智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向近重复文本图像检索的三分支孪生网络

许柏祥, 刘丽, 邱桃荣

引用本文:
许柏祥,刘丽,邱桃荣. 面向近重复文本图像检索的三分支孪生网络[J]. 智能系统学报, 2022, 17(3): 515–522.
XU Boxiang,LIU Li,QIU Taorong. Near–duplicate document image retrieval based on three–stream convolutional Siamese network[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(3): 515–522.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202105018

DOI: 10.11992/tis.202105018
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220303.1131.002.html

面向近重复文本图像检索的三分支孪生网络

许柏祥, 刘丽, 邱桃荣
(南昌大学 信息工程学院, 江西 南昌 330031)

摘 要: 针对传统近重复文本图像检索方法需人工事先确定近重复文本图像之间存在的变换类型, 易受到人主观性影响这一问题, 提出一个面向近重复文本图像检索的三分支孪生网络, 能自动学习图像之间存在的各种变换。该网络输入为三元组, 包括查询图像、查询图像的近重复图像以及其非近重复图像, 训练时采用三元损失使得查询图像和近重复图像之间的距离小于查询图像与非近重复图像之间的距离。提出的方法在两个数据集上的 mAP (mean average precision) 分别达到 98.76% 和 96.50%, 优于目前已有方法。

关键词: 近重复文本图像; 图像检索; 三分支孪生网络; 三元损失函数; 图像变换; 三元组; 特征提取; 鲁棒性
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2022)03−0515−08

中文引用格式: 许柏祥, 刘丽, 邱桃荣. 面向近重复文本图像检索的三分支孪生网络 [J]. 智能系统学报, 2022, 17(3): 515–522.
英文引用格式: XU Boxiang, LIU Li, QIU Taorong. Near-duplicate document image retrieval based on three-stream convolutional Siamese network[J]. CAAI transactions on intelligent systems, 2022, 17(3): 515–522.

Near-duplicate document image retrieval based on three-stream convolutional Siamese network

XU Boxiang, LIU Li, QIU Taorong
(School of Information Engineering, Nanchang University, Nanchang 330031, China)

Abstract: In the traditional near-duplicate document image retrieval methods, the variations among the near-duplicate document images had to be manually identified beforehand, which can be easily influenced by human subjectivity. To solve this problem, we propose a three-stream convolutional Siamese network orienting toward the near-duplicate text-image retrieval, which can automatically learn the variation types among the near-duplicate document images. The input to this network is a triplet, consisting of a query image, its near-duplicate image, and its non-near-duplicate image. Using the triplet loss, the distance between the query image and its near-duplicate image is guaranteed to be smaller than that between the query and its non-near-duplicate image. This approach achieves promising results with the mAP of 98.76% and 96.50% on two datasets, respectively, thereby greatly outperforming the state-of-the-art near-duplicate document image retrieval methods.

Keywords: near-duplicate document image; image retrieval; three-stream convolutional Siamese network; triplet loss; image variations; triplet; feature extraction; robustness

随着通信技术和图像获取设备的发展, 图像的数量迅猛增长, 与此同时出现了大量的近重复图像。近重复图像是指从同一个目标在不同条件下拍摄得到的图像, 图像之间在光照、倾斜角度以及视角等各方面均有差异。本文主要研究目标为近重复文本图像检索。文本图像是一种特殊的图像, 其主体内容一般为文字。近重复文本图像检索在文本图像分析与理解中起着非常重要的作用, 而且在很多领域有重要应用, 例如, 在建立数字图书馆过程中, 大量纸质文档被扫描或拍摄, 并以图像格式存储。由于同一个纸质文档可能分散在多处, 经常出现将同一个纸质文档进行多次扫描的情况, 产生了近重复文本图像。这样一

收稿日期: 2021−05−13. **网络出版日期**: 2022−03−06.
基金项目: 国家自然科学基金青年项目 (61603256).
通信作者: 刘丽. E-mail: liuli_033@163.com.

来，不仅导致了冗余，占用了大量的磁盘空间，而且给后续建立索引等工作带来不便。因此，研究一种有效的近重复文本图像检索方法至关重要。

由于近重复文本图像之间在视角以及光照等方面存在差异，给近重复文本图像检索带来很大的挑战。一般来说，近重复文本图像检索分为两个步骤，即图像特征提取以及相似度计算。其中，特征提取是关键，提取的特征应对近重复文本图像之间存在的各种变换均具有良好的鲁棒性。基于所提取的特征，可以采用不同的相似度计算方法，例如欧氏距离、余弦距离等[1]。

本文将卷积神经网络[2](convolutional neural network, CNN) 应用于近重复文本图像检索这一研究领域，提出一种基于三分支孪生卷积神经网络的近重复文本图像检索方法，3 个分支共享权重。训练时网络输入为一个三元组，包括查询图像、查询图像的近重复图像以及查询图像的非近重复图像。通过采用三元损失使得查询图像和近重复图像之间的距离小于其与非近重复图像之间的距离。本文提出的方法能够自动学习近重复文本图像之间存在的各种变换，包括图像旋转、投影变换以及光照变化等。所提取的特征具有良好的鲁棒性，在新建的两个近重复文本图像数据集上取得了较好的结果。

1　研究现状

现有近重复文本图像检索方法主要采用手工特征描述图像。早期研究通常采用文本图像中的字符来描述图像[3-4]。例如，Spitz[3] 首先分割出文本图像中的所有字符，接下来将字符根据其形状进行编码，进而将图像表示为一个字符串，接下来采用 Levenshtein 距离计算图像之间的相似度。然而字符分割本身就是一个难题，经常出现字符黏连或者过分割等现象，导致后续检索失败。为了解决该问题，学者们直接利用文本图像中的词作为特征[5-7]，从而避免了字符分割这一问题。例如，Nakai 等[5] 提出一种名为 LLAH (locally likely arrangement hashing) 的近重复文本图像检索方法，将图像中每个词的中心点作为特征点，对于每个特征点，根据其与相邻特征点之间的空间位置关系定义多个仿射不变量。利用词作为特征只能应用于印刷体文本图像，在手写体文本图像中，如何准确地将词分割出来目前仍然是个难题。除了字符和词，文本图像的布局由于其具有较强的描述能力也被用于近重复文本图像检索[8-9]。提取图像的布局需要对图像进行分割，然而图像分割本身是一个极具挑战的问题。此外，SIFT (scale-invariant feature transform) 等局部特征[10] 由于其对图像变换，如旋转、缩放以及一定程度的仿射变换均具有良好的鲁棒性，也被用于近重复文本图像检索[11]。但正如文献 [12] 所述，文本图像中由于部分文字经常反复出现给关键点匹配带来很大干扰。近年来，CNN 迅猛发展，在很多领域具有广泛应用，例如图像分类[13]、目标检测以及跟踪[14] 等。目前已有学者将 CNN 应用于解决近重复非文本图像检索[15-19]，主要包括自然场景图像以及建筑物图像等。例如，Babenko 等[15-16] 采用 CNN 作为特征提取器，比较了网络中不同层在描述图像方面的差异，以及微调对近重复图像检索性能的影响，并且基于卷积层提出一种 SPoC(sum pooling of convolutions) 特征用于描述图像。Min 等[17] 设计了一个包含两种不同损失函数的网络，并且提出了一个新的网络层 RGMP。Husain 等[18] 提出了一种名为 ACTNET 的网络，其融合了网络中不同层来表示图像。Gordo 等[19] 将 R-MAC[20] 嵌入到孪生卷积神经网络中，最终以端到端的形式生成图像特征。

2　本文提出的方法

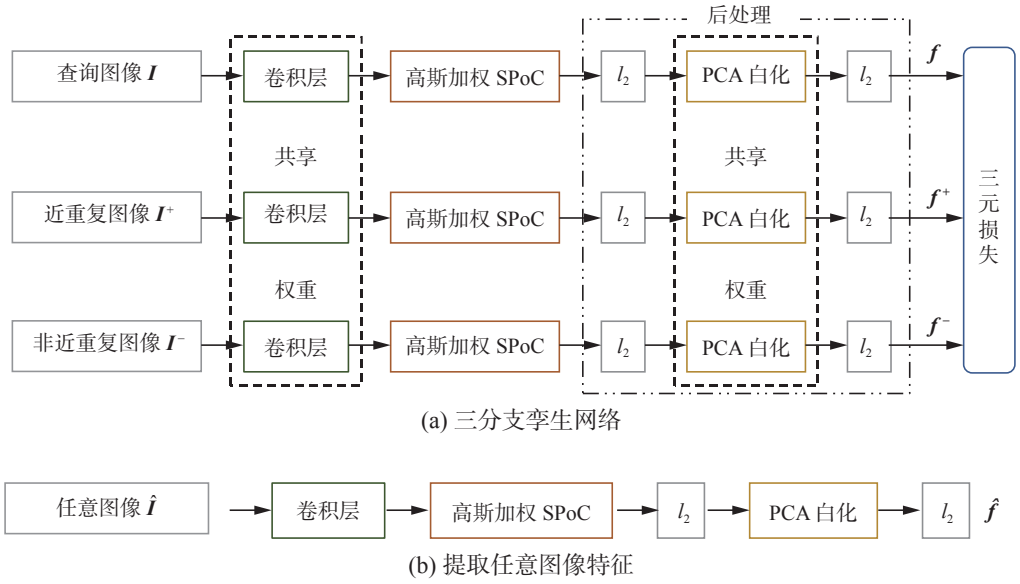
据我们所知，目前还没有研究将 CNN 应用于近重复文本图像检索这一领域。本文提出一种基于三分支孪生网络的近重复文本图像检索方法，创新点主要体现在如下两个方面：

1) 将 CNN 应用于近重复文本图像检索这一领域，能自动学习近重复文本图像之间存在的各种变换，包括图像旋转、投影变换以及光照变化等，进而避免了传统手工提取特征容易受到人主观性的影响，而且通常无法考虑到近重复文本图像之间存在的所有变换这一问题。

2) 所提出的三分支孪生网络专门针对近重复文本图像检索这一任务而设计。通过采用三元损失，可以保证查询图像和近重复图像之间的距离小于其与非近重复图像之间的距离。

2.1　网络结构

网络结构如图 1(a) 所示，网络输入为一个三元组，包括查询图像、查询图像的近重复图像以及查询图像的非近重复图像。每个图像分别经过一个分支，每个分支由卷积层、高斯加权 SPoC 池化以及后处理构成，3 个分支共享权重。网络中每个组成部分具体介绍如下。



(a) 三分支孪生网络

(b) 提取任意图像特征

图 1　本文提出的三分支孪生网络

Fig. 1　Our proposed convolutional siamese network

2.1.1　卷积层

每个分支首先中采用 VGG-16[21] 作为骨干网络, 并且去掉 VGG-16 中的全连接层, 具体结构如图 2 所示。假设输入一个 224×224×3 的图像, 则最终得到 512 个大小为 14×14 的特征图。

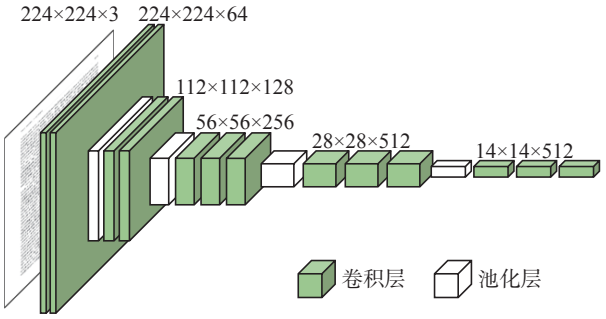


图 2　VGG-16 网络卷积层

Fig. 2　Convolutional layers in VGG-16

2.1.2　高斯加权 SPoC 池化

基于卷积层, 本文采用 SPoC 池化方法将其聚合为一个紧凑的特征向量。具体来说, 令 VGG-16 最后一个卷积层具有 N 个大小为 $H\times W$ 的特征图 $\boldsymbol{P}_n(n=1,2,\cdots,N)$, $\boldsymbol{P}_n[i,j]$ 代表第 n 个特征图中第 i 行第 j 列的值 $(1\leqslant i\leqslant H, 1\leqslant j\leqslant W)$。使用 SPoC 池化将得到一个 N 维的向量: $\boldsymbol{f}=[f_1\ f_2\ \cdots\ f_N]$, 其中 $f_n(n=1,2,\cdots,N)$ 定义为

$$f_n=\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{P}_n[i,j] \tag{1}$$

此外, 鉴于文本图像有效信息大部分都在图像中间部分, 在进行 SPoC 池化时加入高斯加权, 即将式 (1) 修改为

$$f_n=\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{G}(i,j)\boldsymbol{P}_n[i,j] \tag{2}$$

其中 $\boldsymbol{G}(i,j)$ 定义为

$$\boldsymbol{G}(i,j)=\exp\left(-\frac{\left(j-\frac{H}{2}\right)^2+\left(i-\frac{W}{2}\right)^2}{2\sigma^2}\right) \tag{3}$$

根据 3σ 准则, 将 σ 的值设置为图像中心和最近边界距离的三分之一:

$$\sigma=\frac{\min(W,H)}{6} \tag{4}$$

2.1.3　后处理

本文采用的后处理流程为 l_2 归一化、PCA 白化[22] 以及再次 l_2 归一化, 该特征后处理方法被广泛使用并取得了良好的性能[16]。

1) l_2 归一化: 给定特征向量 $\boldsymbol{f}$, 进行 l_2 归一化后为 $\boldsymbol{f}'$, 则:

$$\boldsymbol{f}'=\frac{\boldsymbol{f}}{\|\boldsymbol{f}\|} \tag{5}$$

2) PCA 白化: 可以降低特征之间的相关性, 在网络中可以利用减均值和全连接层来实现[19]。

2.2　三元损失函数

网络训练时采用三元损失函数, 定义如下:

$$\mathrm{Loss}(\boldsymbol{I},\boldsymbol{I}^+,\boldsymbol{I}^-)=\max(0,m+\|\boldsymbol{f}-\boldsymbol{f}^+\|^2-\|\boldsymbol{f}-\boldsymbol{f}^-\|^2) \tag{6}$$

式中: $(\boldsymbol{I},\boldsymbol{I}^+,\boldsymbol{I}^-)$ 为输入的三元组, 令 $\boldsymbol{I}$、$\boldsymbol{I}^+$、$\boldsymbol{I}^-$ 分别表示查询图像、查询图像的近重复图像以及查询图像的非近重复图像; $\boldsymbol{f}$、$\boldsymbol{f}^+$、$\boldsymbol{f}^-$ 分别为 $\boldsymbol{I}$、$\boldsymbol{I}^+$、$\boldsymbol{I}^-$ 经过网络之后得到的特征; m 是一个经验阈值。利用该三元损失函数, 可以保证查询图像与近重复图像之间距离小于查询图像与非近重复图像之间距离。

利用训练好的三分支孪生网络, 可以提取任

意图像的特征。具体来说，将图像$\hat{\boldsymbol{I}}$输入三分支中的任意一个分支，即可得到该图像的特征向量$\hat{\boldsymbol{f}}$，如图 1(b) 所示。

检索时，首先利用本文提出的网络提取查询图像以及图像库中图像的特征向量，接下来利用欧氏距离计算查询图像与图像库中每个图像的相似度，距离越小则认为相似度越高，并且按照相似度从高到低进行排序。

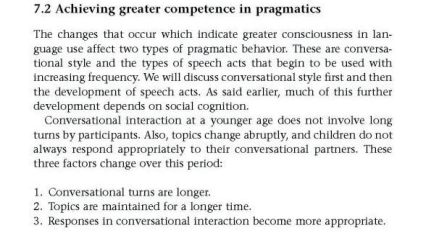
3　实验及结果分析

3.1　数据集

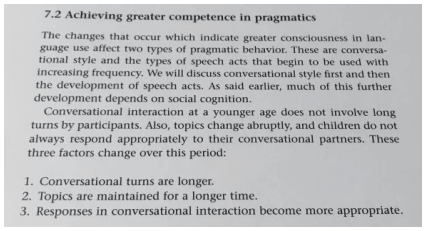
由于目前没有公开的近重复文本图像数据集，本文自建了 2 个近重复文本图像数据集，即 NDDoc_ENG 和 NDDoc_CN，分别包括英文和中文文本图像。文本图像传统获取方式为采用扫描仪进行扫描。随着智能手机的普及，手机拍摄图像越来越多。在本文自建数据集中，不仅包括扫描图像，而且包括手机拍摄图像。相比较扫描图像，手机拍摄图像经常存在旋转，光照不均以及投影变换等问题，给后续图像检索带来很大挑战。每个数据集的具体细节如下：

1) NDDoc_ENG 数据集

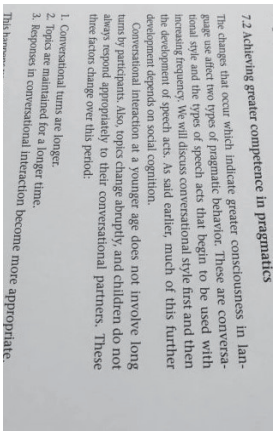
该数据集共包含 2 596 个英文文本图像，分为 518 组，每组有 5~6 个图像，为近重复文本图像。具体来说，每组第 1 个图像由 PDF 版电子书中的一页转换得到，接下来将其打印出来，用手机对其进行不同角度的拍摄获得 3~4 个图像。此外，利用扫描仪扫描得到 1 个图像。所有图像均为灰度图，格式为 JPG，尺寸为 750 像素×1 200 像素。图 3 给出了一组近重复英文文本图像示例。



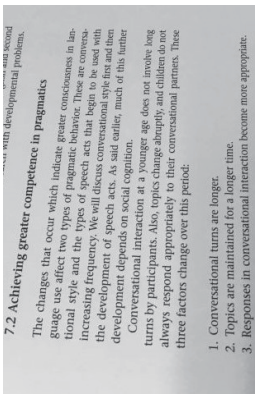
(a) 该组第一个图像



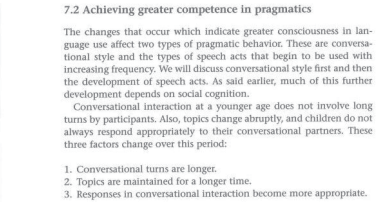
(b) 手机拍摄图像 1



(c) 手机拍摄图像 2



(d) 手机拍摄图像 3



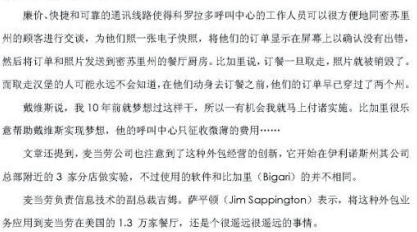
(e) 扫描图像

图 3　NDDoc_ENG 数据集上一组近重复图像示例

Fig. 3　A sample group of images from NDDoc_ENG dataset

2) NDDoc_CN 数据集

该数据集共包含 2 710 个中文文本图像，分为 542 组，每组有 5 个图像，为近重复文本图像，图像的获取方式和上述近重复英文文本图像相似。图 4 给出了一组近重复中文文本图像示例。

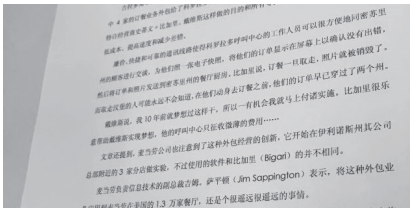

廉价、快捷和可靠的通讯线路使得科罗拉多呼叫中心的工作人员可以很方便地同密苏里州的顾客进行交谈，为他们照一张电子快照，将他们的订单显示在屏幕上以确认没有出错，然后将订单和照片发送到密苏里州的餐厅厨房。比加里说，订餐一旦取走，照片就被销毁了。而取走汉堡的人可能永远不会知道，在他们动身去订餐之前，他们的订单早已穿过了两个州。

戴维斯说，我 10 年前就梦想过这样干，所以一有机会我就马上付诸实施。比加里很乐意帮助戴维斯实现梦想，他的呼叫中心只征收微薄的费用……

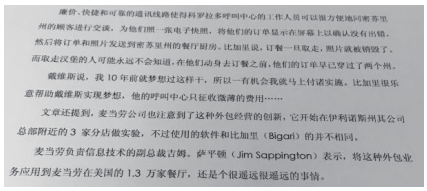
文章还提到，麦当劳公司也注意到了这种外包经营的创新，它开始在伊利诺斯州其公司总部附近的 3 家分店做实验，不过使用的软件和比加里（Bigari）的并不相同。

麦当劳负责信息技术的副总裁吉姆。萨平顿（Jim Sappington）表示，将这种外包业务应用到麦当劳在美国的 1.3 万家餐厅，还是个很遥远很遥远的事情。

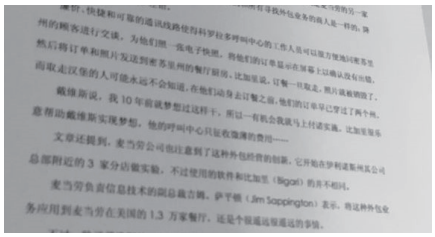
(a) 该组第一个图像



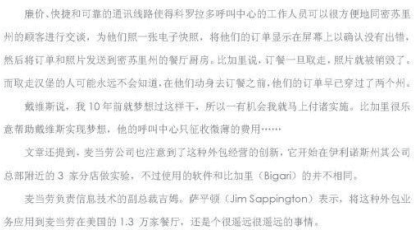
(b) 手机拍摄图像 1



(c) 手机拍摄图像 2



(d) 手机拍摄图像 3



(e) 扫描图像

图 4　NDDoc_CN 数据集上一组近重复图像示例

Fig. 4　A sample group of images from NDDoc_CN dataset

从图 3 和图 4 中可以看出, 近重复文本图像之间在旋转角度、光照以及视角等方面均存在差异。将每个数据集中的图像以组为单位, 按照 6:2:2 的比例随机划分为训练集、验证集与测试集, 训练集主要用于训练本文提出的三分支孪生网络, 验证集用于确定网络中的学习率等超参数, 接下来将在验证集上取得最优结果的参数应用于测试集上。

本文将训练集、验证集与测试集随机划分 5 次, 采用测试集上的平均 mAP[23] 作为评价指标。具体来说, 在测试集上, 用每组的第一个图像作为查询图像进行检索, 计算该查询图像和其他图像之间的相似度, 根据相似度大小由高到低排序。该查询图像的近重复图像排序越靠前, mAP 值越高。

3.2　实现细节

本文首先对训练集进行扩充。具体来说, 针对训练集上的每一组图像, 通过改变光照以及视角的方式进行数据增强, 使得每一组图像的数量达到 600~700 个, 训练集中图像总数扩充为 20 万。如 2.1 节所述, 每个三元组包括一个查询图像, 该图像的近重复图像和该图像的非近重复图像。为了产生三元组用于训练本文提出的三分支孪生网络, 首先在训练集上随机生成大量的三元组, 并且利用式 (6) 计算损失, 接下来根据损失大小进行降序排序, 选取前 M 个三元组作为网络的输入。这样的做法可以保证在反向传播时得到更大的梯度, 加快网络的训练。本文中, 实验 M 取值为 96000。

本文提出的三分支孪生网络中, 卷积层和 PCA 白化层是可训练的。为了初始化卷积层, 将训练集上的每一组图像看成一类, 训练一个用于分类的 VGG-16, 接下来用训练好的网络初始化卷积层。PCA 白化层利用训练集上学习的 PCA 白化矩阵进行初始化。本文提出的网络可以接受任意尺寸的图像, minibatch 设置为 32, 学习率设置为 0.001, 实验使用的 GPU 显卡型号为 NVIDIA Tesla P100。

3.3　不同池化方法对检索性能的影响

如图 1 所示, 本文提出的三分支孪生网络中, 采用 SPoC 这一池化方法。为了验证该池化方法的有效性, 将其与常用的 MAC (maximum activation of convolutions) 池化方法进行比较, 结果如图 5 所示。可以明显看出, 采用 SPoC 池化优于 MAC 池化。

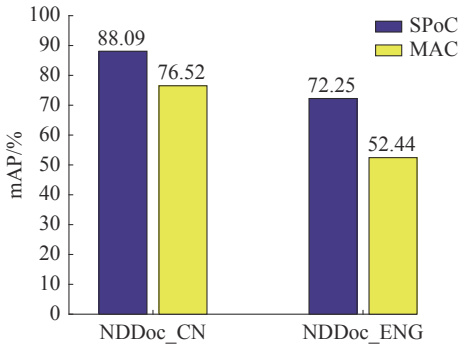


图 5　不同池化方法的性能比较

Fig. 5　Performance comparison between different pooling techniques

3.4　PCA 白化层对检索性能的影响

为了验证本文提出网络中 PCA 白化层的有效性, 比较了有无 PCA 白化层时的检索性能, 以及有 PCA 白化层时不降维与降至 256 维的结果, 具体如表 1 所示。可以看出采用 PCA 白化层可以大大提高 mAP, 主要原因在于其降低了特征之间的相关性。当降至 256 维时, mAP 比不降维时有所下降, 但依然高于无 PCA 白化层时的结果。

表 1　有无 PCA 白化层的性能比较

Table 1　Performance with and without PCA whitening layers　%

数据集	有PCA白化层		无PCA 白化层
	不降维	降至256维	
NDDoc_CN	**88.09**	79.07	77.75
NDDoc_ENG	**72.25**	63.27	51.05

3.5　m 取不同值时对检索性能的影响

图 6 中给出了 m 取不同值时的检索结果，可以看出，在数据集 NDDoc_CN 上，当 m=1.4 时，检索效果最好；在数据集 NDDoc_ENG 上，当 m=1.1 时，检索效果最好。

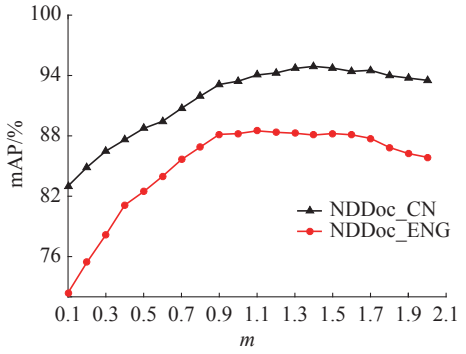


图 6　mAP 随 m 的变化曲线

Fig. 6　mAP w.r.t. different values of m.

3.6　训练不同数目 epoch 对检索性能的影响

图 7 中给出训练不同数目 epoch 时检索性能的变化，可以看出，在数据集 NDDoc_CN 上，训练 18 个 epoch 时取得了最高的 mAP，即 98.76%。而在数据集 NDDoc_ENG 上，训练 11 个 epoch 时取得了最高的 mAP，即 96.50%。

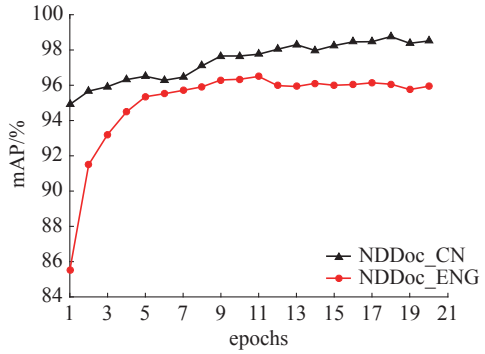


图 7　mAP 随 epoch 数目的变化曲线

Fig. 7　mAP w.r.t. different number of epochs

3.7　与现有方法比较

为了验证本文提出方法的有效性，将其与现有近重复图像检索方法进行比较。具体来说，将本文提出方法与文献 [5, 8, 11, 15, 17-19] 中的方法进行比较。其中文献 [5, 8, 11] 采用传统手工特征进行近重复文本图像检索。而文献 [15, 17-19] 中均采用基于 CNN 的方法。在两个数据集上的具体比较结果如表 2 所示。

表 2　本文提出方法与现有方法检索性能比较

Table 2　Performance comparison between our proposed approach and the other approaches　%

方法	NDDoc_CN	NDDoc_ENG
文献[5]	25.58	**97.75**
文献[8]	6.24	43.56
文献[11]	89.12	86.13
文献[15]	76.67	55.28
文献[17]	89.39	90.03
文献[18]	92.27	89.92
文献[19]	69.62	48.07
本文提出的方法	**98.76**	96.50

根据表 2，可以得出如下结论：

1) 对于传统基于手工特征的近重复文本图像检索方法：文献 [5, 8, 11] 分别利用文字之间的空间位置关系、布局以及关键点来表示图像。其中，文献 [5] 中的方法在英文数据集 NDDoc_ENG 上取得了最高的 mAP，然而，在中文数据集 NDDoc_CN 上的性能很差，主要原因在于对于印刷体中文文本图像来说，文字为方块字，大小基本一致，并且文字之间没有空格，进而使得文字之间的空间位置关系具有较低的区分度[24]。

2) 对于采用 CNN 的方法，根据文献 [15, 17-19] 的实验结果可知，它们在近重复非文本图像检索中具有良好的性能。然而，本文提出的方法在近重复文本图像检索方面优于其他几个方法。此外，文献 [19] 中的方法将 RMAC 池化嵌入到孪生卷积神经网络中，在两个数据集上的 mAP 分别为 69.62% 和 48.07%，远远低于本文提出方法。由此可以看出，SPoC 池化方法更适合文本图像。

3) 本文所提出的三分支孪生网络专门针对近重复文本图像检索这一任务而设计。所提取的特征对近重复文本图像之间存在的旋转、投影变换以及光照变化具有良好的鲁棒性，在中英文两个数据集上均取得了令人满意的效果，mAP 分别达到了 98.76% 和 96.50%。

3.8　大规模检索

为了进一步验证本文提出的近重复文本图像检索方法在大规模数据集下的性能，将测试集中分别增加 10 000、20 000 以及 30 000 个图像作为干扰，这些图像均来自 RVL-CDIP 数据集[25]，实验结

果如表 3 所示。从表中可以看出, 随着干扰图像数量的增加, mAP 有所下降, 但仍然具有较好的检索效果。

表 3　测试集加入干扰后的 mAP

Table 3　mAP of the proposed approach after adding different quantities of distractors to the test set　%

增加干扰图像数量	NDDoc_CN	NDDoc_ENG
10 000	96.29	92.26
20 000	95.49	91.37
30 000	95.11	90.82

4　结束语

本文提出了一种基于三分支孪生网络的近重复文本图像检索方法, 可以自动学习近重复文本图像之间存在的各种变换。训练时网络输入为一个三元组, 包括查询图像、查询图像的近重复图像以及查询图像的非近重复图像。网络中每个分支包括卷积层、高斯加权 SPoC 池化以及后处理, 三个分支共享权重。通过采用三元损失使得查询图像和近重复图像之间的距离小于其与非近重复图像之前的距离。训练好后, 可以将任意图像输入网络的一个分支提取特征, 所提取的特征对近重复文本图像之间存在的各种变换具有良好的鲁棒性。由于目前没有公开的近重复文本图像数据集, 本文新建了 2 个近重复文本图像数据集, 包括不同语言并且近重复文本图像之间在光照以及视角等方面均有较大的差异。本文提出的方法在两个数据集上均取得了令人满意的结果。

参考文献:

[1] 方涛, 陈志国, 傅毅. 神经网络多层特征信息融合的人脸识别方法 [J]. 智能系统学报, 2021, 16(2): 279–285. FANG Tao, CHEN Zhiguo, FU Yi. Face recognition method based on neural network multi-layer feature information fusion[J]. CAAI transactions on intelligent systems, 2021, 16(2): 279–285.

[2] WANG Zi, LI Chengcheng, WANG Xiangyang. Convolutional neural network pruning with structural redundancy reduction[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 14913−14922.

[3] SPITZ L. Duplicate document detection[C]//Proceedings of Document Recognition IV. San Jose, USA, 1997: 88−94.

[4] HULL J J. Document image matching and retrieval with multiple distortion-invariant descriptors[J].Proceedings of the international workshop on document analysis systems, 1994: 379−396.

[5] NAKAI T, KISE K, IWAMURA M. Real-time retrieval for images of documents in various languages using a web camera[C]//2009 10th International Conference on Document Analysis and Recognition. Barcelona: IEEE, 2009: 146−150.

[6] MORALEDA J. Large scalability in document image matching using text retrieval[J]. Pattern recognition letters, 2012, 33(7): 863–871.

[7] LIU Li, LU Yue, SUEN C Y, et al. Modeling local word spatial configurations for near duplicate document image retrieval[C]//2013 12th International Conference on Document Analysis and Recognition. Washington: IEEE, 2013: 235−239.

[8] LIU Li, LU Yue, SUEN C Y. Near-duplicate document image matching: a graphical perspective[J]. Pattern recognition, 2014, 47(4): 1653–1663.

[9] LIU Li, LU Yue, SUEN C Y. Document image matching using probabilistic graphical models[C]//Proceedings of the 21st International Conference on Pattern Recognition. Tsukuba, Japan, 2012: 637−640.

[10] BELLAVIA F, COLOMBO C. Is there anything new to say about SIFT matching?[J]. International journal of computer vision, 2020, 128(7): 1847–1866.

[11] VITALADEVUNI S, CHOI F, PRASAD R, et al. Detecting near-duplicate document images using interest point matching[C]//Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). Tsukuba: IEEE, 2012: 347−350.

[12] ROYER E, BOUCHARA F. Guiding text image keypoints extraction through layout analysis[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, 2017: 9−14.

[13] SUN Yanan, XUE Bing, ZHANG Mengjie, et al. Automatically designing CNN architectures using the genetic algorithm for image classification[J]. IEEE transactions on cybernetics, 2020, 50(9): 3840–3854.

[14] JI Yuzhu, ZHANG Haijun, ZHANG Zhao, et al. CNN-based encoder-decoder networks for salient object detection: a comprehensive review and recent advances[J]. Information sciences, 2021, 546: 835–857.

[15] BABENKO A, SLESAREV A, CHIGORIN A, et al. Neural codes for image retrieval[C]//European Conference on Computer Vision. Cham:Springer, 2014: 584−599.

[16] BABENKO A, LEMPITSKY V. Aggregating deep convolutional features for image retrieval[C]//Proceedings

of International Conference on Computer Vision. Santiago: IEEE, 2015: 1269−1277.

[17] MIN Weiqing, MEI Shuhuan, LI Zhou, et al. A two-stage triplet network training framework for image retrieval[J]. IEEE transactions on multimedia, 2020, 22(12): 3128–3138.

[18] HUSAIN S S, ONG E J, BOBER M. ACTNET: end-to-end learning of feature activations and multi-stream aggregation for effective instance image retrieval[J]. International journal of computer vision, 2021, 129(5): 1432–1450.

[19] GORDO A, ALMAZÁN J, REVAUD J, et al. End-to-end learning of deep visual representations for image retrieval[J]. International journal of computer vision, 2017, 124(2): 237–254.

[20] TOLIAS G, SICRE R, JÉGOU H. Particular object retrieval with integral max-pooling of CNN activations [EB/OL]//(2015–11–18)[2021–05–13]https://arxiv.org/abs/1511.05879.

[21] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL].(2014–09–4)[2021–05–13]https://arxiv.org/abs/1409.1556.

[22] ZHANG Shengdong, NEZHADARYA E, FASHANDI H, et al. Stochastic whitening batch normalization [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021: 10978−10987.

[23] LIU Guanghai, YANG Jingyu. Deep-seated features histogram: a novel image retrieval method[J]. Pattern recognition, 2021, 116: 107926.

[24] TAKEDA K, KISE K, IWAMURA M. Real-time document image retrieval on a smartphone[C]//2012 10th IAPR International Workshop on Document Analysis Systems. Gold Coast: IEEE, 2012: 225−229.

[25] HARLEY A W, UFKES A, DERPANIS K G. Evaluation of deep convolutional nets for document image classification and retrieval[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis: IEEE, 2015: 991−995.

作者简介：

许柏祥，硕士研究生，主要研究方向为文本图像分析与识别、深度学习。

刘丽，讲师，博士，主要研究方向为文本图像分析与识别、机器视觉、深度学习。主持完成国家自然科学基金项目 1 项。发表学术论文 21 篇。

邱桃荣，教授，博士，主要研究方向为模式识别与人工智能、机器学习与脑电信号处理和应用。主持完成国家级项目 2 项、省级项目 6 项。发表学术论文 39 篇。