 DOI: https://doi.org/10.12349/iser.v5i12.4049

Archive Seal Detection and Extraction

Longtai Wang

School of Automation, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, China

Abstract

Archives contain valuable historical information and must be properly preserved. However, traditional archival materials are susceptible to damage from water, fire, and mold, making long-term storage difficult. To address this issue, a digital archive system has been established for management. Therefore, effectively storing, detecting, extracting, and utilizing archival information has become a focus of attention for people. This paper combines the analysis of specific requirements in the digitalization system of enterprise company archives and studies the seal extraction function on archive images. Aiming at the position recognition of archive seal images, YOLOv9 is proposed for seal image object detection; A seal extraction method using U²-Net was proposed for archive seal images. Through a two-stage approach of object detection and image segmentation, seals on archival images can be effectively detected and extracted.

Keywords

digital archive system; object detection; image segmentation

档案印章检测与提取

王龙泰

南京理工大学自动化学院，中国 · 江苏 南京 210094

摘　要

档案包含有价值的历史信息，必须妥善保存。然而，传统的档案材料容易受到水、火和霉菌的破坏，使得长期储存变得困难。为了解决这一问题，建立了数字档案系统进行管理。因此，有效地存储、检测、提取和利用档案信息已成为人们关注的焦点。论文结合对企业公司档案数字化系统中的具体需求分析，研究了档案图像上的印章提取功能。针对档案印章图像的位置识别，提出了使用YOLOv9进行印章图像目标检测；针对档案印章图像的印章提取，提出了使用U²-Net的印章提取方法。通过目标检测及图像分割这两个阶段的方式，可以有效地对档案图像上的印章进行检测与提取。

关键词

数字档案系统；目标检测；图像分割

1 引言

档案包含有价值的历史信息，必须妥善保存。然而，传统的纸质档案面临着诸多挑战，尤其是在物理保存方面。水、火、霉菌等自然灾害和环境因素的侵蚀，使得这些珍贵资料的长期保存变得异常困难，同时也限制了档案信息的广泛传播和利用。为了有效应对这些挑战，数字档案系统的建立显得尤为重要，它不仅能够保护档案免受物理损害，还能通过数字化手段实现档案信息的高效管理和便捷利用。

印章信息的准确提取对于自动化处理档案、提高工作效率具有重要意义。尽管国内外学者对图像的检测与提取进行了深入的研究，但对真实的旧档案中关键信息尤其是印章信息的提取却鲜有关注。针对这一不足，论文提出了一种针对旧档案印章的数字档案印章图像检测与提取技术。通过深入分析企业公司档案数字化系统的具体需求，本研究构建了一个新的档案印章图像数据集，并提出了有效的印章检测与提取方法。

论文包括“档案印章图像检测”“档案印章图像提取”两个主要章节，另外还有“引言”和“结论”。在“档案印章图像检测”中，描述了如何构建印章检测数据集以及介绍了所使用的方法及实验结果。在“档案印章图像提取”中，描述了如何构建印章提取数据集以及介绍了所使用的方法及实验结果。

本研究在该领域的贡献可以总结如下：

①构建了一个新的档案印章图像数据集；

②提出了有效的印章检测与提取方法。

总的来说，本研究为从旧档案中检测和提取印章提供了有效的方法，有效提高了档案图像的保存工作。

2 档案印章图像检测

2.1 数据集构建

在档案印章提取任务中，面临着一个显著的挑战：现

【作者简介】王龙泰（2000–），男，中国山东潍坊人，硕士，从事图像处理研究。

有的公开数据集无法满足算法训练的需求。档案图像通常包含敏感信息，其保密性质导致了可获得性较低。此外，对于印章图像提取这一细分领域，研究相对较少，缺乏相关的公开数据集。因此，构建一个符合实际需求的数据集是很有必要的。

使用标准印章制作软件 Sedwen 制作了 1500 个圆形印章和 1500 个方形印章图像 [1]，并利用 OpenCV 制作了相应的印章 mask 图像（GT 图）。从真实档案图像中随机挑选出 3000 张图片作为加盖印章的背景（背景图）。对 GT 图进行了色彩多样化处理（10 种红色、5 种蓝色、1 种紫色），并应用随机模糊和缺失效果，然后再经过随机旋转加盖在背景图上得到带有印章的档案原图（原图）。生成的档案图像如图 1 所示。

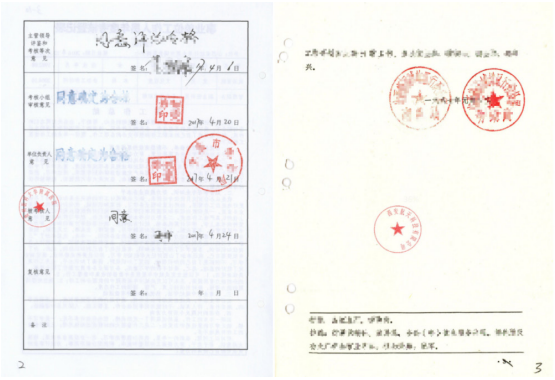

图 1 生成的档案图像

2.2 档案印章图像检测介绍

为了便于后续的模型训练，论文使用 LabelImg 开源工具在电脑本地对档案图形进行印章的手工标注。将全标注的数字化档案印章图像数据集分为训练集（7/10）和验证集（2/10）和测试集（1/10）。将训练图像和标签一起输入 YOLOv9[2] 检测模型中进行训练，经过 300 次训练得到模型的训练结果文件。

为了检测新存档图像中的印章，将图像输入训练好的模型进行推理，并标记和保存邮票图像的位置。印章检测总体框图如图 2 所示。在目标检测框的顶部显示检测到的印章类型和该类型正确的概率。根据检测框和推理得到的坐标信息，对印章区域进行裁剪，将印章区域图像分别存储在不同的文件夹中。

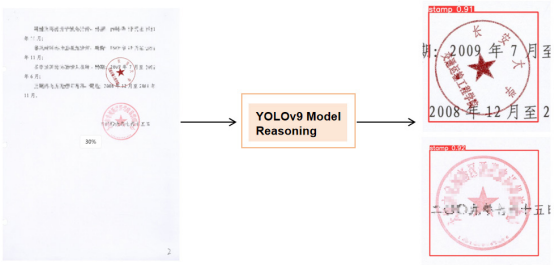

图 2 印章检测推理图

利用 YOLOv9 推理保存的位置信息，可以在随后的标签制作中轻松获得印章区域图像，从而通过简单的手动验证创建大型数字化档案邮票图像数据集。

总体而言，该方法为从档案中检测和提取印章图像提供了一种高效有效的方法，可以创建更全面的数字化档案数据集，用于保存和利用。

2.3 实验结果与分析

如图 3 所示，训练集和验证集的 box_loss、dfl_loss 和 cls_loss 都无限接近于 0，说明该模型可以准确响应印章的检测情况。表 1 中 Precision、Recall、mAP_0.5 和 mAP_0.5:0.95 这四个指标都无限地收敛于 1，表明具有很高的准确性。

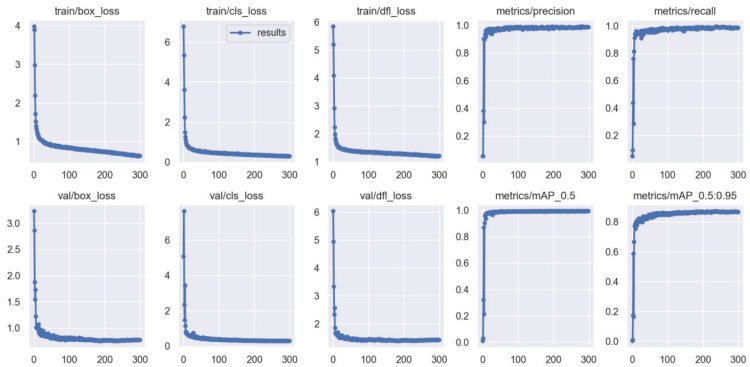

图 3 档案图像印章检测指标

表 1 档案图像印章检测记录

Classification	Training	Validation	precision	Recall	mAP@.5	mAP@.5:95
Circle	1352	357				
Rectangle	1275	324				
Total	2627	681	0.98842	0.98544	0.99402	0.8667

总体而言，YOLOv9 网络具有较高的综合检测率和准确率，非常适合档案图像的印章检测任务。

3 档案印章图像提取

3.1 数据集构建

沿用 2.1 中数据集构建的方法制作 GT 图。但背景图则是从真实档案图像中随机裁剪 1000 张 600 × 600 像素的背景区域（背景图），规定好背景图的大小，这样可以模拟第一阶段 YOLOv9 对印章区域识别后裁剪得到的印章小区域图像。然后再用 2.1 中同样的方法将印章加盖在背景图上得到印章区域图（原图）。

印章提取数据集由 3000 套印章图像组成，包括印章图像和印章掩码标签图像，即原图与 GT 图。其中，训练集包含了 2100 个集合，验证集包含了 900 个集合。印章提取数据集包含两种类型的印章，包括圆形、方形。印章原图、GT 图如图 4 所示。

图 4 印章原图、GT 图

3.2 档案印章图像提取介绍

档案图像处理中的印章提取任务，涉及对图像中特定区域的精确识别和分离，属于计算机视觉领域中的语义分割问题。在众多语义分割算法中，对常见算法如 FCN[3]、U-Net[4] 和 U²-Net[5] 进行了广泛的比较和实验。全卷积网络（FCN）作为早期的语义分割模型，以其端到端的训练方式和直接的像素级分类能力而受到关注。U-Net，以其独特的编码器 - 解码器结构和跳跃连接，在医学图像分割领域取得了显著成效。U²-Net 则是 U-Net 的改进版本，通过引入更深层次的特征融合，进一步提升了分割的精度和鲁棒性。经过综合对比，我们发现 U²-Net 算法在处理档案印章提取任务时，展现出了卓越的性能。U²-Net 不仅能够有效地捕捉到印章区域的细节特征，还能够在保持较高分辨率的同时，实现对印章区域的精确分割。因此，论文选择使用 U²-Net 算法作为论文的印章提取算法。

3.3 实验结果与分析

图 5 显示了使用 U²-Net 算法提取两种形状的印章的效果。图像的第一列是真实档案图像通过裁剪后得到的印章区域的原始图像，图像的第二列是通过 U²-Net 算法得到的标签掩模图像，图像的第三列是通过将原始图像与掩模图像叠加得到的印章前景的提取图像。从图 5 可以看出，印章提取模型具有更好的分割效果，印章前景图像可以更完整地与档案背景分离。

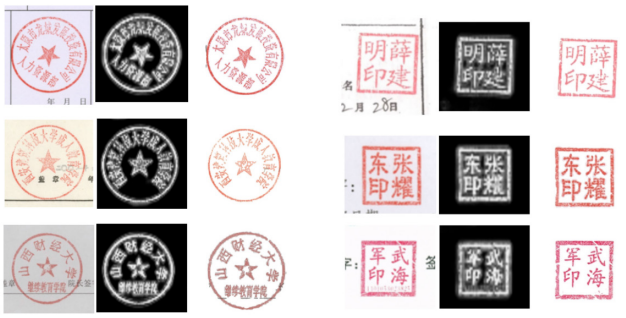

图 5 印章提取效果图

4 结语

综上所述，论文依据档案数字化处理的实际需求，提出了档案图像中印章检测与提取的有效方法。使用了两阶段的算法来完成该工作，通过 YOLOv9 来检测档案图像中的印章区域，通过 U²-Net 来对印章进行像素级提取分割。实验结果验证了论文算法在印章检测与提取领域的有效性，为档案数字化处理的进一步发展提供了技术支持。

参考文献

[1] Jin, X., Mu, Q., Chen, X., Liu, Q., Xiao, C. (2024). Digital Archive Stamp Detection and Extraction. In: Lu, H., Cai, J. (eds) Artificial Intelligence and Robotics. ISAIR 2023. Communications in Computer and Information Science, vol 1998. Springer, Singapore. https://doi.org/10.1007/978-981-99-9109-9_16.

[2] Wang C Y, Yeh I H, Liao H Y M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information[J]. 2024.

[3] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.

[4] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer International Publishing, 2015.

[5] Qin X, Zhang Z, Huang C, et al. U2-Net: Going deeper with nested U-structure for salient object detection[J]. Pattern recognition, 2020,106: 107404.