP.

loss in performance until they were reduced to 16 dimensions. At all levels of compression, the holistic CNN fine-tuned from ImageNet exceeded the performance of every other approach.

IV. Conclusion

This paper established a new state-of-the-art for document image classification and retrieval, using features learned by deep convolutional neural networks. Generic features extracted from a CNN trained on ImageNet exceeded the performance of the state-of-the-art alternatives, and fine-tuning these features on document images pushed results even higher. Interestingly, the experiments showed that given sufficient training data, enforcing region-specific feature-learning is unnecessary. Furthermore, CNN features were shown to be robust to compression. In all, this work showed that the CNN approach to document image representation exceeds the power of the current handcrafted alternatives.

Acknowledgements

This work was supported by NSERC Discovery and Engage grants (held by K.G.D.), and an NSERC USRA (awarded to A.W.H.). The authors thank Palomino System Innovations Inc. for posing the problem and providing data with helpful discussions. The authors gratefully acknowledge the support of NVIDIA Corporation with the donation of a Tesla K40 GPU used for this research.

References

[1] Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *ICML*, pages 17–24. 2007.

[2] A. Babenko, A. Slesarev, A. Chigorin, and V. S. Lempitsky. Neural codes for image retrieval. In *ECCV*, pages 584–599, 2014.

[3] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speeded Up Robust Features. In *ECCV*, pages 404–417, 2006.

[4] S. Branson, G. V. Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. In *BMVC*, 2014.

[5] Y. Byun and Y. Lee. Form classification using DP matching. In *SAC*, pages 1–4, 2000.

[6] N. Chen and D. Blostein. A survey of document image classification: Problem statement, classifier architecture and performance evaluation. *IJDAR*, 10(1):1–16, 2007.

[7] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *ECCV*, pages 1–22, 2004.

[8] A. Dengel and F. Dubiel. Clustering and classification of document structure-a machine learning approach. In *ICDAR*, pages 587–591, 1995.

[9] J. Hu, R. Kashi, and G. Wilfong. Comparison and classification of documents based on layout similarity. *Information Retrieval*, 2(2-3):227–243, 2000.

[10] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. *arXiv*, http://caffe.berkeleyvision.org/, 2013.

[11] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doerman. Convolutional neural networks for document image classification. In *ICPR*, pages 3168–3172, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.

[13] J. Kumar and D. Doermann. Unsupervised classification of structurally similar document images. In *ICDAR*, pages 1225–1229, 2013.

[14] J. Kumar, P. Ye, and D. Doermann. Structural similarity for document image classification and retrieval. *PRL*, 43:119126, 2014.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *PIEEE*, 86(11):2278–2324, 1998.

[17] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard. Building a test collection for complex document information processing. In *SIGIR*, pages 665–666, 2006.

[18] S. Marinai, B. Miotti, and G. Soda. Digital libraries and document image retrieval techniques: A survey. In M. Biba and F. Xhafa, editors, *Learning Structure and Schemas from Documents*, volume 375, pages 181–204. Springer Berlin Heidelberg, 2011.

[19] G. Nagy. Twenty years of document image analysis in PAMI. *PAMI*, 22(1):38–62, 2000.

[20] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: An astounding baseline for recognition. In *CVPR*, pages 512–519, 2014.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge, 2014.

[22] S. Taylor, M. Lipshutz, and R. W. Nilson. Classification and functional decomposition of business documents. In *ICDAR*, pages 563–566, 1995.

[23] University of California, San Francisco. *The Legacy Tobacco Document Library (LTDL)*, 2007.