Document Image Segmentation Using Deep Features

K. V. Jobin and C. V. Jawahar

CVIT, IIIT-Hyderabad, India,
jobin.kv@research.iiit.ac.in, jawahar@iiit.ac.in

Abstract. This paper explores the effectiveness of deep features for document image segmentation. The document image segmentation problem is modelled as a pixel labeling task where each pixel in the document image is classified into one of the predefined labels such as text, comments, decorations and background. Our method first extracts deep features from superpixels of the document image. Then we learn an SVM classifier using these features, and segment the document image. Fisher vector encoded convolutional layer features (FV-CNN) and fully connected layer features (FC-CNN) are used in our study. Experiments validate that our method is effective and yields better results for segmenting document images in comparison to the popular approaches on benchmark handwritten datasets.

1 Introduction

Document image segmentation can be considered as the primary stage of document image analysis and understanding pipeline. The objective of this step is often to segment the image into semantically similar regions such as text, graphics, comments, decorations, backgrounds, etc. This problem is further challenging for historical documents. Analysis and understanding of historical document images is an active area of research. Challenges of historical handwritten document images such as unstructured layouts, degradation, various handwritten styles, etc. are exemplified in Figure 1.

The early approaches for document segmentation (such as [1]) are based on binarization and connected component analysis. This approach fails with the images that have non-uniform background color (Figure 1(b)). To get rid of this issue, instead of taking connected components, researchers started to classify image patches or superpixels to segment the image. Various features such as color and texture [2], SIFT, SURF, LBP, HOG, etc. are extracted for segmentation.

Various methods have been presented in the literature to perform the document image segmentation task. Most of these approaches focus on improving any of the stages of segmentation pipeline such as pre-processing, feature extraction, feature modeling, etc.

The recent paper on historical document images [3] utilizes the idea of autoencoder to learn the features automatically. Generally, an autoencoder is a neural

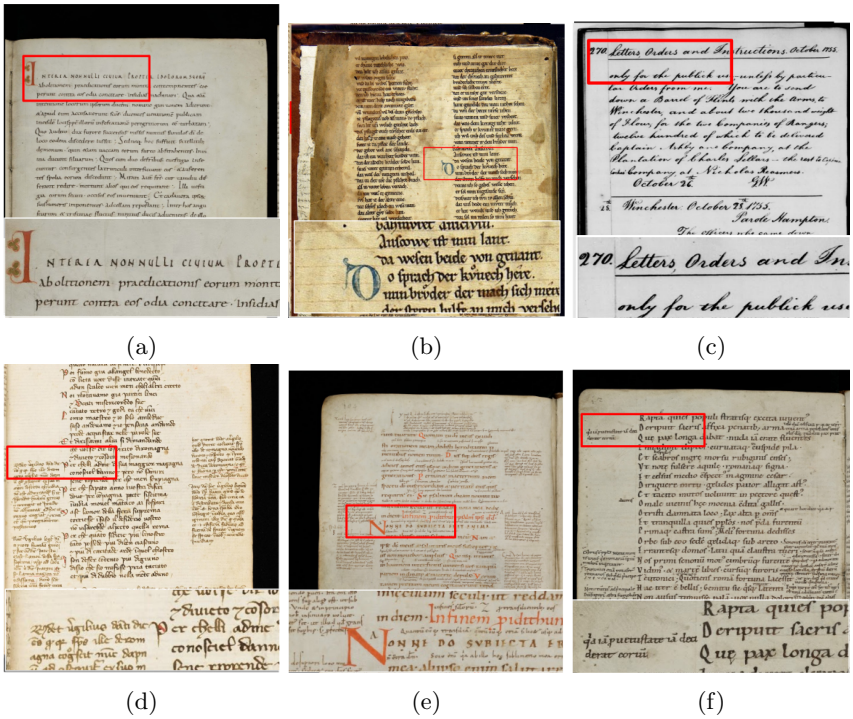


Fig. 1: Sample historical handwritten document images for layout segmentation from the following datasets: (a) St. Gall, (b) Parzival, (c) G. Washington, (d) CB55, (e) CSG18, (f) CSG863. The bottom part of each image is the zoomed view of red rectangular region.

network trained to reconstruct its input. In this paper, the encoder output of the autoencoder is extracted as features and feed to an off-the-shelf classifier to segment an image. In the paper [5], incorporated the SLIC superpixel extraction stage inorder to speed up the process in [3]. Another paper [7] uses Conditional Random Field (CRF) [6] to model the local and contextual information jointly. This helps to refine the segmentation results of [5]. Another recent approach [8] segments document images with a simple CNN architecture. Inspired from [5], we consider the segmentation problem as a pixel labeling problem. In contrast to the above approaches, our work extracts deep features from each superpixel and classifies using SVM. Our results show that deep features are more robust than handcrafted features, the autoencoder based approach [3] and the CNN based approach [8].

In this paper, we explore the effectiveness of deep features in the document segmentation. Document images have many practical difficulties to use CNNs directly. The main difficulty is that the image should be resized to the size of input of the CNN architecture. It may poorly affect the performance of the

Table 1: Details of of the dataset used in our experiments. TR, TE, and VA denotes size of the training, test, and validation sets respectively.

Dataset	Image size (pixels)	TR	TE	VA
G. Washington [10]	2200 × 3400	10	5	4
St. Gall [11]	1664 × 2496	20	30	10
Parzival [12]	2000 × 3008	20	30	10
CB55 [13]	4872 × 6496	20	13	2
CSG18 [13]	3328 × 4992	20	10	10
CSG863 [13]	3328 × 4992	20	10	10

method. Moreover, the time and data required to train deep neural networks are significantly large compared to standard statistical machine learning methods such as Support Vector Machines (SVM). We try to overcome this disadvantage.

We pose the document image segmentation task as a semantic segmentation problem. Semantic segmentation is a popular problem in computer vision in which each pixel is assigned to its most appropriate label from a predefined label set. There are many approaches that use CNN for semantic segmentation [4]. In this work, we first extract deep features from document image pixels and train an SVM classifier to label the pixels. We compare the performance of the proposed approach with the following approaches: (i) local MLP [5], (ii) CRF [7] and (iii) CNN [8] on six different historical document image datasets. Our proposed method gives superior quantitative results consistently.

2 Proposed Method

We pose the document image segmentation problem as a pixel labeling problem. Consider an image I of size $w \times h \times d$, where w, h and d are the width, height and the number of color channel of the image respectively. Let $x_{i,j}$ is the pixel of image I at position (i, j), where $i \in \{1, \ldots, w\}$ and $j \in \{1, \ldots, h\}$. We train a statistical model which learns the label l from a set of labels L for each pixel $x_{i,j}$. The label set $\mathcal{L}$ for historical document images is set as $L = \{$body text, comments, decoration, background$\}$ similar to the past methods. We extract deep features proposed by Cimpoi *et al.* [9] from image patch surrounding each pixel x_{ij} in document image I. Finally, using an SVM classifier, on top of the extracted features, we assign a label l to each pixel x_{ij}.

2.1 Image pre-processing

In our approach, we apply the superpixel segmentation algorithm SLIC. In document images, most of the pixels in a superpixel share the same label. The superiority of the superpixel based labeling approach over the pixel labeling approach for the page segmentation task has been demonstrated in [5].

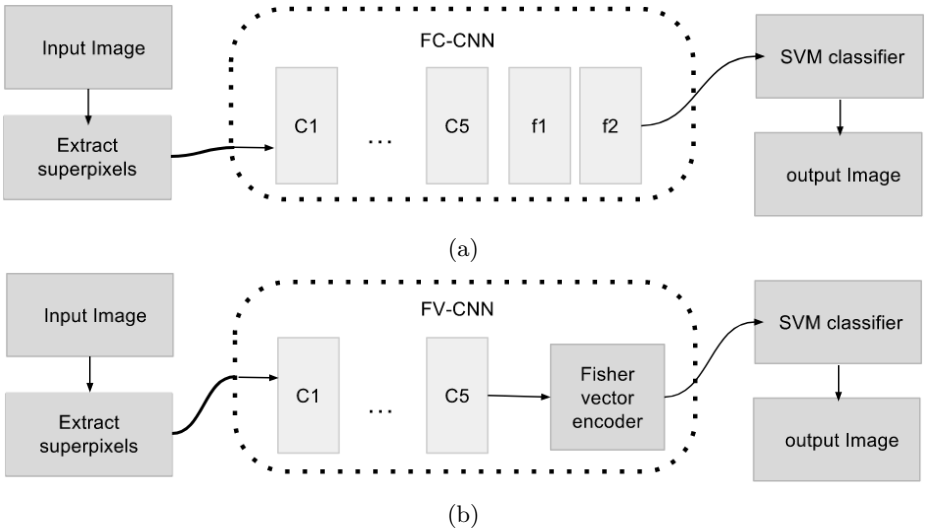


Fig. 2: The proposed approaches: (a) and (b) are separate pipelines for segmenting a document image using FC-CNN and FV-CNN features. The dotted box represents the feature extraction module, c1, ..., c5 represent convolutional layers of CNN and f1 and f2 represent the fully connected layers of CNN.

2.2 Texture of document images

In early works such as [14], texture is characterized with the arrangement of local patterns by the distribution of local filter bank responses. The filter banks are capable of capturing edges, spots, and bars at different scales and orientations. In the work [15] propose *textons* which are define by combining the filter responses. The idea of *textons* is improved by new pooling schemes such as soft-assignment [16] and Fisher Vectors (FVs) [18]. The work [17] uses FV encoding on the SIFT features, which achieved the state-of-the-art result in textures, objects and scene detection tasks.

One of the key observations made while visualizing the weights of a learned CNN [19] is that the initial layer learns to detect low level patterns such as dots, line edges, strokes, etc. While later layers in the network learn the higher level structures from images like face, the shape of an object, etc. Even if the CNN is trained with a different dataset or non-document images, the filters are capable of emphasizing various texture patterns in an image.

This paper proposes two types of deep features, namely FC-CNN and FV-CNN inspired from [9]. Both the descriptors are based on the same CNN features [20] obtained from an off-the-shelf CNN pre-trained on the ILSVRC 2012 dataset [21]. We evaluate the performance of segmentation using both features.

The FC-CNN descriptor is obtained by extracting the output of the penultimate fully connected layer of a CNN, including the nonlinear gating function, applied to an input image. This feature can be considered as an object descriptor

because a fully connected layer captures the overall shape of an object. However it has some drawbacks in case of document images: i) since it is using a fully connected neural network, the input image patch should be resized, which makes the feature faulty ii) the FC-CNN feature represents the entire shape of an object rather than the texture. However, in document images, there is no importance in the shape of the object in a patch because it may very randomly.

The FV-CNN feature overcomes the above disadvantages because the feature is extracted by FV pooling of the convolutional filter response rather than the fully connected layer. Hence, the FV-CNN is a more efficient way to describe the texture of an image than the FC-CNN. Since the feature responses are taken from the convolutional layer, no rescaling of input patch/image is required. In the work[9], it is proved empirically that the efficiency of FV-CNN feature improve from the initial convolutional layer to final convolutional layer monotonically. Hence, in all our experiments, we use the final convolutional layer of the pre-trained neural network.

2.3 Document image segmentation

From the given training data, first, we generate superpixels by running the SLIC algorithm using an appropriate region size and regularization parameters. To get the contextual information about a superpixel, we crop each superpixel into patches of size $p \times p$ with the center aligned to the center of the superpixel. The optimum patch size p is calculated by minimizing the superpixel classification error (Figure 4). Then we extract the deep features for each superpixel region from the corresponding patches.

In FV-CNN, from the densely pooled response of the input image in the convolutional layer, we learn a Fisher Vector (FV) encoder with 64 Gaussian components. This encoder is used for creating the FV-CNN features from the densely pooled response of the convolutional layer of pre-trained CNN architecture. In all our experiments, we use VGG-M [23] architecture trained on the ILSVRC 2012 dataset. The feature dimension in FV-CNN depends on the number of components of Fisher Vector encoder and the number of filters in the convolutional layer. However, the dimension of FC-CNN is the same as the dimension of the fully connected layer.

To model the extracted features, we train an SVM classifier with the predefined labels. In the testing stage, first we extract features from each superpixel. Then we classify the feature and label the superpixel region in the output image with the corresponding label of the superpixel. The proposed approaches using features FC-CNN and FV-CNN are described in the Figure 2.

3 Experiments

In this section, we use various datasets to demonstrate the effectiveness of the proposed algorithm on different kinds of document images. We use the standard evaluation scheme to analyze the performance of the proposed method. Since

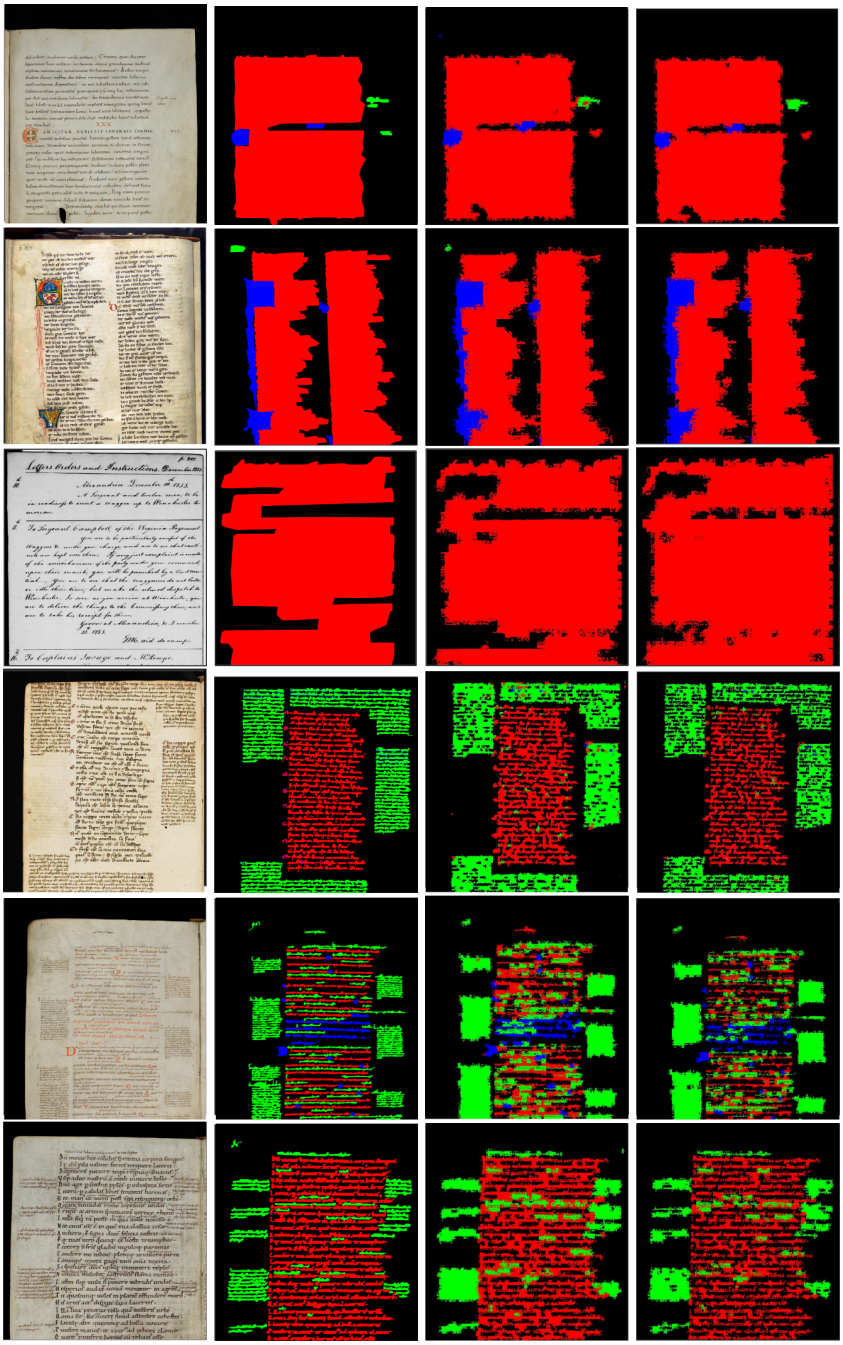


Fig. 3: Qualitative results of the proposed method. The first column shows the original test images from various datasets (1) St. Gall, (2) Parzival, (3) G. Washington, (4) CB55, (5) CSG18, (6) CSG863 from top to botom respectively. The second column shows the ground-truth images. The third and fourth columns are the output of the proposed methods with FC-CNN and FV-CNN respectively. The black, red, green and blue color of ground truth and output images represent page/background, text, comments and decoration respectively.

document image segmentation can be considered as a semantic segmentation task, for comparisons, we use pixel accuracy and region of intersection over union(IoU)[24] as the evaluation metrics. Let n_{ij} and n_{cl} are the number of pixels of class i predicted to belong to class j and the total number of classes respectively. Let $t_i = \sum_j n_{ij}$ is the total number of pixels of class i. The following are the metrics we use to evaluate the segmentation:

- pixel accuracy: $\sum_i n_{ii} / \sum_i t_i$
- mean accuracy: $(1/n_{cl}) \sum_i n_{ii} / \sum_i t_i$
- mean IoU: $(1/n_{cl}) \sum_i n_{ii} / (\sum_i t_i + \sum_j n_{ji} - n_{ii})$
- frequency weighted IoU: $(\sum_k t_k)^{-1} \sum_i t_i n_{ii} / (\sum_i t_i + \sum_j n_{ji} - n_{ii})$

Table 2: Comparison on results of historical document image segmentation on the six different dataset with Local MLP, CRF, CNN and the proposed methods FC-CNN and FV-CNN.

Datasets	G. Washington				Parzival				St.Gall			
Evaluation	pixel acc.	mean acc.	mean IoU.	mean IoU.	pixel acc.	mean acc.	mean IoU.	mean IoU.	pixel acc.	mean acc.	mean IoU.	mean IoU.
Local MLP [5]	87	89	75	83	91	64	58	86	95	89	84	92
CRF [7]	91	90	76	85	93	70	63	88	97	88	84	94
CNN [8]	91	91	77	86	94	75	68	89	98	90	87	96
FC-CNN (ours)	94	92	80	89	**97**	74	**71**	**94**	**99**	**91**	**88**	**98**
FV-CNN (ours)	**95**	**93**	**81**	**91**	**97**	**76**	**71**	**94**	**99**	**91**	**88**	**98**

Datasets	CB55				CSG18				CSG863			
Evaluation	pixel acc.	mean acc.	mean IoU.	mean IoU.	pixel acc.	mean acc.	mean IoU.	mean IoU.	pixel acc.	mean acc.	mean IoU.	mean IoU.
Local MLP [5]	83	53	42	72	83	49	39	73	84	54	42	74
CRF [7]	84	53	42	75	86	47	37	77	86	51	42	78
CNN [8]	86	59	47	77	87	53	41	79	87	58	45	79
FC-CNN (ours)	91	64	52	86	89	64	52	85	91	66	55	87
FV-CNN (ours)	**95**	**73**	**64**	**91**	**92**	**72**	**60**	**89**	**94**	**71**	**61**	**91**

3.1 Evaluation on historical document images

To evaluate the proposed method, we use six different historical handwritten document datasets whose details can be found in Table 1. G. Washington, Parzival, and St. Gall document images are from the IAM historical document database [25]. We use the annotations described in [26]. For the experiments, we choose the following four types of regions from annotation: page/background, text block, decoration, and comment. A new datasets with more complex layout introduced [13] are CB55, CSG18 and CSG863. The CB55 dataset consists of

manuscripts from the 14th century which are written in Italian and Latin languages by one writer. The CSG18 and CSG863 datasets consist of manuscripts from the 11th century which are written in the Latin language. The number of writers of both these datasets is not specified. The details of the three datasets are presented in [27]. In SLIC superpixel extraction stage, we choose 10 pixels as the region size and the regularizing parameter as 0.01. The proposed method is compared with the current state-of-the-art method [8] and other two methods[7] and [5].

The quantitative results are shown in Table 2 and the qualitative results are shown in Figure 3. From Table 2, we can see that the proposed method with FV-CNN feature achieves maximum of 9%, 19%, 19%, 14% improvements from the current state-of-the-art method [8] in pixel accuracy, mean accuracy, mean IoU and frequency weighted IoU respectively.

3.2 Deep feature analysis

To validate the utility of deep features in the document image, we visualize the deep features using t-SNE[22] method. The Figure 5 represents the deep feature visualization of a sample image taken from the dataset CB55. From the figure, we can observe that even though the feature extraction technique is unsupervised, the feature representations of each class are clustered together.

The patch is a rectangular region we chose around the super pixels before feeding it to the CNN for feature extraction. Large patch size gives more contextual information about the superpixel it covers, it eventually reduces the importance of the superpixel because the local features are extracted uniformly from a patch. Hence the patch size needs to be optimum. The graph shown in Figure 4 explains the effect on SVM classification accuracy while increasing the patch size for various document regions.

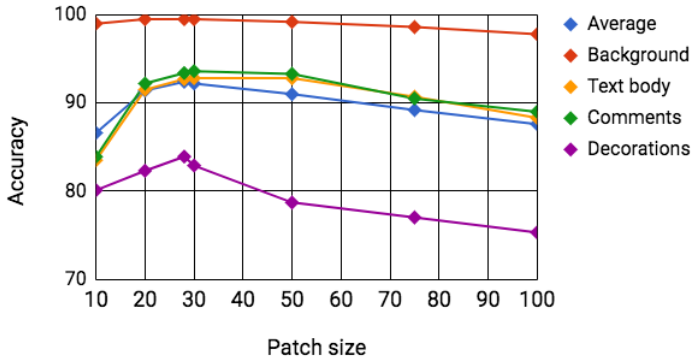
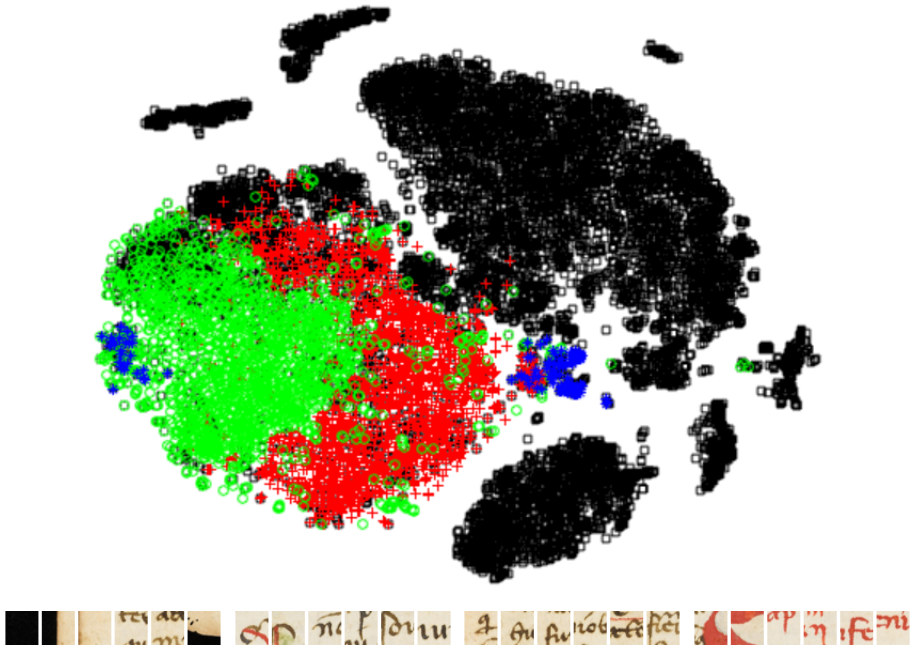


Fig. 4: The graph showing SVM classification accuracy of superpixels by extracting deep features with various patch sizes.



(a) (b) (c) (d)

Fig. 5: T-SNE [22] visualization of deep feature (FC-CNN) extracted from a sample image from the CB55 dataset. The black squares, red pluses, green rounds and blue stars are representing background, text body, comments, and decorations features respectively. The subfigure a, b, c, d are the corresponding sampled image patches of these regions respectively (better viewed in colour).

4 Conclusion

We have proposed a deep feature based document image segmentation approach. This approach has the following advantages compared to other methods: (i) It provides better feature representation for document image regions compared to the previous methods. (ii) Since the proposed approach uses a pre-trained network, the training time is significantly reduced compared to an end-to-end CNN training approach [8] because the proposed method uses SVM to train the model. As a future work, we will try the FCN[24] architecture which wil be more suites in document segmentation.

References

1. Yu Zhong, K. Karu, A.K. Jain.: Locating text in complex color images ICDAR(1995)

2. Kai Chen, Hao Wei, Jean Hennebert, Rolf Ingold, Marcus Liwicki.: Page Segmentation for Historical Handwritten Document Images Using Color and Texture Features ICFHR(2014)
3. Kai Chen, Mathias Seuret, Marcus Liwicki, Jean Hennebert, Rolf Ingold.: Page segmentation of historical document images with convolutional autoencoders. ICDAR(2015)
4. Yaroslav Ganin, Victor Lempitsky.: N4 -Fields: Neural Network Nearest Neighbor Fields for Image Transforms. ACCV(2015)
5. Kai Chen, ChengLin Liu, Mathias Seuret, Marcus Liwicki, Jean Hennebert, Rolf Ingold.: Page Segmentation for Historical Document Images Based on Superpixel Classification with Unsupervised Feature Learning. DAS(2016)
6. John Lafferty, Andrew McCallum, Fernando Pereira.: Conditional Random Fields: Probabilistic Models for Segmenting, Labeling Sequence Data ICML(2001)
7. Kai Chen, Mathias Seuret, Marcus Liwicki, Jean Hennebert, ChengLin Liu, Rolf Ingold.: Page Segmentation for Historical Handwritten Document Images Using Conditional Random Fields. ICFHR(2016)
8. Kai Chen, Mathias Seuret.: Convolutional Neural Networks for Page Segmentation of Historical Document Images. arXiv:1704.01474(2016)
9. Mircea Cimpoi, Subhransu Maji, Andrea Vedaldi.: Deep filter banks for texture recognition, segmentation. CVPR(2015)
10. Fischer, Andreas, Keller, Andreas, Frinken, Volkmar, Bunke, Horst.: Lexicon-free handwritten word spotting using character HMMs. P.R. Letters(2012)
11. Fischer, Andreas, Frinken, Volkmar, Fornés, Alicia, Bunke, Horst.: Transcription alignment of Latin manuscripts using hidden Markov models. Workshop on HDIP(2011)
12. Fischer, Andreas, Wuthrich, Markus, Liwicki, Marcus, Frinken, Volkmar, Bunke, Horst, Viehhauser, Gabriel, Stolz, Michael.: Automatic transcription of handwritten medieval documents. Virtual Systems, Multimedia(2009)
13. Simistira, Foteini, Seuret, Mathias, Eichenberger, Nicole, Garz, Angelika, Liwicki, Marcus, Ingold, Rolf.: Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts. ICFHR(2016)
14. T. Leung, J. Malik.: Recognizing surfaces using three-dimensional textons. CVPR(1999)
15. B. Julez, J.R. Bergen.: Human Factors, Behavioral Science: Textons, The Fundamental Elements in Preattentive Vision and Perception of Textures. Readings in Computer Vision(1987)
16. Lingqiao Liu, Lei Wang, Xinwang Liu.: In defense of soft-assignment coding. ICCV(2011)
17. D.G. Lowe.: Object recognition from local scale-invariant features. ICCV(1999)
18. Florent Perronnin and Jorge Sanchez and Thomas Mensink.: Improving the Fisher Kernel for Large-Scale Image Classification. ECCV(2010)
19. Matthew D. Zeiler, Rob Fergus.: Visualizing and Understanding Convolutional Networks. ECCV(2014)
20. Krizhevsky, Alex, Sutskever, Ilya, Hinton, Geoffrey E.: Imagenet classification with deep convolutional neural networks. PAMI(2012)
21. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, Li Fei-Fei.: ImageNet Large Scale Visual Recognition Challenge. IJCV(2015)
22. Imagenet classification with deep convolutional neural networks.: Visualizing data using t-SNE. JMLR(2008)

23. Chatfield, Ken, Simonyan, Karen, Vedaldi, Andrea, Zisserman, Andrew.: Return of the devil in the details: Delving deep into convolutional nets. BMVC(2014)
24. Jonathan Long, Evan Shelhamer, Trevor Darrell.: Fully convolutional networks for semantic segmentation. CVPR(2015)
25. http://www.fki.inf.unibe.ch/databases/iam-historical-document-database
26. Kai Chen, Mathias Seuret, Hao Wei, Marcus Liwicki, Jean Hennebert, Rolf Ingold.: Ground truth model, tool, and dataset for layout analysis of historical documents. DRR(2015)
27. Foteini Simistira , Mathias Seuret , Nicole Eichenberger, Angelika Garz, Marcus Liwicki, Rolf Ingold.:DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts ICFHR(2016)