[ICDAR2017] Convolutional Neural Networks for Page Segmentation of Historical Document Images

**摘要**

本文提出了一种基于卷积神经网络（CNN）的手写历史文档图像页面分割方法。我们认为页面分割是一个像素标记问题。每个像素被分类为预定义的类别之一。这一领域的传统方法依赖于精心制作的特征或大量的先验知识。相比之下，我们建议使用CNN从原始图像像素中学习特征。虽然许多研究人员致力于开发深度CNN架构来解决不同的问题，但我们只训练了一个简单的CNN，只有一个卷积层。我们表明，在不同的公共数据集上，简单的体系结构与其他深度体系结构相比具有竞争力。实验还证明了该方法的有效性和优越性。

**Ⅰ. INTRODUCTION**

页面分割是文档图像分析和理解的重要前提步骤。目标是将文档图像分割为感兴趣的区域。与机器打印文档图像的分割相比，历史文档图像的页面分割由于布局结构、装饰、书写风格和退化等多种变化而更具挑战性。我们的目标是开发一种通用的手写历史文档图像分割方法。在这种方法中，我们把分割问题看作是一个像素标记问题。对于给定的文档图像，每个像素都标记为预定义类之一。

最近，一些页面分割方法被开发出来。这些方法依赖于手工制作的特征[1]、[2]、[3]、[4]或先验知识[5]、[6]、[7]、[8]或将手工制作的特征与领域知识相结合的模型[9]、[10]。相比之下，在本文中，我们的目标是开发一种更通用的方法，自动从文档图像的像素中学习特征。单词的笔划、句子中的单词、段落中的句子等元素具有从低到高的层次结构。因为这些模式在文档的不同部分重复。基于这些特性，特征学习算法可以用来学习文档图像的布局信息。

卷积神经网络（CNN）是一种前馈人工神经网络，它在同一层的神经元之间共享权重。通过加强相邻层神经元之间的局部连接模式，CNN可以发现空间上的局部相关性[11]。CNN拥有多个卷积层和池层，在各个领域取得了许多成功，例如：手写识别[12]、图像分类[13]、自然图像中的文本识别[14]和句子分类[15]。

在我们之前的工作[16]中，我们设计了一个自动编码器，用于在训练图像上自动学习特征。自动编码器是一种前馈神经网络。其主要思想是，通过训练自动编码器重建其输入，可以在隐藏层上发现特征。然后，一个现成的分类器可以用学习到的特征进行训练，以将像素预测到不同的预定义类中。通过使用超像素作为标记单位[17]，该方法的速度得到了提高。在[18]中，应用了条件随机场（CRF）[19]，以便联合建模局部和上下文信息，以细化[17]中获得的分割结果。按照相同的想法[18]，我们认为分割问题作为一个图像贴标问题。采用超像素算法生成图像块。与[16]、[17]、[18]相比，在这项工作中，我们着重于开发一种端到端方法。我们将特征学习和分类器训练结合在一起。图像块用作输入，用于训练CNN进行标记任务。在训练期间，用于预测图像块标签的特征在CNN的卷积层上学习。

虽然许多研究人员致力于开发非常深入的CNN来解决各种问题[13]、[20]、[21]、[22]、[23]，但在所提出的方法中，我们训练了一个简单的单卷积层CNN。在公共历史文档图像数据集上的实验表明，尽管该方法结构简单，超参数调整少，但与其他CNN结构相比，该方法取得了相当的效果。

论文的其余部分组织如下。第二节对一些相关工作进行了概述。第三节介绍了建议的CNN用于分割任务。第四节报道了实验结果，第五节给出了结论。

**Ⅱ. RELATED WORK**

本节回顾了一些具有代表性的历史文档图像分割的最新方法。与当代机器印刷文档的分割不同，手写历史文档的分割更具挑战性，因为其书写风格多样，装饰丰富，退化，噪音大，布局不受限制。因此，传统的页面分割方法不能直接应用于手写历史文档。手写历史文档的分割方法很多，主要分为基于规则的分割和基于机器学习的分割。

有些方法依赖于基于文档结构的先验知识预定义的阈值。Van Phan等人[6]使用Voronoi图的面积来表示连接部件（CCs）的邻域和边界。通过应用预定义的规则，通过对相邻Voronoi区域进行分组来提取字符。Panichkriangkrai等人[7]提出了一种日本历史木版印刷书籍的文本行和字符提取系统。通过在二值化图像上使用垂直投影来分隔文本行。为了提取汉字，采用基于规则的集成来合并或拆分CCs。Gatos等人[8]提出了一种手写历史文档的文本区域和文本行分割方法。基于文档结构的先验知识，通过分析文档图像的垂直规则线和垂直空白来检测垂直文本区域。在检测到的文本区域上，采用基于Hough变换的文本行分割方法对文本行进行分割。所有这些方法在特定的文档数据集上都取得了良好的分割效果。然而，常见的限制是必须仔细定义一组规则，并假定遵守文档结构。利用结构的先验知识，调整阈值以归档良好的性能。换言之，由于其通用性，基于规则的方法不能直接应用于其他类型的历史文档图像（此前方法的不足）。

为了提高页面分割方法的通用性和鲁棒性，采用了机器学习技术。在这种情况下，通常将分割问题视为像素标记问题。特征表示是基于机器学习方法的关键。精心设计的特征是为了在标记的训练集上训练现成的分类器。Bukhari等人[2]提出了一种阿拉伯历史文档图像的文本分割方法。他们把CCS的归一化高度、前景区域、相对距离、方向和邻域信息作为特征。然后利用这些特征训练多层感知器（MLP）。最后，使用训练好的MLP将CCs分类到相关的文本类别中。Cohen等人[9]在多尺度二值化图像上应用高斯拉普拉斯算子来提取CCs。基于先验知识，选择合适的阈值以去除噪声。使用能量最小化方法和特征，例如边界框大小、面积、笔划宽度和估计的文本行距离，将每个CC标记为文本或非文本。Asi等人[10]提出了一种两步分割阿拉伯历史文档图像的方法。他们首先用Gabor滤波器提取主文本区域。然后通过最小化一个能量函数来细化分割。与基于规则的方法相比，基于机器学习的方法的优点是所需的先验知识较少。然而，现有的基于机器学习的方法依赖于手工特征工程，为特定任务获取合适的手工特征非常麻烦。

**Ⅲ. METHODOLOGY**

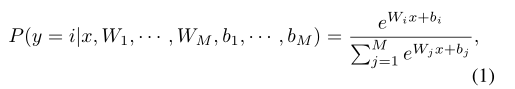
为了创建一般的页面分割方法，而不使用文档的布局结构的任何先验知识，我们将页面分割问题视为像素标记问题。我们建议使用CNN进行像素标记任务。其主要思想是学习一组特征检测器，并根据特征检测器提取的特征训练非线性分类器。通过一组特征检测器和分类器，可以将未经过预习的文档图像上的像素划分为不同的类别。

1. **Preprocessing**

为了加快像素标注过程，对于给定的文档图像，我们首先应用超像素算法来生成超像素。超像素是包含属于同一对象的像素的图像块。然后，我们不再标记所有像素，而是只标记每个超像素的中心像素，并将该超像素中的其余像素分配给相同的标签。在页面分割任务中，超像素标记方法相对于像素标记方法的优越性已经在[17]中得到了证明。在前人工作[17]的基础上，应用简单线性迭代聚类(SLIC)算法作为预处理步骤，为给定的文档图像生成超像素

1. **CNN Architecture**

CNN架构图1给出了我们CNN的架构。其结构可概括为28×28×1−26×26×4−100−M、 这里M是类别的数量。输入是灰度图像块。图像块的大小为28×28像素。我们的CNN架构只包含一个由4个内核组成的卷积层。每个内核的大小为3×3像素。与其他传统的CNN体系结构不同，我们的体系结构中不使用池层。然后，100个神经元的一个完全连接的层跟随卷积层。最后一层由逻辑回归和softmax组成，该回归输出每个类别的概率，例如



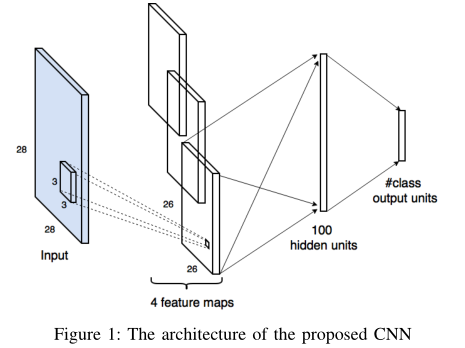
其中，x是全连接层的输出，和是该层中第个神经元的权重和偏差，M是类的数量。预测类是具有最大概率的类，因此



在CNN的卷积层和完全连接层中，整流线性单元（RELU）[24]被用作神经元。ReLU的形式如下：



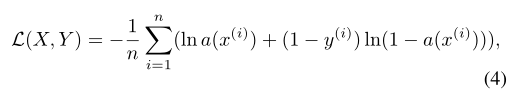
x是神经元的输入。在 [13] 中证明了在 CNN 中使用 ReLUs作为神经元优于传统的 sigmoid神经元。



1. **Training训练**

为了训练 CNN，对于每个超像素，我们生成一个以该超像素为中心的patch(块)。patch被认为是网络的输入。每个patch的大小为28×28像素。每个patch的标签是其中心像素的标签。训练图像的patch用于训练网络。

在CNN中，成本函数定义为交叉熵损失，因此



其中是训练图像块，是对应的标签集。 训练图像patch的数量是n。 对于每个， 是等式1中定义的 CNN 的输出。CNN 使用随机梯度下降 (SGD) 和 dropout [25] 技术进行训练。 dropout 的目标是通过向训练样本引入随机噪声来避免过拟合。 这样在训练过程中，神经元的输出以 0.5 的概率被屏蔽。

**IV. EXPERIMENT 实验**

在六个公共手写历史文档图像数据集上进行了实验。

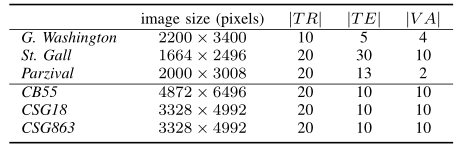
1. **Datasets 数据集**

这些数据集的性质非常不同。TheG. Washington由纸上用墨水用英文书写的页面组成，图像为灰度。另外两个数据集，即Parzival和St.Gall数据集由羊皮纸上用墨水书写的手稿图像组成，图像是彩色的。Parzival数据集由13世纪三位作家所写的页面组成。St.Gall数据集由一份用拉丁文写成的中世纪手稿组成。有关ground true的详细信息见[26]。

最近创建了三个布局更复杂的新数据集[27]。CB55数据集由14世纪的手稿组成，由一位作者用意大利语和拉丁语撰写。CSG18和CSG863数据集由11世纪的手稿组成，这些手稿用拉丁语书写。未指定两个数据集的作者数量。这三个数据集的细节见[27]。

在实验中，所有图像都使用缩放因子2^-3进行缩放。表一给出了六个数据集的训练集、测试集和验证集的详细信息。

表一：训练、测试和验证集的详细信息。TR、TE和VA分别表示训练集、测试集和验证集。



1. **Metrics 指标**

为了评估历史文档图像的页面分割方法，最常用的度量是精确度(precision)、查全率(recall)和像素级准确率(pixel level accuracy)。相反，除了标准的度量外，我们还采用了定义良好的度量，这些度量在常见的语义分割和场景分析评估中得到了广泛的应用。这些度量是像素精确度(pixel accuracy)和联合区域交集(IU)的变化。它们已在[28]中提出。因此，实验中使用的度量是：像素精度、平均像素精度、平均IU和频率加权IU(f.w.IU)。

为了获得这些指标，我们定义了变量：

* nc：类别的数量
* nij：预测属于类别j的类别i的像素数量，对于类别i：

- nii：正确分类的像素数（true positives）

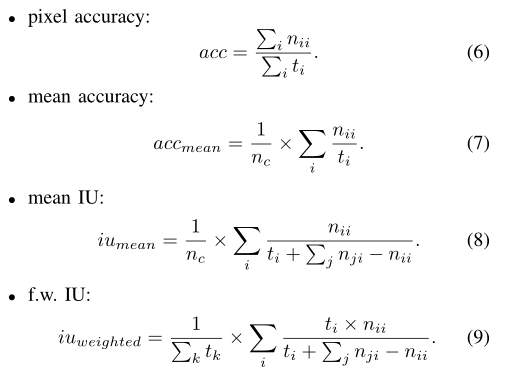
- nij：错误分类像素的数量（false positives）

- nji：错误未分类像素的数量（false negative）

* ti：类别i中的像素总数，即



通过定义的变量，我们可以计算：

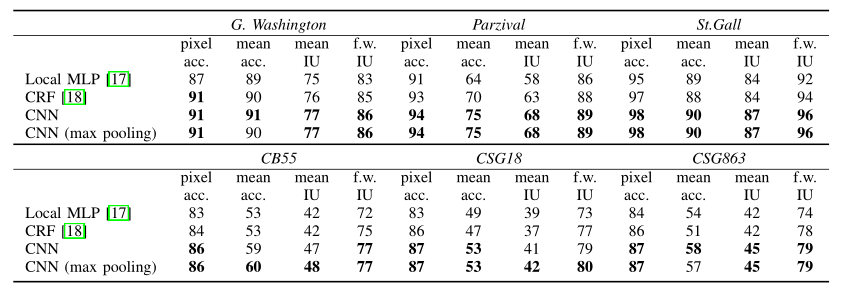


1. **Evaluation 评估**

我们将所提出的方法与我们以前的方法[17]、[18]进行了比较。与所提出的方法类似，超像素被认为是标记的基本单元。在[17]中，利用堆叠的卷积自动编码器以无监督的方式在随机选择的灰度图像块上学习特征。然后利用超像素的特征和标签训练分类器。利用训练好的分类器，将超像素分类为不同的类别。在[18]中，应用条件随机场(CRF)来为超像素标记任务联合建模局部和上下文信息。[17]中训练的分类器被认为是[18]中的局部分类器。然后利用局部分类器训练上下文分类器，将局部分类器的输出作为输入，输出给定标签的分数。利用局部分类器和上下文分类器，训练CRF来标记给定图像的超像素。在实验中，我们使用一个多层感知器(MLP)作为[17]、[18]中的局部分类器，并使用另一个多层感知器(MLP)作为上下文分类器。采用简单线性迭代聚类算法(SLIC)[29]生成超像素。SLIC算法相对于其他超像素算法的优越性在[17]中得到了证明。在实验中，每幅图像生成3000个超像素。

表II报告了三种方法中的像素精度、平均像素精度、平均IU和f.w.IU。结果表明，所提出的CNN方法比以往的方法具有更好的性能。图2给出了这三种方法的分割结果。我们可以看到，与其他方法相比，CNN在视觉上取得了更准确的分割结果。

表 II：仅使用局部 MLP、CRF 和提议的 CNN 进行超像素标记的性能（百分比）。



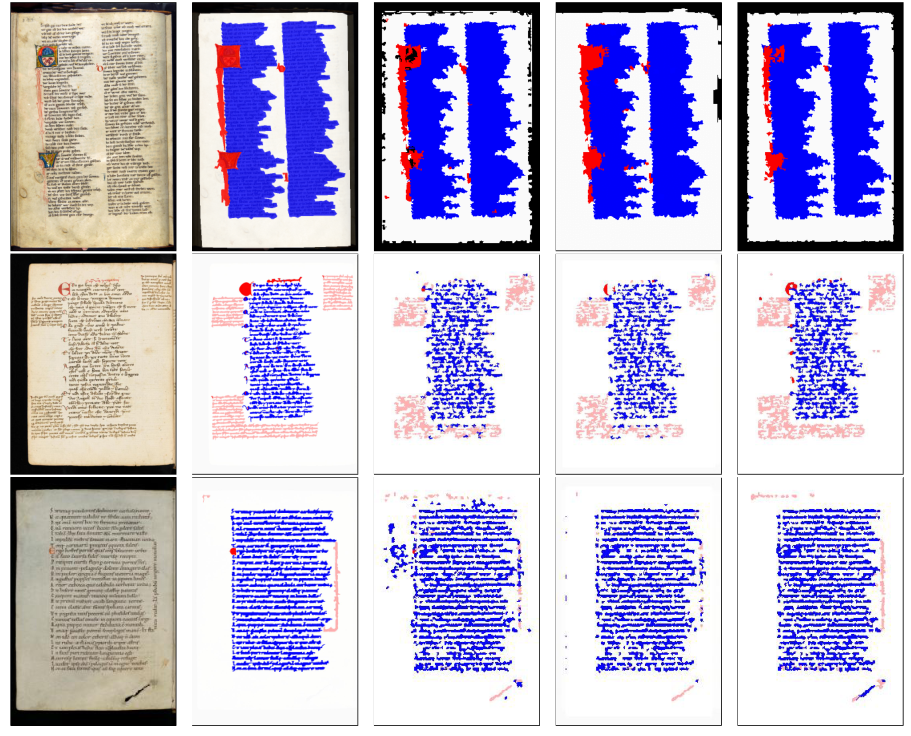


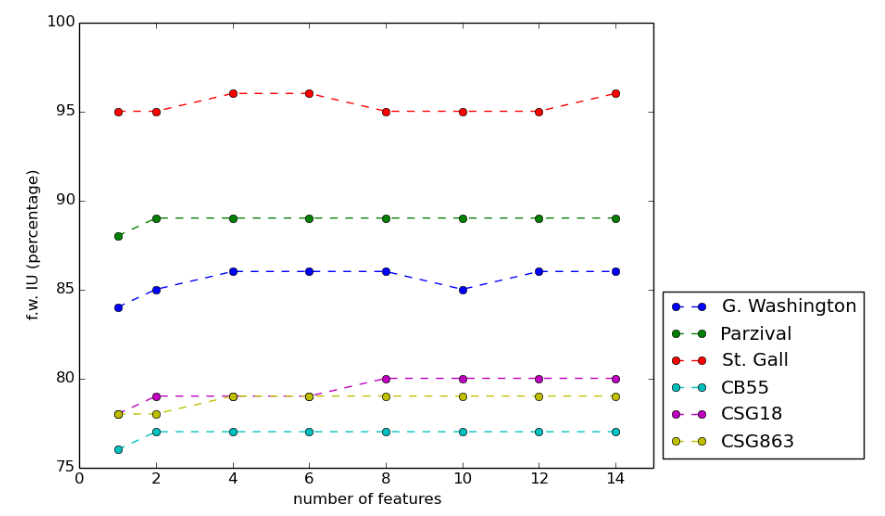
图2. 从上到下分别对Parzival、CB55和CSG863数据集的分割结果。颜色：黑色、白色、蓝色、红色和粉色分别用于表示：外围、页面、文本、装饰和注释。从左到右的列分别是：输入、ground true、局部MLP、CRF和CNN的分割结果。

1. **Max Pooling**

池化是CNN中广泛使用的技术。 最大池化是最常见的池化类型，用于减少表示的空间大小以减少网络参数的数量。 为了展示最大池化对分割任务的影响。 我们在卷积层之后添加了一个最大池化层。 池化大小为2×2像素。 表 II 报告了具有最大池化层的 CNN 的性能。 我们可以看到，仅在CB55数据集上，平均像素精度和平均IU略有提高。 一般来说，增加一个最大池化层并不能提高分割任务的性能。 原因是对于某些计算机视觉问题，例如自然图像中的对象识别和文本提取，特征的确切位置不如其相对于其他特征的粗略位置重要。 但是，对于给定的文档图像，要标记patch中心的像素，仅知道该patch中某处是否有文本是不够的，还要知道文本的位置。 因此，特征的准确位置有助于页面分割任务。

1. **Number of Kernels**

为了展示卷积层的核数对分割任务的影响。 我们将内核数定义为 K。 在实验中，我们设置K∈{1,2,4,6,8,10,12,14}。 图 3 报告了具有不同核数的一个卷积层 CNN 的 f.w. IU。我们可以看到，除了在 CS18 数据集上，当 K≥4 时，性能没有提高。

  
 图3. 一个卷积层 CNN 在不同数量的滤波器上的f.w. IU。

1. **Number of Layers**

为了展示卷积层数对页面分割任务的影响。 我们逐步添加卷积层，这样当前层上的内核比前一层多两个。 图 4 报告了具有不同卷积层数的 CNN 的 f.w. IU。 结果表明，层数不影响分割任务的性能。 但是，在 G. Washington 数据集上，层数越多，性能略有下降。 原因是与其他数据集相比，G. Washington数据集的训练图像较少。

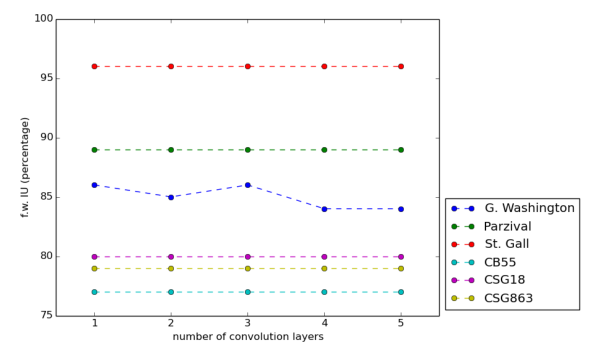


图4. CNN 在不同卷积层数上的 f.w. IU。

1. **Numbers of Training Images**

为了展示不同数量的训练图像下的性能。 对于每个数据集，我们在训练集中选择 N 个图像来训练 CNN。 对于每个实验，批次数设置为5000。图 5 报告了 不同N值下的f.w. IU，使得N∈{1,2,4,8,10,12,14,16,18,20}。我们可以看到，一般来说，当 N > 2 时，性能并没有提高。 然而，在G. Washington数据集，训练图像越多，性能略有下降。原因是与其他数据集相比，在G. Washington数据集页面更加多样化，ground true不太固定。

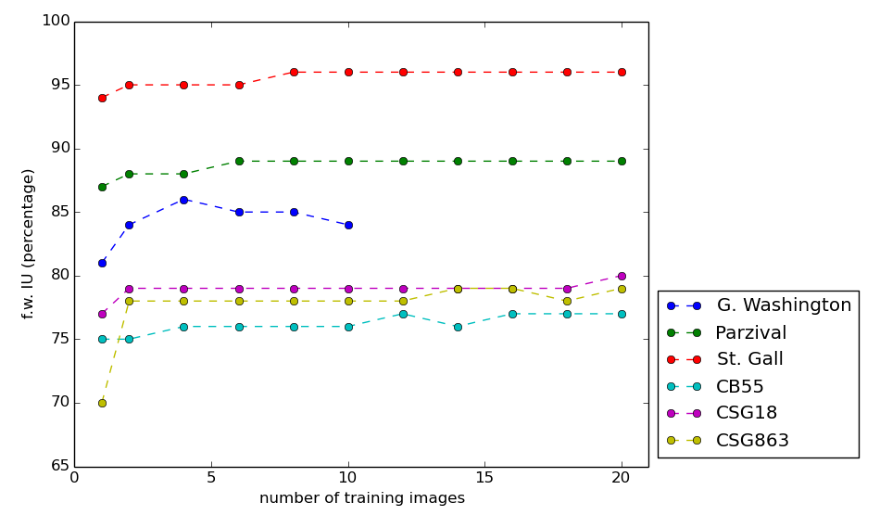


图5. CNN 在不同数量的训练图像上的 f.w. IU

1. **Run Time**

提出的CNN是用python库Theano实现的[30]。这些实验是在一台采用Intel Core i7-3770 3的PC上进行的。4 GHz处理器和16 GB RAM。平均而言，对于每幅图像，CNN需要大约1秒的处理时间。超级像素标记法[17]和CRF模型[18]分别需要2秒和5秒。

**Ⅴ. CONCLUSION**

在本文中，我们提出了一种用于手写历史文档图像的页面分割的卷积神经网络 (CNN)。 与依赖于使用手工特征训练的现成分类器的传统页面分割方法相比，所提出的方法直接从图像块中学习特征。 此外，特征学习和分类器训练合二为一。 在公共数据集上的实验表明，所提出的方法优于以前的方法。 虽然许多研究人员专注于将非常深的 CNN 架构应用于不同的任务，但我们表明，与其他网络架构相比，通过简单的一个卷积层 CNN，我们实现了可比的性能。

REFERENCES

1. C. Grana, D. Borghesani, and R. Cucchiara, “Automatic segmentation of digitalized historical manuscripts,”Multimedia Tools and Applications, vol. 55, no. 3, pp. 483–506, 2011.
2. S. S. Bukhari, T. M. Breuel, A. Asi, and J. El-Sana, “Layout analysis for arabic historical document images using machine learning,” inFrontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012, pp. 639–644.
3. K. Chen, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, “Robust text line segmentation for historical manuscript images using color and texture,” in2014 22nd International Conference on Pattern Recognition (ICPR). IEEE, 2014, pp. 2978–2983.
4. K. Chen, H. Wei, J. Hennebert, R. Ingold, and M. Liwicki, “Page segmentation for historical handwritten document images using color and texture features,” inFrontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014, pp. 488–493.
5. M. Bulacu, R. van Koert, L. Schomaker, and T. van der Zant, “Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen,” inNinth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 1. IEEE,2007, pp. 357–361.
6. T. V an Phan, B. Zhu, and M. Nakagawa, “Development of nom character segmentation for collecting patterns from historical document pages,” in Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. ACM, 2011, pp. 133–139.
7. C. Panichkriangkrai, L. Li, and K. Hachimura, “Character segmentation and retrieval for learning support system of japanese historical books,” in Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing. ACM, 2013, pp. 118–122.
8. B. Gatos, G. Louloudis, and N. Stamatopoulos, “Segmentation of historical handwritten documents into text zones and text lines,” in Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014, pp. 464–469.
9. R. Cohen, A. Asi, K. Kedem, J. El-Sana, and I. Dinstein, “Robust text and drawing segmentation algorithm for historical documents,” in Proceedings of the 2nd International Workshop on Historical Document Imaging and Processing. ACM, 2013, pp. 110–117.
10. A. Asi, R. Cohen, K. Kedem, J. El-Sana, and I. Dinstein, “A coarse-tofine approach for layout analysis of ancient manuscripts,” inFrontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on. IEEE, 2014, pp. 140–145.
11. Y . LeCun, L. Bottou, Y . Bengio, and P . Haffner, “Gradient-based learning applied to document recognition,”Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
12. Y . LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,”Neural computation, vol. 1, no. 4, pp. 541–551, 1989.
13. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” inAdvances in neural information processing systems, 2012, pp. 1097–1105.
14. T. Wang, D. J. Wu, A. Coates, and A. Y . Ng, “End-to-end text recognition with convolutional neural networks,” inPattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012, pp. 3304– 3308.
15. Y . Kim, “Convolutional neural networks for sentence classification,” EMNLP, pp. 1746–1751, 2014.
16. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, “Page segmentation of historical document images with convolutional autoencoders,” inDocument Analysis and Recognition (ICDAR), 2015 13th International Conference on. IEEE, 2015, pp. 1011–1015.
17. K. Chen, C.-L. Liu, M. Seuret, M. Liwicki, J. Hennebert, and R. Ingold, “Page segmentation for historical document images based on superpixel classification with unsupervised feature learning,” inDocument Analysis System (DAS), 2016 12th IAPR International Workshop on. IEEE, 2016, pp. 299–304.
18. K. Chen, M. Seuret, M. Liwicki, J. Hennebert, C.-L. Liu, and R. Ingold, “Page segmentation for historical handwritten document images using conditional random fields,” inFrontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on. IEEE, 2016, pp. 90–95.
19. J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in Proceedings of the eighteenth international conference on machine learning, ICML, vol. 1, 2001, pp. 282–289.
20. M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” inEuropean conference on computer vision. Springer, 2014, pp. 818–833.
21. K. Simonyan and A. Zisserman, “V ery deep convolutional networks for large-scale image recognition,”arXiv preprint arXiv:1409.1556, 2014.
22. C. Szegedy, W. Liu, Y . Jia, P . Sermanet, S. Reed, D. Anguelov, D. Erhan, V . V anhoucke, and A. Rabinovich, “Going deeper with convolutions,” inProceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
23. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” inProceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
24. V . Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” inProceedings of the 27th international conference on machine learning (ICML-10), 2010, pp. 807–814.
25. N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting.”Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.
26. K. Chen, M. Seuret, H. Wei, M. Liwicki, J. Hennebert, and R. Ingold, “Ground truth model, tool, and dataset for layout analysis of historical documents,” inIS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2015, pp. 940 204–940 204.
27. F. Simistira, M. Seuret, N. Eichenberger, A. Garz, M. Liwicki, and R. Ingold, “Diva-hisdb: A precisely annotated large dataset of challenging medieval manuscripts,” inFrontiers in Handwriting Recognition (ICFHR), 2016 15th International Conference on. IEEE, 2016, pp. 471–476.
28. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” inProceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
29. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P . Fua, and S. Süsstrunk, “Slic superpixels compared to state-of-the-art superpixel methods,”IEEE transactions on pattern analysis and machine intelligence, vol. 34, no. 11, pp. 2274–2282, 2012.
30. J. Bergstra, O. Breuleux, F. Bastien, P . Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y . Bengio, “Theano: a cpu and gpu math expression compiler,” inProceedings of the Python for Scientific Computing Conference (SciPy), 2010.