文章编号 1004-924X(2017)03-0792-07

结合主体检测的图像检索方法

熊昌镇[1],单艳梅[1],郭芬红[2*]
(1.城市道路交通智能控制技术北京市重点实验室,北京 100144;
2.北方工业大学 理学院,北京 100144)

摘要:为解决图像背景复杂造成图像检索效果差的问题,提出了一种结合主体检测的图像检索方法。该方法首先训练用于目标检测的深度卷积神经网络模型,利用训练好的模型检测查询图像中的物体类别、类别概率和其所在区域坐标及特征。根据物体的类别概率和其所在区域的坐标判断图像主体后,在数据库中查找和主体类别相同的图像。计算查询图像与检索的同类别图像之间区域特征的余弦距离,结合类别概率对所有检索图像进行打分排序,返回分值最高的前 10 幅图像作为检索结果。最后在 VCO2007 数据集和自己收集的书页数据集上进行算法验证。实验结果表明,在随机选取的 1 000 幅测试图片检索结果的全正确率为 96.5%,比现有方法提升了 6.6 个百分点。本文方法可有效排除图像背景的干扰,得到更加准确的检索结果和定位精度。

关 键 词:深度学习;特征提取;图像检索;余弦距离
中图分类号:TP391.41 **文献标识码**:A **doi**:10.3788/OPE.20172503.0792

Image retrieval method based on image principal part detection

XIONG Chang-zhen[1], SHAN Yan-mei[1], GUO Fen-hong[2*]

(1. *Beijing Key Laboratory of Urban Intelligent Control, Beijing* 100144, *China*;
2. *School of Sciences, North China University of Technology, Beijing* 100144, *China*)
* *Corresponding author, E-mail: gfh@ncut.edu.cn*

Abstract: Aimed at the problem-poor result of image retrieval arising from the complexity of image background, a kind of image retrieval method combined with subject detection was put forward. This method has initially trained the deep Convolutional Neural Network (CNN) model used in object detection and used the model detection well trained to inquiry the object class, class probability and the coordinate and feature of region where it was placed in the image. After the image subject estimated in accordance with the object's class probability and coordinate of region where it was placed, the image similar to the subject in the database was found. The cosine distance of region feature between the image inquired and similar image retrieved was caculated, combined with the class probability to carry out grading and sorting for all images retrieved and returned the top 10 images with the highest scores to be as the retrieved result. Finally verification of algorithm was conducted on VCO2007 dataset and paper dataset collected by myself. The experiment result shows that the total

收稿日期:2016-12-02;**修订日期**:2017-01-14.
基金项目:北京市属高等学校青年拔尖人才培育计划资助项目(No. CIT&TCD201404009);科技创新服务能力建设—科技成果转化—提升计划项目(PXM2016_014212_000036)

accuracy for retrieved result of 1 000 test images is 96.5%, which has raised 6.6 percent points than the existing method. The proposed method can effectively exclude the disturbance of image background and get more accurate retrieved result and location accuracy.

Key words: deep learning; feature extract; image retrieval; cosine distance

1 引　言

多媒体技术在各个领域的应用越来越广泛，数字图像的数量随之急剧增长，如何从海量的图像中快速有效地检索出想要的图像内容已成为国内外研究热点。图像检索包括特征提取和匹配两大主要环节[1]，早期主要利用图像的底层视觉特征（如颜色、形状、纹理等）[2-3]或人工设计的特征（如尺度不变特征变换（Scale-invariant feature transform，SIFT）[4]、方向梯度直方图（Histogram of Oriented Gradient，HOG）[5]等局部不变特征，作为特征描述子对图像进行分类检索。上述方法的主要缺点是手工选取特征的质量多数靠经验和运气，此外还需要大量时间调节参数，而且与用户的高层语义概念之间也存在巨大差距，导致检索效果并不理想。

随着深度学习的出现，卷积神经网络（Convolutional Neural Network，CNN）广泛应用于图像分类[6-7]、物体检测[8-10]和语义分割等领域。不同于传统特征提取算法，CNN 可以通过训练提取图像的高级语义特征，并对图像平移、缩放、倾斜等变形有很高的抵抗力，可直接将原始的像素数据作为 CNN 的输入，具有更高的灵活性和普适性。Girshick 等人提出的 R-CNN（Regions with CNN features，R-CNN）方法将物体检测问题转化为分类问题[8]，使用 Selective Search[11]算法在图像上提取出约 2 000 个候选框，将它们输入网络，判断每一个候选框是否为物体，但存在很多重复计算，速度较慢。何凯明等提出的空间金字塔池化（Spatial pyramid pooling，SPP）网络使用空间金字塔[12]。该方法通过在整幅图像的特征映射上池化出每个候选框区域的特征，从而显著降低了检测时间。Girshick 后来提出的 Fast R-CNN 采用了同样的策略[9]，并将支持向量机替换为 Softmax 分类器，但是候选框提取方法不变。在后续 Faster R-CNN 算法[10]中，研究人员又加入了区域侯选网络（Region Proposal Network，RPN）直接得出物体位置和对应的类别得分，使速度和精度都得到了很大提升。由于使用 CNN 提取的卷积层或全连接层特征可以很好地表示图像的语义特征，故以图像作为输入的图像检索技术也取得了较大进展[13]。

图像检索的另一关键技术是特征相似度匹配，文献[14]通过计算查询图像和数据库中图像的余弦距离获得特征相似度，由于采用浮点数计算，在数据量较大时速度较慢。Kevin Lin 等人[15]使用哈希检索方法来解决此问题，即他们在预训练好的网络的倒数第二层和分类器中间插入一个全连接层直接学习二值哈希码。但由于在训练过程中没有考虑样本之间的相对位置关系，不能保证汉明距离近的点在语义上也相近。除此之外，还有人通过 CNN 学习图像的相似度量[16]。但以上算法主要应用于整幅图像，对于背景复杂或图像主体较小的图像，背景信息的特征较显著，检测效果较差。文献[14]在整幅图像上提取特征进行余弦距离排序，之后采用局部特征进行空间重排，从而提高了检索精度。但由于采用的是已训练好的网络或仅使用少量的检索图像来微调网络，区域匹配效果不太理想。

针对上述情况，受文献[14]的启发，本文提出了先检测图像的主体类别，然后在数据库中检索同类别图像，并采用类别分值和特征余弦距离相结合的方法对检索图像进行排序，取得了较好的结果。其主要贡献为：

（1）分析了所使用数据集对物体检测的 CNN 进行重新训练结果的影响。

（2）提出了一种使用图像主体类别代表整幅图像类别进行检索的方法，这样可以排除与查询图像完全不相关的图像，也避免了图像背景的干扰。在主体类别相同的图像范围内按照相似度进行排序，减小了运算量，提高了检索效率和准确率；RPN 检测出主体位置后提取区域特征，从而保证检测出的图像在语义上相似。

(3)通过图像主体类别分数和主体区域特征间的余弦距离这两项指标相乘得到排序的相似度指标，与仅用余弦距离相比，准确度得到提升。

2　基于图像主体的检索方法

本文算法的总体框架如图 1 所示，主要包括图像主体检测和特征匹配两部分，其中主体检测利用物体检测 CNN 提取图像中所有对象的类别、概率分值和类别区域特征，再通过类别区域与图像中心位置之间的关系判断图像的主体类别。特征匹配部分则主要是计算同类别图像与原图像的相似度，在此采用余弦距离和类别概率相结合的方法。详细的算法介绍见 2.1 至 2.3 节。

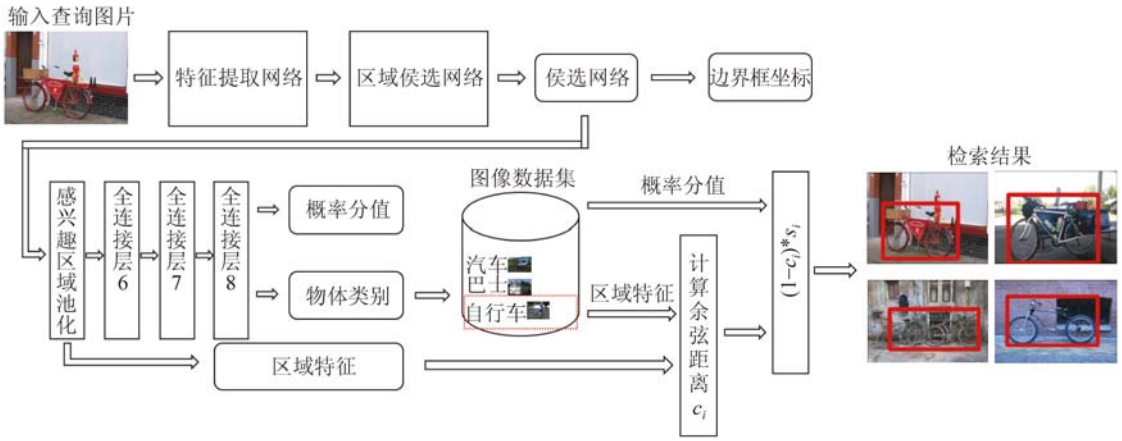


图 1　基于主体检测的图像检索算法框架

Fig. 1　Framework of image retrieval algorithm based on principal part detection

2.1　卷积神经网络模型

图像主体检测算法是利用图像的主体类别进行图像检索的，而不是直接将整幅图像作为查询对象。因此本文采用 Faster-R-CNN 在 ImageNet 数据集上预训练的网络模型[8]，并根据实际需求，重新微调网络。对于 VOC2007 数据集，可直接下载已经训练好的网络模型；对于书页检索数据集，首先对每一页图书页，选择区别于其他书页的页码、图画、表格等 70 类显著特征进行标记，将标记数据分别作为训练样本、评估样本和测试样本。首先使用牛津大学视觉几何组(Visual Geometry Group，VGG)提出的 16 层网络(记为 VGG16)模型参数对网络参数进行初始化，对未训练的参数采用随机初始化。接着，通过随机梯度下降法更新权重，进行不断迭代训练、评估和测试得到满足要求的网络模型。然后，将数据库中的图像作为 CNN 的输入，进行一次完整的前向运算。将类别得分大于给定阈值的候选框的类别分值、候选框位置及候选框区域的卷积特征存储到数据库中，一幅图像可能属于多个类别，供后续算法调用。

2.2　图像主体类别检测算法

查询图像包含背景区域和前景对象，当背景区域较大前景对象较小时，背景会对检索效果造成较大的影响，因此可先对查询图像进行主体类别检测，再将主体区域作为该图像的检索区域，以减少背景区域的影响。具体实现过程如下：

步骤 1：查询图像通过 VGG16 卷积神经网络后[8]，经过一个 RPN 可寻找出可能的侯选块，对每一个侯选区域都进行池化操作，得到区域特征和坐标。

步骤 2：将所有候选块的区域特征输入到后面的三层全连接层加 SoftMax 分类层的网络，得到每个候选区域的对象类别和类别概率值。

步骤 3：计算出第 i 类区域中心点(x_i, y_i)与图像中心点(x_0, y_0)的距离 d_i：

$$d_i=1-\frac{\sqrt{(x_i-x_0)^2+(y_i-y_0)^2}}{\sqrt{x_0^2+y_0^2}}. \tag{1}$$

步骤 4：使 d_i 乘以类别概率值 s_i 即得到第 i 类的分值 $d_i * s_i$，取分值最高的类作为查询图像的主体。

2.3　特征匹配算法

检测完查询图像的主体类别后，进行检索图

像同类别区域的特征相似性度量。为了便于比较各种不同度量方法[12-14]的效果，在此将余弦距离作为基本的度量方法。在 2.1 节将物体类别概率值大于给定阈值的图像存储为该类别图像，在这种情况下，若只采用余弦距离进行度量则会误检出类别错误的图像，因此本文采用描述同类别区域特征的余弦距离和表述类别的概率分值相结合的方法，来提高相似度量的精度。具体的算法如下：

步骤 1：从数据库中找出所有与原查询图像主体类别相同的图像。

步骤 2：计算查询图像与步骤 1 查找出的第 i 幅图像同类别区域特征的余弦距离 c_i；

步骤 3：计算结合 c_i 和类别概率分值 s_i 的相似性度量 $(1-c_i) * s_i$，对同类别的图像进行排序，取分值最大的前 n（实验中 $n=10$）幅图像作为检索结果。

3　实验结果

实验中采用包括飞机、巴士等 20 类物体 9 000 多张图像的 VOC2007 数据集和包括页码、显著标志等 70 类 5 000 张的书页图像数据集对算法进行验证。与文献[14]一样，将所有图像均缩放成短边为 600 pixel 的大小，所有实验均在具有 6 GB 显存的 Nvidia GTX980 TI 的 GPU 上运行。为了验证算法的有效性，设计了三个实验。实验一对比微调网络和不经过微调网络的结果，验证经过重新微调训练后的网络效果。实验二对比仅使用余弦距离和使用余弦距离与类别概率分值相结合的检索结果。实验三对比本文算法与文献[14]算法的结果。

3.1　网络微调训练实验

在书页图像数据集上进行微调网络对检索结果的影响实验。先提供一幅查询图像及感兴趣区域的坐标，采用 VGG16 模型提取整幅图像的 Conv5_3 层池化后的 512 维特征，计算余弦距离排在前 100 名的 100 幅图像，然后从选取的区域中提取池化后的 512 维特征，与这 100 幅图像所有侯选区域的特征进行比较，将余弦距离最小的侯选区域作为该图像的定位区域，然后对所有图像定位区域的余弦距离进行排序，将最相似的前 10 幅图像保留下来。检索结果如图 2（彩图见期刊电子版）和图 3（彩图见期刊电子版）所示，图中左侧蓝色框为查询图片，感兴趣区域用红色框圈出。其它 10 幅图像作为检索结果，定位框也使用红色框圈出，排列顺序为从左至右，从上到下，后续图例也采用相同的排列方式。图 2 是没有经过微调的检索结果，可见定位出的候选框区域偏差较大，图 3(a) 显示经过微调后的结果，其候选框位置基本正确。图 3(b) 为将候选框位置放大显示的结果。可以看出，经过微调的网络模型在候选框位置检测上有较大的提高。经过观察数据集后发现，这主要是因为在其他数据集上很少见到书页类图像，故训练网络对此类图像的检测效果并不是很理想。

图 2　未微调的网络模型的检索结果

Fig. 2　Retrieval results without fine tuning

(a)按照整幅图像显示的检索结果

(a)Retrieval results by showing the whole image

(b)将定位区域放大显示的结果

(b)Results by only showing the location area

图 3　微调网络模型后的检索结果

Fig. 3　Retrieval results for network model with fine tuning

3.2　不同相似度量的图像检索实验

为了验证结合余弦和类别概率方法的度量效果，对于给定的查询图像根据 2.2 节的算法先提取出图像的主体类别、类别概率和区域特征，然后计算查询图像和同类别图像候选框区域特征的余

弦距离，按照余弦距离对图像进行排序。结果见图 4，查询图像的主体类别为狗，使用余弦距离排序的结果中存在猫的图像且排在前面。因为它虽然和查询图像主体类别之间的余弦距离很小，但是类别概率分值较低，从而造成结果错误。若使用余弦距离结合概率分值作为相似度评判标准，检索结果如图 5 所示，结果中前 10 张图像候选框区域全部为狗，检索精度有了明显的提升。可见图 4 和图 5 中有些检索图像中存在多个定位框，这是因为在生成排序后将图片中包括同类别的物体都标注了出来，对后续图中的定位框也采用相同的标注方法。

此外，也可通过提高阈值来排除类别误判的影响，但会造成漏检。因此本文结合分数和余弦距离两项指标对同类图像进行排序，来提高检索的准确率。

图 4　仅使用余弦距离排序的结果

Fig. 4　Ranked result only using cosine distance

图 5　结合余弦距离和类别概率分值的实验结果

Fig. 5　Ranked result using cosine distance and class probability

3.3　与其它算法的对比实验

文献[14]也采用了 Faster-RCNN 训练的 VGG16 网络模型进行图像对象检索。它先计算整幅图像间的余弦距离对图像做初排序，然后再通过局部特征的余弦距离对结果进行重排。本文算法与它的不同之处如下：(1)先检测出图像的主体，避免了背景的干扰；(2)排序阶段采用余弦距离和类别概率相结合的排序方法来提高检索准确度。在 VOC2007 数据集上随机抽取 1 000 张图像作为测试图片，文献[14]算法有 101 张测试图片在返回前 10 张检索图像时出现错误，其全正确率(检索返回的前 10 幅图像全部为同类图像的正确率)为 89.9%；本文算法有 35 张测试图片的检索结果出现错误，全正确率为 96.5%，较文献[14]算法提高了 6.6 个百分点。图 6 和图 7 显示了文献[14]算法的检索结果，图 6 的原图像背景色彩比较单一，目标几乎占据了整幅图像，排序结果较好。图 7 的原图像背景较为复杂，虽然物体在图像中间，但检索结果较差。可以看出，对于背景相对单一或图像中物体较大的图像，文献[14]可以得到不错的检索效果。但是对于背景比较复杂或图像中物体较小的图像则效果较差。

图 6　色彩单一的图像检索结果

Fig. 6　Retrieval results of simple image hue

图 7　背景复杂的图像检索结果

Fig. 7　Retrieval results of image with complex background

本文算法对图 7 的查询图像进行检索，结果如图 8 所示。可以看出，本文算法的效果明显优于文献[14]的结果。在书页数据集上测试的结果如图 9 所示，与图 3 相比，可以看出候选框位置更加准确。

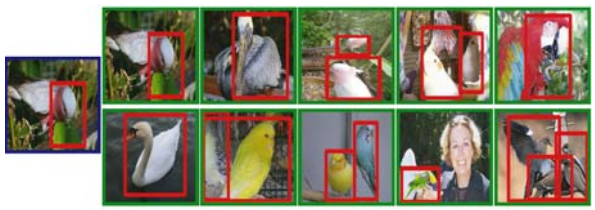


图 8　本文算法的图像检索结果

Fig. 8　Retrieval results of proposed method in image with complex background

(a)按照整幅图像显示的检索结果

(a)Retrieval results by showing the whole image

(b)将定位区域放大显示的结果

(b)Results by only showing the location area

图 9　本文算法在书页数据集上的检索结果

Fig. 9　Retrieval results of proposed method on page dataset

以上实验说明本文算法可以准确地检索出图像并定位出图像的主体类别区域,并且避免了图像背景干扰,得到更为准确的检索结果。

4　结　论

本文提出了一种结合主体检测的图像检索方法,利用训练好的 Faster R-CNN 模型检测图像的主体类别,在数据库中查找同类图像,综合类别的概率分值和区域特征的余弦距离对图像进行排序,并返回检索结果。实验中分析了经过微调与未经过微调的网络模型对检索结果的影响,对比了主体检测和其它图像检索算法的结果,给出了不同相似度量方法的比较实验。实验结果表明主体检测算法可以很好地避免图像背景的干扰,综合余弦距离和类别概率分值算法在随机选择的 1 000 幅图像上的检索结果的全准确率达 96.5%。与文献[14]算法相比,本文算法的全准确率提高了 6.6%,可以更准确地定位出图像中的物体。

由于本文仅选择了图像中单个对象作为图像的主体类别,没有考虑查询图像中多个类别及特征对检索结果的影响。因此下一步工作将分析图像中多个物体类别的综合检索方法,提高检索的准确率。

参考文献:

[1] 吴晓雨,何彦,杨磊,等. 基于改进形状上下文特征的二值图像检索[J]. 光学 精密工程,2015,23(1):302-309.
WU X Y, HE Y, YANG L, *et al.*. Binary image retrieval based on improved shape context algorithm [J]. *Opt. Precision Eng.*, 2015, 23(1): 302-309. (in Chinese)

[2] 刘丽,匡纲要. 图像纹理特征提取方法综述[J]. 中国图象图形学报,2009,14(4):622-635.
LIU L, KUANG G Y. Overview of image textural feature extraction methods [J]. *Journal of Image and Graphics*, 2009,14(4): 622-635. (in Chinese)

[3] 赵爱罡,王宏力,杨小冈,等. 纹理粗糙度在红外图像显著性检测中的应用[J]. 光学 精密工程,2016,24(1):220-228.
ZHAO AI G, WANG H L, YANG X G, *et al.*. Application of texture coarseness in saliency detection of infrared image [J]. *Opt. Precision Eng.*, 2016, 24(1): 220-228. (in Chinese)

[4] CHEN C C, HSIEH S L. Using binarization and hashing for efficient SIFT matching [J]. *Journal of Visual Communication and Image Representation*, 2015, 30: 86-93.

[5] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), NJ: IEEE, 2005, 1: 886-893.

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. *CoRR*, *abs*/1409.1556, 2014.

[7] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]. *In Advances in Neural Information Processing Systems* (NIPS), US: MIT Press, 2012: 1097-1105.

[8] GIRSHICK R, DONAHUE J, DARRELL T, *et al.*. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), NJ: IEEE, 2014: 580-587.

[9] GIRSHICK R. Fast R-CNN [C]. *Proceedings of the IEEE International Conference on Computer Vision* (ICCV), NJ: IEEE, 2015: 1440-1448.

[10] REN SH Q, HE K M, GIRSHICK R, *et al.*. Faster R-CNN: Towards real-time object detection with region proposal networks [C]. *In Advances in Neural Information Processing Systems* (NIPS), US: *MIT Press*, 2015:91-99.

[11] UIJLINGS J RR, SANDE K E A, GEVERS T, *et al.*. Selective search for object recognition[J]. *International Journal of Computer Vision*, 2013, 104(2): 154-171.

[12] HE K M, ZHANG X Y, REN SH Q, *et al.*. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(9): 1904-1916.

[13] BABENKO A, SLESAREV A, CHIGORIN A, *et al.*. Neural codes for image retrieval [C]. *Proceedings of the European Conference on Computer Vision* (ECCV), *Berlin*: *Springer*, 2014:584-599.

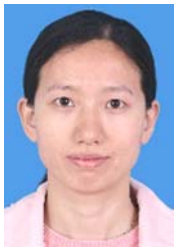
[14] SALVADOR, AMAIA, GIRO-I-NIETO, *et al.*. Faster R-CNN features for instance search [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) *Workshops*, NJ: IEEE, 2016.

[15] LIN K, YANG H F, HSIAO J H, *et al.*. Deep learning of binary hash codes for fast image retrieval [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) *Workshop*, NJ: *IEEE*, 2015:27-35.

[16] Han X F, LEUNG T, JIA Y Q, *et al.*. Matchnet: unifying feature and metric learning for patch-based matching [C]. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), NJ: *IEEE*, 2015: 3279-3286.

作者简介：

熊昌镇(1979－)，男，福建建宁人，博士，副教授，硕士生导师，2004 年于北方工业大学获得硕士学位，2007 年于中山大学获得博士学位，主要从事交通图像处理和机器学习方面的研究。E-mail: xczkiong@163.com



郭芬红(1980－)，女，山东肥城人，博士，讲师，2004 年于北京邮电大学获得硕士学位，2011 年于中山大学获得博士学位，主要从事图形图象处理方面的研究。E-mail: gfh@ncut.edu.cn

单艳梅(1992－)，女，河北唐山人，硕士研究生，2015 年进入北方工业大学学习，主要研究方向为深度学习和图像检索。E-mail: minions0315@163.com