计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):33−39 [doi: 10.15888/j.cnki.csa.006479]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

深度先验图像特征在城市遥感大数据中的应用①

申金晟, 池明旻

(复旦大学 计算机科学技术学院, 上海 201203)
(上海市数据科学重点实验室, 上海 201203)

摘　要: 图像特征提取始终是计算机视觉和图像处理的核心任务. 随着深度学习的快速发展, 卷积神经网络逐渐取代传统图像特征算子, 成为特征提取的主要算法. 本文针对城市遥感数据众包标记系统中的数据关联问题, 结合卷积神经网络和池化编码, 提出基于深度先验的图像特征提取方法. 该特征能有效聚焦室外图像近处物体, 并通过图像检索实验验证了其对室外图像的良好表征能力.

关键词: 图像特征提取; 城市遥感大数据; 和池化; 深度先验

引用格式: 申金晟,池明旻.深度先验图像特征在城市遥感大数据中的应用.计算机系统应用,2018,27(9):33–39. http://www.c-s-a.org.cn/1003-3254/6479.html

Application of Image Feature Extraction Based on Depth Prior in Urban Remote Sensing Big Data

SHEN Jin-Sheng, CHI Ming-Min

(School of Computer Science, Fudan University, Shanghai 201203, China)
(Shanghai Key Laboratory of Data Science, Shanghai 201203, China)

Abstract: Image feature extraction is always the core task of computer vision and image processing. With the rapid development of deep learning, the Convolutional Neural Network (CNN) has gradually replaced the traditional image feature operator and became the main algorithm for feature extraction. Combined with CNN and sum pooling, we propose a new image feature extraction algorithm based on depth prior aiming at the data association problem in the crowd sourcing labeling system for urban remote sensing data. The feature can effectively focus on the objects in the vicinity of outdoor images and verify their good characterization of outdoor images via image retrieval experiments.

Key words: image feature extraction; urban remote sensing big data; sum pooling; depth prior

1 引言

近年来遥感技术正朝着高时间分辨率、高空间分辨率、高光谱分辨率快速发展, 遥感数据已经进入到大数据时代[1]. 在国内城镇化快速推进的过程中, 遥感技术作为动态获取城市地形地貌、城市建设、土地利用与覆盖等信息的重要手段, 为城市公关安全、自然灾害预警、环境污染等提供必要的监测信息, 为相关政府部门对城市系统规划和决策提供科学依据.

在城市遥感大数据中, 一项重要任务为土地覆盖分类任务[2], 即为遥感数据 (尤其对高空间分辨率多光谱遥感图像) 中土地的物理覆盖类型 (诸如裸土、建筑、植被、水体等) 进行分类. 作为一项分类任务, 城市遥感数据面临着严重缺乏标记数据且标记困难的现状. 为此, 我们已提出了基于社交媒体数据的城市遥感众包标记系统[3], 其系统框架图如图 1, 其中矩形框部分为本文工作在框架中位置. 在该系统中, 我们使用主动学习技术, 从城市遥感图像中挑选最需要标记的样本点. 对这些待标记样本点, 在社交媒体照片数据集中

① 基金项目: 国家重点研发计划 (2016YFE0100300)
Foundation item: National Key Research and Development Program of China (2016YFE0100300)
收稿时间: 2017-12-06; 修改时间: 2017-12-27; 采用时间: 2018-01-16; csa 在线出版时间: 2018-07-26

(这里指带有 GPS 信息的城市室外照片) 通过 GPS 信息关联周围的城市照片, 经过图像聚类为众包标记者提供最能反映标记点地物特征的照片, 完成一轮迭代.

在该系统中, 我们需要面对数据关联问题, 即照片所呈现的内容不一定反映标记点内容. 为此, 我们希望在图像特征提取的过程中, 生成的特征不仅能有效表征图像, 更能反映图像中近景的特征, 从而减小数据关联带来的误差.

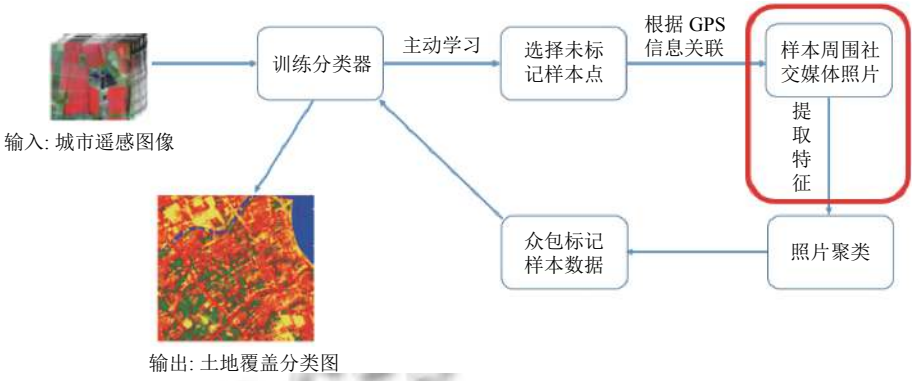


图 1 城市遥感数据众包标记系统框架图[3]

为此, 在本文中我们提出了基于深度先验的特征提取算法, 该算法在特征提取过程中加入了图像深度信息, 从而聚焦照片近处的物体特征; 同时, 该特征能很好地表征室外照片, 并被用于其他图像任务.

2 相关工作

图像的特征提取始终是计算机视觉领域的一个核心问题, 特征的好坏往往决定了目标任务的最终结果. 通常, 图像的特征提取可以分为两个步骤:

1) 提取图像原始特征. 这些特征往往是局部不变性特征, 具有较强的表达能力, 诸如 SIFT 特征[4]、SURF 特征[5]、HOG 特征[6]等等.

2) 为图像的原始特征进行编码. 通常, 图像的原始特征规模往往比较庞大, 因此需要进一步为这些特征进行编码, 起到去噪、压缩等功能, 使得图像特征的表达更为紧凑, 方便后续目标任务. 常见的图像特征编码算法包括 BoW 算法[7]、Fisher Vector 算法[8]、VLAD 算法[9]等. 本步骤并不是必须的, 在目标任务中直接使用图像原始特征也完全可以.

自 Krizhevsky 等人开创性地提出了 AlexNet[10]后, 卷积神经网络 (以下简称 CNN) 在图像分类、目标检测等一系列计算机视觉任务中都取得了突破性进展. CNN 的网络结构主要由卷积层和全连接层堆叠而成, 其卷积层的输出为卷积图张量, 全连接层的输出为固定维度的向量. Razavian 等人[11]使用预训练的 CNN 模型提取图像特征, 结合分类算法 (SVM 等), 在多个图像任务和数据集上都取得了优于传统图像特征 (SIFT 特征等) 的结果, 证明了 CNN 模型具有良好的特征表达能力和泛化能力.

Girshick 等人[12]在目标检测任务中, 直接使用 CNN 的全连接层输出作为候选区域的图像特征, 并结合 SVM 做分类. Babenko 等人[13]则在图像检索任务中使用 AlexNet 模型中的全连接层输出作为图像特征, 并用 PCA[14]降维. 作者分别比较了不同全连接层输出的效果, 同时使用自己收集的地标建筑物数据集微调 (fine-tune) 网络提升检索效果. Gong 等人[15]提出了 MOP Pool 特征提取算法, 该算法对图像的不同分块使用 CNN 提取全连接层输出作为图像原始特征, 结合 VLAD 算法对其编码, 并验证了 CNN 特征具有一定的尺度不变性和旋转不变性.

另一方面, 相较于使用全连接层输出, 卷积层输出保留了更多的图像空间信息和底层细节信息, 因此受到越来越多的关注. Ng 等人[16]使用卷积图张量结合 VLAD 算法, 在图像检索任务上得到了优于 MOP Pool 的结果. Tolias 等人[17]提出了 RMAC 池化方法, 通过对卷积图进行滑块做最大池化 (max pooling), 并对池化后的局部特征合并降维得到最终特征.

Babenko 等人[18]提出了和池化 (sum pooling) 方法替代最大池化, 对 CNN 中的特征图进行编码, 并设置中心先验权重来提升检索效果, 即 SPOC 算法. 同时,

该文对不同卷积层输出的特征张量进行对比, 验证越是靠后的卷积层提取的原始特征张量对图像的表达能力越强.

到目前为止的主要特征提取算法都是对图像进行全局地表征, 而在本文任务中, 我们需要在图像特征提取过程中有区别的对待不同距离的景物, 并更关注图像近景内容, 目前尚未有专门的算法. 因此我们需要提出一个新的特征提取算法.

3 基于深度先验的图像特征提取

本节将给出基于深度先验的图像特征构造算法. 算法框架图如图 2 所示. 对于输入图像, 我们将基于以下流程来提取图像特征:

1) 使用单目图像深度估计算法为图像还原深度信息.

2) 使用 CNN 卷积层部分提取特征图张量作为图像的原始特征.

3) 结合图像的深度先验和特征图张量计算卷积图权重, 并使用和池化方法对原始特征进行编码.

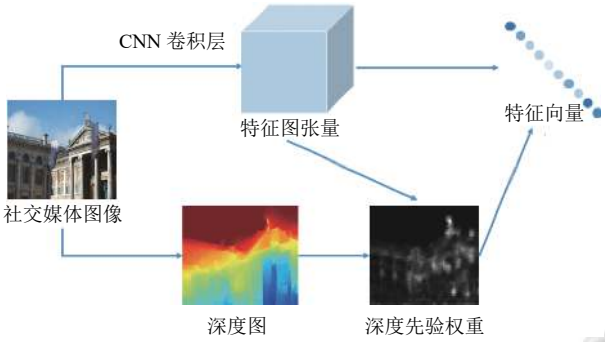


图 2 基于深度先验的图像特征提取算法框架图

3.1 图像深度先验提取

对于输入图像, 我们使用单目图像深度估计算法来还原图像的深度信息. 这里的深度信息指图像中拍摄的物体距离相机镜头的真实距离, 因此该任务是典型的不适定问题. 常见的深度估计算法主要包括基于马尔科夫随机场方法 (以下简称 MRF)[19–21]、基于大规模 RGB-D 数据集的无参估计[22,23]以及卷积神经网络方法[24,25]. 本文的目标任务在为室外城市照片提取特征, 故选择结合 CNN 的 MRF 算法来还原图像深度信息, 而得到的深度图即图像特征提取的先验信息.

3.2 图像原始特征的构造

由先前的论述可知, CNN 卷积层的输出比全连接层输出保留了更多的图像空间信息, 具有更强的图像表达能力. 此外, 在 CNN 处理图像时, 往往需要对输入图像预处理 (裁剪或形变缩放), 将其规约到固定大小 (比如 VGG 网络中规定图像输入尺寸为 224×224), 以保证最后的全连接层参数一致. 使用卷积层输出作为图像的原始特征则意味着不再需要将输入图片规约到固定大小. 常见的输入图像规约预处理有三种:

1) 双步缩放: 将图像分别在长宽两个维度上缩放至固定值, 比如 224×224.

2) 单步缩放: 将图像在长或宽上缩放至固定值, 另一维度上同比例缩放.

3) 图像原始尺寸输入.

显然, 图像原始尺寸输入即不会改变图像的空间信息, 也不会损失图像的细节信息. 本文即采用原始图像输入来提取原始特征.

另一方面, 使用原始图像输入会使不同尺寸的图像在同一个卷积层所得到的卷积图大小不同. 这里, 我们对得到的卷积图再使用单尺度的 Spatial Pyramid Pooling[26]进行规约, 使得所有图像所获得的最终原始特征尺寸相同, 使用的池化方法为最大值池化.

3.3 基于深度先验的和池化编码

在使用 CNN 获得了图像的原始特征后, 我们提出一种结合图像深度先验的池化方法来提取最终的图像特征. 整个和池化编码过程如图 3, 其中, ×表示每个卷积特征图分别与深度先验权重做和池化.

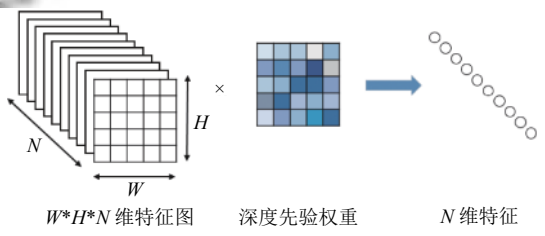


图 3 基于深度先验的全局和池化示意图

对于输入图像, 在使用某种卷积神经网络对其进行训练后, 其中某一个卷积层 L 的输出结果为特征张量记为 $F \in R^{W \times H \times N}$, $W \times H$ 为特征图的长和宽, N 为该卷积层的卷积核数量. 定义由第 k 个卷积核产生的特征图为 $f_k \in R^{W \times H}$, 特征图中位置 (x, y) 处的值为 $f_{x,y,k}$, 构造卷积层 L 所输出的第 k 个特征图的特征值 φ_k 如下:

$$\varphi_k = \sum_{y=1}^{H}\sum_{x=1}^{W} \omega_{x,y} \cdot d_{x,y} \cdot f_{x,y,k} \tag{1}$$

其中, $w_{x,y}$ 为特征图中点 (x, y) 处的响应权重, $d_{x,y}$ 为该点的深度权重. 对于响应权重矩阵 $W \in R^{W \times H}$, 我们使用特征图 f_k 来构造:

$$W = \sum_{k=1}^{N} f_k \tag{2}$$

对于深度权重, 我们使用图像的深度信息为权重赋值. 我们先将输入图像的深度图缩放至 $W \times H$, 则:

$$d_{x,y} = \left(\frac{d_{\max} - d(x,y)}{d_{\max} - d_{\min}}\right) + \varepsilon \tag{3}$$

其中, $d_{\max}$ 为深度最大值, $d_{\min}$ 为深度最小值, $d(x, y)$ 为 (x, y) 处的深度信息. ε 为一个极小量, 保证单目图像深度估计过程中对极远处的误判. 该权重值使图像特征向近景倾斜.

对于获得的权重矩阵, 我们使用 L2 范数归一化, 即:

$$V = \frac{V}{\|V\|_2} \tag{4}$$

其中, 权重矩阵 $V \in R^{W \times H}$, 定义如下:

$$v_{x,y} = \omega_{x,y} \cdot d_{x,y} \tag{5}$$

我们对卷积层 L 输出的所有 N 个特征图经过上述和池化计算, 获得该卷积层的一个 N 维的特征向量 φ, 并使用同维度的 PCA 白化, 并对得到的白化特征再进行 L2 范数归一化, 最终得到 N 维图像特征.

作为比较, 我们将 SPOC 算法摘录如下:

$$\varphi_k = \sum_{y=1}^{H}\sum_{x=1}^{W} a_{x,y} \cdot f_{x,y,k} \tag{6}$$

其中, $a_{x,y}$ 为高斯中心先验, 其权值设置如下:

$$a_{x,y} = \exp\left(-\frac{\left(y-\frac{H}{2}\right)^2 + \left(x-\frac{W}{2}\right)^2}{2\sigma^2}\right) \tag{7}$$

其中, σ 为分布协方差, 设置为特征图中心距最近边界长度的三分之一. 可以看到 SPOC 算法是在 sum pooling 的基础上加上了高斯中心先验, 并不能有效反映图像中的图标物体, 更不能反映近景的特征.

4 实验

我们的目标任务在于图像特征提取过程中更关注近景, 但该任务并不是一个标准的图像任务, 所以没有一个标准数据集可以进行对比. 但在一定程度上, 本文任务和基于内容的图像检索 (Content Based Image Retrieval, CBIR) 任务相近, 二者都要求特征对图像中的内容物体有良好的表征能力. 因此, 我们将在图像检索任务上与其他特征提取算法进行对比.

4.1 数据集

我们使用 Oxford5k Buildings dataset[26]作为实验数据集. 该数据集包含了 5063 张图片和 55 个查询. 图片采集自 Filckr 上反映牛津地区的照片, 本身即为社交媒体图像, 而查询内容均为牛津地标建物的室外照片, 与我们的社交媒体数据集中城市室外照片非常相近.

4.2 实验设置

我们使用 VGG16-NET[27]模型作为提取图像原始特征的网络结构. 所有网络参数均训练自 ImageNet 数据集[28], 不对网络做重训练或微调. 我们使用 VGG16 中最后一个卷积层 (即 conv5_3) 输出的特征张量作为原始图像特征. 在深度权重的计算中, ε 设置为 0.0001. 该卷积层共有 512 个卷积核, 得到 512 个卷积特征图. 在进行池化编码后, 最终获得一个 512 维的图像特征.

对于查询区域的特征提取, 通常将查询区域从原图中进行剪裁, 再对剪裁图像进行特征提取. 这里我们借鉴 SPP-NET[29]中从图像到卷积特征图之间区域位置映射关系, 直接从全图的卷积图上剪裁查询区域的特征再进行池化编码.

在 SPOC 中提到, 数据集中查询内容往往靠近图像的中心区域, 因此设置了高斯中心先验来减少近处景物的噪声 (比如近处的植被、门框等). 为此, 我们将靠近图像边界的近景的深度信息设置为极远处从而提高图像检索的准确率.

在相似度上, 我们使用余弦距离来计算图像间的相似度. 在实验结果的评价上, 使用平均正确率均值 (mean Average Precision, mAP) 作为评价标准, 其中图像检索的结果依据余弦距离进行排名.

4.3 权重矩阵

图 4 展示了单目图像深度估计获得的图像深度图,上排为输入的室外城市照片, 下排为对应的使用单目图像深度估计算法还原的深度图. 这里使用的深度估计算法为自行设计的结合 CNN 的条件随机场算法, 在 Make3D 数据集[19]上的相对误差为 0.276. 我们可以看到, 通过深度估计所获得的深度先验能基本反映出

图像中物体的基本轮廓和远近关系, 从而帮助我们获取图像的近处景物.

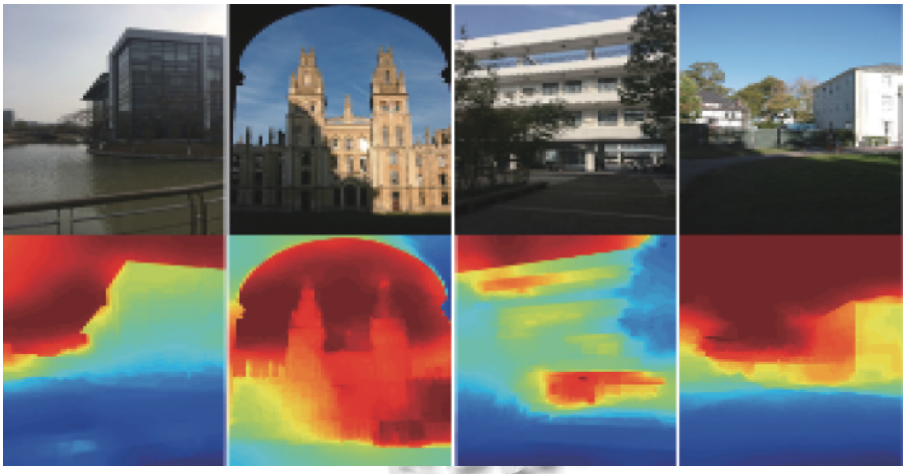
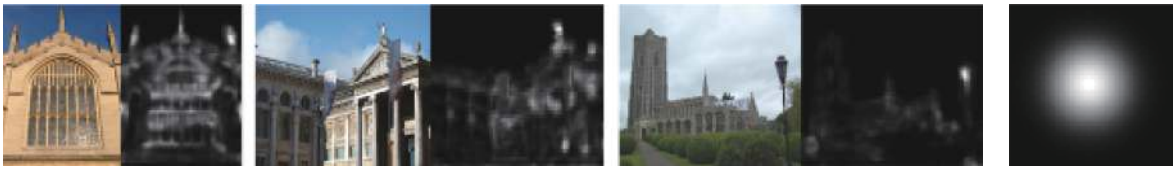


图 4 单目图像深度估计可视化结果

图 5 展示了本文设计的权重矩阵, 图 5(a) 是输入的室外城市照片和对应的深度先验权重矩阵, 图 5(b) 是 SPOC 高斯中心先验, 对所有输入图像, SPOC 的全局和池化权重矩阵相同. 这里的权重矩阵未经中心化去近景操作. 作为对比, 我们展示了 SPOC 算法的高斯先验中心权重. 我们可以看到, 本文基于深度先验的权重矩阵能有效地反映出图像中的主要物体, 同时, 深度先验可以使特征聚焦近处景物.



(a) 城市照片和深度先验权重 (b) SPOC 高斯中心先验权重

图 5 全局和池化权重矩阵

4.4 实验结果与分析

表 1 展示了不同图像特征提取算法在 Oxford5k Buildings 数据集上图像检索的结果, 带*表示使用全图查询的结果. 该表中结果除本文算法外其余结果皆来自原论文. 该结果表明本文特征提取算法对室外照片有良好的表征能力. 同时, 我们可以看到使用 CNN 提取的图像特征相比于传统特征总体上有所提高. 此外, 通过池化编码 CNN 卷积层输出的特征张量能大幅减小特征规模, 极大地降低了后续任务的计算量, 同时, 得益于深度学习的 GPU 加速, 图像的原始特征规模和计算时间也大幅降低. 本文方法在 Oxford5k Buildings 数据集上取得了较好的结果, 主要得益于以下几点.

首先, Oxford5k Buildings 数据集的查询目标都是室外建筑, 而查询结果也是室外图片, 图像的深度估计算法可以较好地得到物体的轮廓和深度信息, 其结果本身就能有效反映图像中的主要物体. 其次, 查询区域特征全部为特征图剪裁保证了图像原始特征的统一性. 最后, 原始图像输入最大程度地保证了原始特征的空间信息和细节信息, 这一点我们将进行对比实验.

表 1 Oxford5k Buildings 数据集图像检索结果

方法	特征维度	原始特征	mAP
BoW[7]	20 000	SIFT	0.364
VLAD[9]	32 000	SIFT	0.555
Neural codes[13]	512	AlexNet	0.435
Sum pooling[18]	256	VGG19	0.531
SPOC[18]	256	VGG19	0.657*
RMAC[17]	512	VGG16	0.669
本文方法	512	VGG16	0.672

图 6 则展示了对输入图像进行缩放的实验结果.

我们将图像缩放到 224×224, 使用的是预训练的 VGG16 网络, 得到 conv5_3 卷积层输出的 14×14 特征图, 并分别使用 sum pooling、SPOC 和深度先验方法提取特征, 最终特征的维度为 512 维. 可以看到, 实验结果都分别低于表 1 中的结果, 说明对于较高分辨率的图像而言, 在输入时进行缩放 (主要是缩小) 会使 CNN 图像特征的表达能力大幅下降, 一方面是缩小到固定尺寸破坏了原有图像的空间信息, 同时大尺度地缩小图像使得图像的细节信息被模糊.

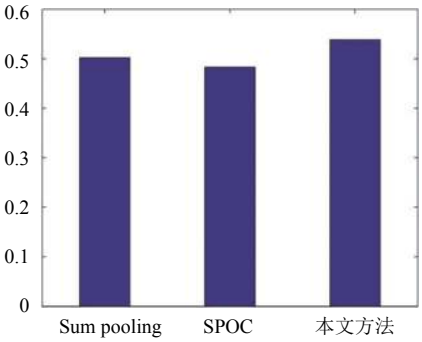


图 6 对输入图像缩放的检索结果

5 结论

本文提出了基于深度先验的图像特征提取算法, 该算法可以有效聚焦近处景物特征. 而图像检索实验验证了该特征能有效表征室外照片, 可以被使应用于其他相关图像任务中. 同时, 我们发现在图像特征提取过程中, 原始图像输入对特征的表征能力有极大地提高, 并且再一次验证了 CNN 模型的良好泛化能力和表征能力.

参考文献

1 宋维静, 刘鹏, 王力哲, 等. 遥感大数据的智能处理: 现状与挑战. 工程研究-跨学科视野中的工程, 2014, (3): 259–265.

2 Turner BL, Skole D, Sanderson S, *et al*. Land-use and land-cover change: Science/research plan. Global Change Report, 1995, 43(1995): 669–679.

3 Chi M, Sun Z, Qin Y, *et al*. A novel methodology to label urban remote sensing images based on location-based social media photos. Proceedings of the IEEE, 2017, 105(10): 1926–1936. [doi: 10.1109/JPROC.2017.2730585]

4 Lowe DG. Object recognition from local scale-invariant features. Proceedings of the Seventh IEEE International Conference on Computer Vision. IEEE. 1999. 1150–1157. [doi: 10.1109/ICCV.1999.790410]

5 Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features. Computer Vision–ECCV 2006. Gray, Austria. 2006. 404–417.

6 Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA. 2005. 886–893. [doi: 10.1109/CVPR.2005.177]

7 Jegou H, Perronnin F, Douze M, *et al*. Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(9): 1704–1716. [doi: 10.1109/TPAMI.2011.235]

8 Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA. 2007. 1–8. [doi: 10.1109/CVPR.2007.383266]

9 Arandjelovic R, Zisserman A. All about VLAD. Computer Vision and Pattern Recognition, 2013: 1578–1585.

10 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM, 2013, 60(2): 2012.

11 Razavian AS, Azizpour H, Sullivan J, *et al*. CNN features off-the-shelf: An astounding baseline for recognition. Computer Vision and Pattern Recognition, 2014: 512–519.

12 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587. [doi: 10.1109/CVPR.2014.81]

13 Babenko A, Slesarev AV, Chigorin A, *et al*. Neural codes for image retrieval. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8689. Springer. Cham. 2014. 584–599. [doi: 10.1007/978-3-319-10590-1_38]

14 Jegou H, Chum O. Negative evidences and co-occurences in image retrieval: The benefit of PCA and whitening. European Conference on Computer Vision. Lecture Notes on Computer Science, vol 7573. Sringer. Berlin, Heidelberg. 2012. 774–787. [doi: 10.1007/978-3-642-33709-3_55]

15 Gong Y, Wang L, Guo R, *et al*. Multi-scale orderless pooling of deep convolutional activation features. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol 8695. Springer, Cham. 2014. 392–407. [doi: 10.1007/978-3-319-10584-0_26]

16 Ng JY, Yang F, Davis LS, *et al*. Exploiting local features from deep networks for image retrieval. 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Boston, MA, USA. 2015. 53–61. [doi: 10.1109/CVPRW.2015.7301272]

17 Tolias G, Sicre R, Jégou H. Particular object retrieval with integral max-pooling of CNN activations. arXiv preprint arXiv, 2015: 1511. 05879.

18 Babenko A, Lempitsky V. Aggregating local deep features for image retrieval. Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1269–1277. [doi: 10.1109/ICCV.2015.150]

19 Saxena A, Chung SH, Ng AY, *et al*. Learning depth from single monocular images. Neural Information Processing Systems, 2006: 1161–1168.

20 Saxena A, Sun M, Ng A. Make3D: Learning 3D scene structure from a single still image. IEEE TPAMI, 2009, 31: 824–840. [doi: 10.1109/TPAMI.2008.132]

21 Liu B, Gould S, Koller D, *et al*. Single image depth estimation from predicted semantic labels. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA. 2010. 1253–1260. [doi: 10.1109/CVPR.2010.5539823]

22 Karsch K, Liu C, Kang SB, *et al*. Depth extraction from video using non-parametric sampling. In: Fitzgibbon A, Lazebnik S, Perona P, *et al*., eds. Computer Vision – ECCV 2012. Lecture Notes in Computer Science, vol 7576. Springer. Berlin, Heidelberg. 2012. 775–788. [doi: 10.1007/978-3-642-33715-4_56]

23 Konrad J, Wang M, Ishwar P, *et al*. 2D-to-3D image conversion by learning depth from examples. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, RI, USA. 2012. 16–22. [doi: 10.1109/CVPRW.2012.6238903]

24 Eigen D, Puhrsch C, Fergus R, *et al*. Depth map prediction from a single image using a multi-scale deep network. Neural Information Processing Systems, 2014: 2366–2374.

25 Liu F, Shen C, Lin G, *et al*. Deep convolutional neural fields for depth estimation from a single image. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA. 2015. 5162–5170. [doi: 10.1109/CVPR.2015.7299152]

26 Philbin J, Chum O, Isard M, *et al*. Object retrieval with large vocabularies and fast spatial matching. 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA. 2007. 1–8. [doi: 10.1109/CVPR.2007.383172]

27 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv, 2015: 1409.1556.

28 Deng J, Dong W, Socher R, *et al*. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA. 2009. 248–255. [doi: 10.1109/CVPR.2009.5206848]

29 He K, Zhang X, Ren S, *et al*. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1904–1916. [doi: 10.1109/TPAMI.2015.2389824]