

基于 k 密集近邻算法的局部 Fisher 向量编码方法

冀治航^{1,2}, 胡小鹏¹, 杨博¹, 田云云¹, 王凡^{*1}

(1. 大连理工大学 计算机科学与技术学院, 辽宁 大连 116024;

2. 河南科技大学 信息工程学院, 河南 洛阳 471023)

摘要: 在基于视觉词包模型的图像分类方法中, Fisher 向量编码是常用的图像表示方法之一. 该方法利用每一个特征关于所有高斯子模型似然函数的梯度信息来构建图像表达. 而在编码过程中, 每一个特征都会被投影到所有的高斯子模型上并进行编码, 同时子模型之间的内在差异也未被考虑, 这些不足削弱了 Fisher 向量的表达能力. 为此, 提出一种基于 k 密集近邻算法的局部 Fisher 向量编码方法. 在编码过程中该方法引入局部性约束原则, 并利用图像特征空间中高斯子模型间的拓扑结构差异. 在多个数据集上进行测试, 结果表明改进方法能够有效提升分类的准确率.

关键词: 视觉词包模型; 图像分类; Fisher 向量编码; k 密集近邻算法

中图分类号: TP391

文献标识码: A

doi: 10.7511/dllgxb202004010

0 引言

图像分类一直是计算机视觉研究领域的热点问题. 近年来, 许多学者致力于图像表达方法的设计和研究, 它是图像分类系统中的一个重要环节. 迄今为止, 多种不同的图像表达模型被相继提出, 而目前常用的则是视觉词包模型^[1-4]和卷积神经网络^[5-7]. 近年来, 卷积神经网络因其在视觉应用方面的卓越表现而获得了越来越多的关注. 但是, 该类方法性能的提升往往依赖大规模的数据和算力, 且缺乏推理能力, 在一定程度上限制了它的应用和部署^[8-10].

此外, 视觉词包模型也一直是图像分类任务中常用的图像表达方法之一^[1-4], 该模型构建图像表达的过程主要包含以下几个步骤: (1) 在训练数据集上随机提取一组图像特征, 并利用 k -means 或高斯混合模型 (Gaussian mixture model, GMM) 算法构建视觉字典; (2) 对图像中的局部特征进行编码并汇集, 从而形成图像的表达向量. 为了获取具有区分能力的图像表达向量, 学者们相继提出了稀疏编码 (sparse coding, SC)^[11]、局

部约束线性编码 (locality-constrained linear coding, LLC)^[12]、局部聚合描述符向量 (vector of locally aggregated descriptors, VLAD)^[13] 以及 Fisher 向量 (Fisher vector, FV) 编码^[14] 等方法. 其中, Fisher 向量编码以其优异的性能, 得到了很多关注.

然而, 在构建 Fisher 向量的过程中, Fisher 向量编码方法要计算每一个局部特征描述符相对于所有高斯子模型似然函数的梯度信息, 而忽视了编码的局部性原则^[12,15]. 该方法不仅削弱了图像表达向量的辨别能力, 也增加了计算量. 此外, 视觉字典中相似子模型的存在即近义词现象, 也会给编码结果带来影响^[16].

为了提高 Fisher 向量编码结果的表达能力, Nakayama^[15] 提出了基于局部高斯度量 (local Gaussian metrics, LGM) 的 Fisher 向量编码改进方法. 该方法首先利用 k -means 对特征空间进行划分, 并利用单高斯对每一个聚类区域进行建模, 从而构成视觉字典. 在编码过程中, LGM 会将每一个特征投影到与它邻接的 k 个高斯子模型上, 并分别进行 Fisher 向量编码, 从而形成具有局部

收稿日期: 2020-01-20; 修回日期: 2020-05-19.

基金项目: 国家重大专项资助项目 (2018YFA0704605); “十三五”重大专项资助项目 (2017ZX05064).

作者简介: 冀治航 (1979-), 男, 博士生, E-mail: jizhihang2000@163.com; 王凡* (1975-), 女, 博士, 副教授, E-mail: wangfan@dlut.edu.cn.

特点的 Fisher 向量图像表达. 此外, Garg 等^[17]将稀疏原则与 Fisher 向量编码相结合, 提出了稀疏判别 Fisher 向量 (sparse discriminative Fisher vectors, SDFV) 编码. Lu 等^[18]同时引入稀疏以及局部性原则, 并将它们与 Fisher 向量编码相结合, 提出了稀疏 Fisher 向量 (sparse Fisher vector, SFV) 编码.

近年来, 许多学者将卷积神经网络与 Fisher 向量编码方法相结合, 以此构建图像表达. Cimpoi 等^[19]将 CNN 预训练模型的最后一个卷积层输出作为图像特征, 并利用 Fisher 向量编码的方式来构建图像表达. Sydorov 等^[20]借鉴 CNN 多层构架思想, 将 Fisher 向量编码与 SVM 分类器的训练过程结合在一起, 优化了 GMM 字典模型的生成过程, 提升了分类效果. 与原始 Fisher 向量编码方法相比, 上述两个方法分别采用了不同的特征, 优化了字典生成过程, 但其并未考虑 Fisher 向量编码方法中存在的问题. 为了区分图像中不同特征之间的重要性, Pan 等^[21]通过学习方式对不同 GMM 视觉单词赋予了不同的权重, 以此提高了 Fisher 表达向量的辨别能力. 此外, 为了使得 Fisher 向量编码方法能够适应高维 CNN 特征, Liu 等^[22]将稀疏表达与 Fisher 向量编码相结合, 提出了稀疏 Fisher 向量编码. 文献[21-22]所提方法是否对传统手工特征也有效, 作者并未探讨. Tang 等^[23]将 Fisher 向量编码过程进行分解, 并融合到卷积神经网络中的多个层中, 从而将特征学习与 Fisher 向量编码的相关参数统一在一起进行端到端的学习, 以此获取图像表达. 然而该方法需要依赖大规模的数据和算力.

上述这些方法从不同角度对 Fisher 向量编码方法进行了改进, 在一定程度上提升了 Fisher 向量的表达能力, 但并未考虑高斯子模型间拓扑结构的差异, 近义词现象并没有得到解决. 为此, 本文提出一种改进的 Fisher 向量编码方法, 即基于 k 密集近邻算法的局部 Fisher 向量 (Fisher vector based on k -dense neighborhood algorithm, KDN-FV) 编码. 在编码阶段, KDN-FV 编码方法会依据高斯子模型间拓扑结构的差异以及编码局部性原则, 利用 k 密集近邻算法^[24]为每一个特征描述符筛选编码时所需的高斯子空间. KDN-FV 编码方法引入了局部性约束, 并且考虑了子空间之间的拓扑结构差异.

1 Fisher 向量编码方法

Fisher 向量编码方法的设计原理源自 Fisher 核理论, 该理论能够将生成式模型与判别式模型紧密地结合起来. 简单地说, Fisher 核能够利用一个生成模型的参数信息来定义一个核函数, 进而可以构造一个分类器^[25].

假设 $\mathbf{x} = \{x_i; i=1, \dots, t\}$ 是一组观测样本, p_{λ} 是由 $\mathbf{x}$ 中样本所构造的生成模型对应的概率密度函数, $\boldsymbol{\lambda} = (\lambda_1 \ \lambda_2 \ \dots \ \lambda_n) \in \mathbf{R}^n$, 是函数 p_{λ} 的 n 维参数向量, 则样本 $\mathbf{x}$ 关于 p_{λ} 似然函数的梯度可定义为

$$\mathbf{g}_{\lambda}^{\mathbf{x}} = \nabla_{\lambda} \log p_{\lambda}(\mathbf{x}) \quad (1)$$

该梯度反映了似然函数关于每一个参数的最快变化方向, 指出了如何调整参数 $\boldsymbol{\lambda}$ 才能使得模型 p_{λ} 更好地拟合样本数据 $\mathbf{x}$. 值得注意的是, 这里 $\mathbf{g}_{\lambda}^{\mathbf{x}}$ 的维度为 n , 即它仅与 $\boldsymbol{\lambda}$ 中参数的个数有关, 而与样本个数无关.

根据信息几何理论, 概率分布空间 $H = \{p_{\lambda}\}$ 可以形成一个黎曼流形, 它在参数 $\boldsymbol{\lambda}$ 处的局部度量函数可由 Fisher 信息矩阵 $\mathbf{F}_{\lambda} \in \mathbf{R}^{n \times n}$ 给出:

$$\mathbf{F}_{\lambda} = E_{\mathbf{x} \sim p_{\lambda}} (\mathbf{g}_{\lambda}^{\mathbf{x}} \mathbf{g}_{\lambda}^{\mathbf{x}^T}) \quad (2)$$

在 Fisher 信息矩阵给定的情况下, 由样本 $\mathbf{x}$ 和 $\mathbf{y}$ 之间的相似性可以导出 Fisher 核的定义:

$$\Gamma(\mathbf{x}, \mathbf{y}) = \mathbf{g}_{\lambda}^{\mathbf{x}^T} \mathbf{F}_{\lambda}^{-1} \mathbf{g}_{\lambda}^{\mathbf{y}} \quad (3)$$

由于 $\mathbf{F}_{\lambda}$ 是一个半正定矩阵, 可将其进行 Cholesky 分解: $\mathbf{F}_{\lambda}^{-1} = \mathbf{L}_{\lambda}^T \mathbf{L}_{\lambda}$, 那么式(3)就可以重新表示为点积的形式, 即

$$\Gamma(\mathbf{x}, \mathbf{y}) = \boldsymbol{\eta}_{\lambda}^{\mathbf{x}^T} \boldsymbol{\eta}_{\lambda}^{\mathbf{y}} \quad (4)$$

其中

$$\boldsymbol{\eta}_{\lambda}^{\mathbf{x}} = \mathbf{L}_{\lambda} \mathbf{g}_{\lambda}^{\mathbf{x}} = \mathbf{L}_{\lambda} \nabla_{\lambda} \log p_{\lambda}(\mathbf{x}) \quad (5)$$

式(5)定义的梯度向量为样本 $\mathbf{x}$ 的 Fisher 向量.

在文献[14]中, Sánchez 等将 Fisher 向量的思想引入到图像编码中. 此时样本集 $\mathbf{x} = \{x_i; i=1, \dots, t\}$ 将具体地表示从一幅图像中提取到的局部描述符, 例如 SIFT、SURF 等; 而 p_{λ} 表示含有 k 个子模型的混合高斯模型, 其参数向量 $\boldsymbol{\lambda} = \{\omega_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j; j=1, \dots, k\}$, 其中 ω_j 、 $\boldsymbol{\mu}_j$ 和 $\boldsymbol{\Sigma}_j$ 分别表示第 j 个高斯子模型对应的混合权重、均值向量以及协方差矩阵. 根据式(5)分别计算出对每一个参数 ω_j 、 $\boldsymbol{\mu}_j$ 以及 $\boldsymbol{\Sigma}_j$ 的梯度, 并将其串联起来构成图像的 Fisher 表达向量.

2 基于 k 密集近邻算法的局部 Fisher 向量编码方法

从上述 Fisher 向量编码方法的介绍中可以看出,该方法在对局部特征描述符进行编码时,会估计每一个局部特征相对于所有高斯子模型的似然函数,并计算相应梯度信息.该方法忽略了编码的局部性原则,削弱了表达向量的区分能力^[12,15].为此,本文提出一种基于 k 密集近邻算法的局部 Fisher 向量(KDN-FV)编码方法.该方法能够依据特征空间的分布,以及各个子模型间拓扑结构的异同性,仅选择每一个局部特征的邻近且拓扑结构差异较大的子模型进行编码.

2.1 KDN-FV 编码方法的视觉字典构造

与原始 Fisher 向量编码方法相同,KDN-FV 编码方法也采用高斯混合模型方法构建视觉字典.不同的是,改进方法会利用主成分分析法计算每一个高斯子模型中局部特征描述符的主方向向量,并以该向量衡量各个子模型间拓扑结构的差异.具体如图 1 所示.

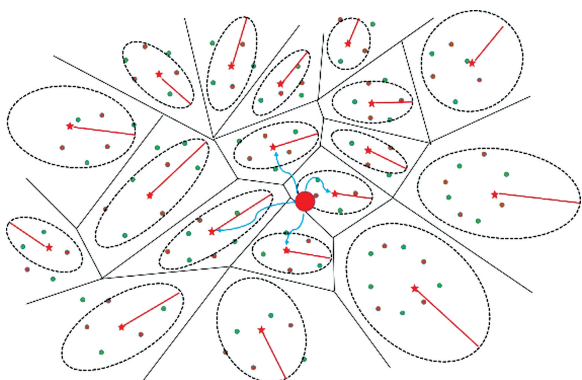


图 1 KDN-FV 编码方法的视觉字典示意图

Fig. 1 Illustration of the visual dictionary for KDN-FV encoding method

2.2 KDN-FV 编码方法的理论基础和方法介绍

假设 $\mathbf{d} = (d_1 \ d_2 \ \cdots \ d_n)$ 是由高斯混合模型方法构成的含有 n 个单词的视觉字典,集合 $\mathbf{p} = \{\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_n\}$ 中包含了每一个单词对应的最大主成分向量.在 KDN-FV 编码方法中,将局部描述符到高斯子模型均值向量之间的欧氏距离称为描述符到子模型的距离.同时,任意两个子模型 d_i 与 d_j 的主方向向量夹角 $\angle \mathbf{p}_i \mathbf{p}_j$ 被用来度量两个子模型的拓扑结构差异性, $\angle \mathbf{p}_i \mathbf{p}_j$ 越大,差

异越大.

定义 1 对于每一个局部特征描述符,将与其最近邻的 m 个高斯子模型称为该描述符的候选子模型(candidate sub-model, CM)组.

定义 2 已知特征描述符 f 及其候选子模型组 $\mathbf{c} = (d_{f1} \ d_{f2} \ \cdots \ d_{fm})$,若从中选取的一个含有 k ($k \leq m$) 个子模型的子集 (k -candidate sub-model, KCM) $\mathbf{c}_k$ 能够使得

$$\text{diff}(\mathbf{c}_k) = \sum_{d_i \in \mathbf{c}_k} \sum_{d_j \in \mathbf{c}_k} F(d_i, d_j);$$

$$|\mathbf{c}_k| = k, \mathbf{c}_k \subseteq \mathbf{c} \quad (6)$$

取得最大值,那么该子集 $\mathbf{c}_k$ 被称为有效子模型(valid sub-model, VM)组,记为 $\mathbf{v}_m$.

这里 $\text{diff}(\cdot)$ 表示 $\mathbf{c}_k$ 中两两子模型之间主方向夹角累加和, $F(\cdot)$ 表示任意两个高斯子模型间主方向向量夹角的度量函数,定义为

$$F(d_i, d_j) = \arccos \frac{\langle \mathbf{p}_i, \mathbf{p}_j \rangle}{\|\mathbf{p}_i\| \cdot \|\mathbf{p}_j\|} \quad (7)$$

其中 $\mathbf{p}_i$ 为 d_i 的主方向向量.

为了克服原始 Fisher 向量编码方法的不足,在编码阶段,KDN-FV 编码方法仅利用每一个描述符邻接的高斯子模型来进行编码,同时需要这些子模型的拓扑结构存在着较大差异.也就是说,KDN-FV 编码方法在进行编码时,首先要确定每一个局部描述符的有效子模型组.为此,本文提出了一种构建有效子模型组的方法.

已知一个特征描述符 f_i 及其所对应的候选子模型组 $\mathbf{c}_i = (d_{i1} \ d_{i2} \ \cdots \ d_{im})$,进而可以得到一个 $m \times m$ 的矩阵 $\mathbf{a}$,其中 $a_{ij} = F(d_i, d_j)$, $1 \leq i, j \leq m$.假设 $\mathbf{y}$ 是一个 m 维指示向量,可用来指出 $\mathbf{c}_i$ 中的子模型是否属于 $\mathbf{v}_{mi}$,即:

$$y_j = \begin{cases} 1; & d_{ij} \in \mathbf{v}_{mi} \\ 0; & d_{ij} \notin \mathbf{v}_{mi} \end{cases}; j = 1, 2, \cdots, m \quad (8)$$

式(6)中的累加和 $\sum_{d_i \in \mathbf{c}_k} \sum_{d_j \in \mathbf{c}_k} F(d_i, d_j)$ 就可等价于 $\mathbf{y}^T \mathbf{a} \mathbf{y}$.如果进一步假设描述符 f_i 的最近邻高斯子模型为 d_{i1} ,且 d_{i1} 一定属于 $\mathbf{v}_{mi}$,那么式(6)的最大值求解问题就可以转化为

$$\arg \max_{\mathbf{y}} \mathbf{y}^T \mathbf{a} \mathbf{y}$$

$$\text{s. t. } \sum_{j=1}^m y_j = k; y_j \in \{0, 1\}, y_1 = 1 \quad (9)$$

式(9)是一个 k 密集近邻求解问题,可由文献[17]中给出的方法求得.

通过求解式(9),可以获得描述符 f_i 的有效子模型组 v_{mi} . 此后,对于图像中的每一个描述符 f_i ,KDN-FV 编码方法仅用其对应 v_{mi} 中的高斯子模型进行 Fisher 向量编码,并将所有描述符的编码结果按照原始 Fisher 向量编码的方式汇集在一起,由此构成图像的 KDN-FV 表示向量.

2.3 KDN-FV 编码方法的主要步骤

输入: 图像 I 的局部特征描述符集合 $f = (f_1$

$f_2 \cdots f_l)$

视觉字典 $d = (d_1 \ d_2 \ \cdots \ d_n)$

候选子模型组参数 m , 有效子模型组参数 k

输出: 图像 I 的 KDN-FV 编码向量 k_{fv}

子模块 1: 计算集合 D 中每一个描述符的 KDN-FV 编码

for $i=1,2,\dots,l$

步骤 1.1 利用 KNN 算法计算描述符 f_i 的候选子模型组 c_i

$index_i \leftarrow \text{KNN}(f_i, d, m)$

$c_i \leftarrow d[index_i]$

步骤 1.2 依据式(7)计算描述符 f_i 所对应的矩阵 a_i ,

for $o=1,2,\dots,m$

for $q=1,2,\dots,m$

$g_o, g_q \leftarrow c_i(o), c_i(q)$

%分别提取 c_i 中的第 o 和第 q 个元素

$a_{ioq} \leftarrow F(d_i, d_j)$

%依据式(7)进行计算

end

end

步骤 1.3 通过求解式(9),计算描述符 f_i 的指示向量 y_i

$y_i \leftarrow \text{KDN}(a_i, m, k)$

步骤 1.4 依据向量 y_i 确定 f_i 的有效子模型组 v_{mi}

$index_i \leftarrow \text{find}(y_i)$ %查找

y_i 中非零元素对应的下标

$v_{mi} \leftarrow c_i[index_i]$

步骤 1.5 根据 v_{mi} 中的每一个子模型,利用式(5)对描述符 f_i 进行 Fisher 向量编码,从而得到 f_i 的 KDN-FV 编码 k_{fvi}

$\alpha_i \leftarrow \text{fit}(f_i, v_{mi})$ %计算描述符 f_i 在有效子模型中各个子空间上分配的权重

$k_{fvi} \leftarrow \text{FV}(f_i, v_{mi}, \alpha_i)$ %对描述符 d_i 进行 Fisher 向量编码

end

子模块 2: 将向量 k_{fvi} ($i=1,2,\dots,l$) 进行累加并求平均,得到图像 I 的 KDN-FV 编码向量 k_{fv}

$k_{fv} \leftarrow \text{sum}(k_{fv1}, k_{fv2}, \dots, k_{fvl}) / l$

2.4 基于 KDN-FV 编码方法的图像分类过程

基于 KDN-FV 编码方法的图像分类过程的关键步骤如下:

(1)在训练数据集中随机选取一组图像,并从每一幅图像中提取局部特征.

(2)使用 GMM 方法将所收集的局部特征进行聚类,构建 GMM 字典.

(3)依次提取数据集中每一幅图像的局部特征,依据已得到的 GMM 字典,采用 2.3 节的 KDN-FV 编码方法对每一幅图像进行编码,从而构建图像表达.

(4)利用训练数据集中图像的表达向量训练 SVM 分类器.

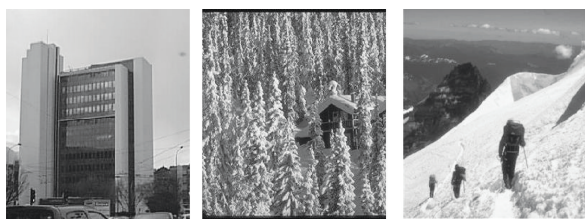
(5)对新输入的图像进行局部特征提取,然后利用 KDN-FV 编码方法进行编码并构建图像的表达向量,最后将图像的表达向量输入到分类器中进行判别,从而完成图像分类.

3 实验结果及分析

为了展示 KDN-FV 编码方法的有效性,分别在 15-Scene^[26]、Pascal VOC 2007^[27]、Caltech 101^[28]、Caltech 256^[29] 以及 MIT Indoor 67^[30] 5 个数据集上对其进行验证. 图 2 展示了相关数据集的图像示例.

3.1 实验参数设置

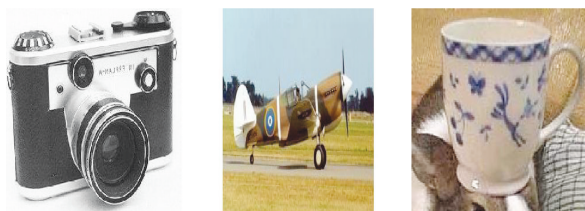
在实验过程中,DenseSIFT 描述符以及 CNN 局部特征被分别用来构建图像表达. 所有方法和实验都利用 VLFeat^[31]、MatConvNet^[32] 工具包来完成. 在 DenseSIFT 特征提取过程中,每隔 4 个像素进行一次采样. 同时,采用空间金字塔匹配 (spatial pyramid matching, SPM)^[33] 方法,



(a) 15-Scene



(b) Pascal VOC 2007



(c) Caltech 101



(d) Caltech 256



(e) MIT Indoor 67

图2 实验数据集图像示例

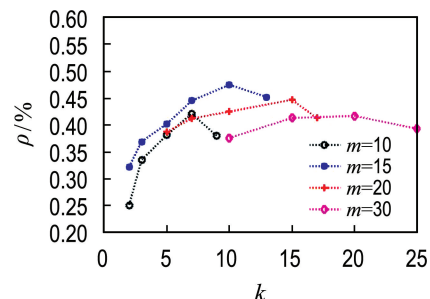
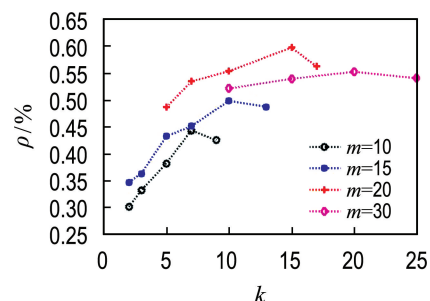
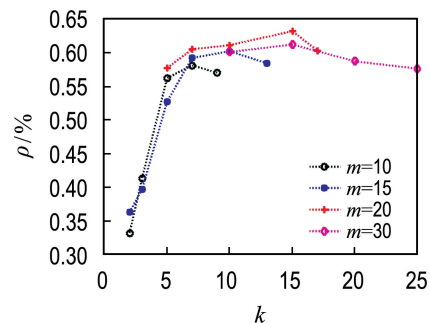
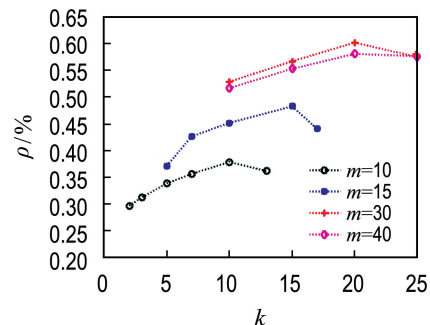
Fig. 2 Sample images taken from the datasets used in the experiments

将图像划分为多个规则的矩形区域, 这里选择 1×1 、 2×2 以及 3×1 的 3 层模式. 在识别阶段, 所有实验都采用 SVM 线性分类器进行分类识别.

3.2 KDN-FV 编码方法关键参数的设置

采用与文献[12, 21-22]相同的策略, 即采用交叉验证的方式来确定 KDN-FV 编码方法中参

数 n 、 m 、 k 的取值. 图 3 所示为在采用 DenseSIFT 局部特征的情况下, n 、 m 、 k 采用不同组合时, KDN-FV 编码方法在 Pascal VOC 2007 数据集上的分类精度. 由图所示, 当 $n=256$, $m=20$, $k=$

(a) $n=64$ (b) $n=128$ (c) $n=256$ (d) $n=512$ 图3 参数 n 、 m 、 k 对分类精度的影响Fig. 3 Influence of the parameter n , m , k on the classification accuracy

15 时,改进方法能够取得较好的结果.

3.3 基于 DenseSIFT 特征的 KDN-FV 图像表达

本节构建基于 DenseSIFT 特征的 KDN-FV 图像表达,与经典 ScSPM^[11]、LLC^[13]和 FV^[14]方法进行比较.此外,为了进一步验证所提方法的有效性,与其他学者提出的局部 Fisher 向量编码方法进行对比,如 LGM^[15]、SDFV^[17]和 SFV^[18].表 1 列出了相应的比较结果.

表 1 基于 DenseSIFT 特征的 KDN-FV 图像表达分类性能

Tab. 1 Classification performance of the KDN-FV image representation based on DenseSIFT feature

方法	分类精度 / %			
	Pascal VOC 2007	Caltech 101	Caltech 256	15-Scene
ScSPM ^[11]	51.9	73.2	37.5	80.3
LLC ^[13]	57.6	73.4	45.3	81.5
FV ^[14]	61.8	77.8	52.1	85.3
LGM ^[15]	—	78.3	—	80.1
SDFV ^[17]	61.9	76.1	—	86.3
SFV ^[18]	58.2	70.8	—	—
KDN-FV	64.6	79.7	55.2	86.9

在 Pascal VOC 2007 数据集上,从每类图像中随机抽取 200 幅进行训练,然后在测试集上进行测试.由表 1 可以看出,KDN-FV 编码方法取得了 64.6% 的分类精度($n=256, m=20, k=15$),比 ScSPM 和 LLC 分别高出 12.7%、7.0%;比原始 Fisher 向量编码方法(FV)高出 2.8%.与具有相似思路的改进方法相比,如 SDFV、SFV, KDN-FV 编码方法依然有明显优势.

在 Caltech 101 数据集上,从每种类别中随机

选取 30 幅图像进行训练,剩余图像用于测试.同时,相关参数依次设置为 $n=256, m=20, k=15$.与其他编码方法相比,KDN-FV 编码方法取得了更高的分类结果,比原始 Fisher 向量编码方法高出 1.9%.

此外,KDN-FV 编码方法在 Caltech 256 和 15-Scene 数据集上也取得了较好的分类结果.它优于原始 Fisher 向量编码方法,并在两个数据集上分别高出 3.1% 和 1.6%.同时,与 SDFV 和 LGM 方法相比,KDN-FV 在 15-Scene 数据集上的分类结果也有一定的提升.

3.4 基于 CNN 局部特征的 KDN-FV 图像表达

将 AlexNet^[5]预训练模型最后一个全连接层的输出作为特征,并与 3 种不同的表达策略进行比较:(1)全局 CNN 特征,即直接使用全连接层的输出作为图像表达.(2)CNN 多尺度无序汇集^[34],即 multi-scale orderless pooling CNN (MOP-CNN).这种方法以多尺度采样的方式在图像中提取 CNN 局部特征,并将这些特征进行 VLAD 或 Fisher 向量编码,形成图像表达.(3)CNN 多尺度金字塔汇集^[35],即 multi-scale pyramid pooling CNN (MPP-CNN).该方法以多尺度采样的方式在图像中提取 CNN 局部特征,并引入金字塔策略,最终将这些特征进行 Fisher 向量编码,形成图像表达.在实验过程中,使用与待比较方法相同的采样策略和设置进行 CNN 局部特征采样,并用 KDN-FV 编码代替对应的编码方法来构建图像表达.其中对于 MOP-CNN (KDN-FV)以及 MPP-CNN (KDN-FV)方法,所采用的参数设置为 $n=512, m=40, k=25$.表 2 列出了相应的比较结果.

表 2 基于 CNN 特征的 KDN-FV 图像表达分类性能

Tab. 2 Classification performance of the KDN-FV image representation based on CNN feature

方法	分类精度 / %			
	Pascal VOC 2007	Caltech 101	Caltech 256	MIT Indoor 67
Global-CNN	72.2	85.1	69.8	53.7
MOP-CNN(VLAD) ^[34]	75.7	87.3	73.1	68.9
MOP-CNN(FV)	76.2	89.1	73.5	70.2
MPP-CNN(FV) ^[35]	79.5	90.7	77.3	75.7
MOP-CNN(KDN-FV)	77.1	92.4	73.9	71.3
MPP-CNN(KDN-FV)	82.1	93.8	78.2	78.6

从表2可以看出,当用KDN-FV编码方法替代比较方法中的编码方法后,新构建的图像表达具有更强的区分能力,这也说明了KDN-FV编码方法同样适用于CNN局部特征。

3.5 结果分析

从上述实验结果可以看出,KDN-FV编码方法既适用于DenseSIFT手工特征,也适用于由CNN网络提取出来的深度特征。文献[15,17-18]指出局部性约束的引入可以提升Fisher向量的区分能力。而KDN-FV在引入局部性约束的同时,又利用子模型拓扑结构的差异对视觉单词进行了筛选。它在Pascal VOC 2007和Caltech 256数据集上分别取得了64.6%和55.2%的分类精度,进一步提升了Fisher向量的表达能力。由此可以看出所提方案的有效性。

3.6 时耗性能分析

与原始Fisher向量编码方法相比,改进的KDN-FV编码方法增加了单词筛选步骤。为了分析KDN-FV编码方法的时间复杂性,采用与文献[14,21]相同的策略,比较了KDN-FV编码方法与原始Fisher向量编码方法的单幅图像平均编码时耗。在实验过程中,从Pascal VOC 2007数据集中选取了1 000幅图像,并将所有图像缩放为 200×200 ,字典中单词个数为128。硬件配置为Intel i5 6 600 KB 3.5 GHz CPU,16 GB内存。经测试,Fisher向量编码方法单幅图像编码时间约为1 530 ms,而KDN-FV则为1 317 ms。实验表明,KDN-FV编码方法的时耗较少,这是因为即使KDN-FV中增加了单词筛选环节,但在编码阶段,KDN-FV将每一个局部特征仅仅投影到所筛选出的视觉单词即可,而原始Fisher向量编码方法需要计算所有视觉单词上的投影。由于单个单词的投影过程计算量较大,而KDN-FV编码方法仅需计算部分投影,这就减少了该方法的编码时间。

4 结 语

针对原始Fisher向量编码方法在编码过程中存在的全局编码以及未考虑视觉单词之间拓扑结构差异的问题,本文提出了一种基于 k 密集近邻算法的局部Fisher向量编码方法。该方法依据图像特征空间的流形结构,利用 k 密集近邻算法为每一个特征描述符筛选编码时所需的高斯子空

间。KDN-FV编码方法在编码过程中引入了局部性约束,并且考虑了子空间之间的异同性,克服了原始Fisher向量编码方法存在的问题。通过在多个数据集上进行测试,结果显示改进后的方法增强了图像表达向量的区分能力,提升了分类精度,提高了编码效率。

参考文献:

- [1] CSURKA G, DANACE C R, FAN Lixin, *et al.* Visual categorization with bags of keypoints [C] // **Proceedings of the 8th European Conference on Computer Vision**. Prague: Springer, 2004.
- [2] 张立和,潘磊,刘涛,等. 基于核拉普拉斯稀疏编码的图像分类[J]. 大连理工大学学报, 2015, **55**(2): 192-197.
ZHANG Lihe, PAN Lei, LIU Tao, *et al.* Kernel Laplacian sparse coding for image classification [J]. **Journal of Dalian University of Technology**, 2015, **55**(2): 192-197. (in Chinese)
- [3] LIU Li, CHEN Jie, FIEGUTH P, *et al.* From BoW to CNN: Two decades of texture representation for texture classification [J]. **International Journal of Computer Vision**, 2019, **127**(1): 74-109.
- [4] ZHENG Liang, YANG Yi, TIAN Qi. SIFT meets CNN: A decade survey of instance retrieval [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2018, **40**(5): 1224-1244.
- [5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [J]. **Advances in Neural Information Processing Systems**, 2012, **2**: 1097-1105.
- [6] 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.
- [7] 吕飞,韩敏. 基于深度极限学习机的高光谱遥感影像分类研究[J]. 大连理工大学学报, 2018, **58**(2): 166-173.
LV Fei, HAN Min. Hyperspectral remote sensing image classification based on deep extreme learning machine [J]. **Journal of Dalian University of Technology**, 2018, **58**(2): 166-173. (in Chinese)
- [8] ARANDJELOVIĆ R, GRONAT P, TORII A, *et al.* NetVLAD: CNN architecture for weakly

- supervised place recognition [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2018, **40**(6): 1437-1451.
- [9] BRENDEN W, BETHGE M. Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet [C] // **7th International Conference on Learning Representations, ICLR 2019**. New Orleans: International Conference on Learning Representations, 2019.
- [10] PASSALIS N, TEFAS A. Training lightweight deep convolutional neural networks using bag-of-features pooling [J]. **IEEE Transactions on Neural Networks and Learning Systems**, 2019, **30**(6): 1705-1715.
- [11] YANG Jianchao, YU Kai, GONG Yihong, *et al.* Linear spatial pyramid matching using sparse coding for image classification [C] // **2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009**. Miami: IEEE Computer Society, 2009: 5206757.
- [12] WANG Jinjun, YANG Jianchao, YU Kai, *et al.* Locality-constrained linear coding for image classification [C] // **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**. San Francisco: IEEE Computer Society, 2010: 5540018.
- [13] JÉGOU H, PERRONNIN F, DOUZE M, *et al.* Aggregating local image descriptors into compact codes [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2011, **34**(9): 1704-1716.
- [14] SÁNCHEZ J, PERRONNIN F, MENSINK T, *et al.* Image classification with the Fisher vector: Theory and practice [J]. **International Journal of Computer Vision**, 2013, **105**(3): 222-245.
- [15] NAKAYAMA H. Aggregating descriptors with local Gaussian metrics [C] // **Proceedings of NIPS 2012 Workshop on Large Scale Visual Recognition and Retrieval**. Cambridge: MIT Press, 2012.
- [16] VAN GEMERT J C, VEENMAN C J, SMEULDERS A W M, *et al.* Visual word ambiguity [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2010, **32**(7): 1271-1283.
- [17] GARG V, CHANDRA S, JAWAHAR C V. Sparse discriminative Fisher vectors in visual classification [C] // **Proceedings - 8th Indian Conference on Computer Vision, Graphics and Image Processing, ICVGIP 2012**. Mumbai: Association for Computing Machinery, 2012.
- [18] LU Xiankai, FANG Zheng, XU Tao, *et al.* Efficient image categorization with sparse Fisher vector [C] // **2015 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015 - Proceedings**. Brisbane: IEEE, 2015: 7178220.
- [19] CIMPOI M, MAJI S, VEDALDI A. Deep filter banks for texture recognition and segmentation [C] // **IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015**. Boston: IEEE Computer Society, 2015: 7299007.
- [20] SYDOROV V, SAKURADA M, LAMPERT C H. Deep Fisher kernels-end to end learning of the Fisher kernel GMM parameters [J]. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, 2014(1 of 1): 1402-1409.
- [21] PAN Yongsheng, XIA Yong, SHEN Dinggang. Foreground Fisher vector: Encoding class-relevant foreground to improve image classification [J]. **IEEE Transactions on Image Processing**, 2019, **28**(10): 4716-4729.
- [22] LIU Lingqiao, WANG Peng, SHEN Chunhua, *et al.* Compositional model based Fisher vector coding for image classification [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2017, **39**(12): 2335-2348.
- [23] TANG Peng, WANG Xinggang, SHI Baoguang, *et al.* Deep FisherNet for image classification [J]. **IEEE Transactions on Neural Networks and Learning Systems**, 2019, **30**(7): 2244-2250.
- [24] LIU Hairong, YANG Xingwei, LATECKI L J, *et al.* Dense neighborhoods on affinity graph [J]. **International Journal of Computer Vision**, 2012, **98**(1): 65-82.
- [25] JAAKKOLA T S, HAUSSLER D. Exploiting generative models in discriminative classifiers [J]. **Advances in Neural Information Processing Systems**, 1999: 487-493.
- [26] OLIVA A, TORRALBA A. Modeling the shape of the scene: A holistic representation of the spatial envelope [J]. **International Journal of Computer Vision**, 2001, **42**(3): 145-175.
- [27] EVERINGHAM M, VAN L, WILLIAMS C K I,

- et al.* The PASCAL visual object classes challenge 2007 (VOC 2007) results. (2007-04-07) [2020-01-20]. <http://host.robots.ox.ac.uk/pascal/VOC/voc2007/>.
- [28] FEI-FEI L, FERGUS R, PERONA P. One-shot learning of object categories [J]. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, 2006, 28(4): 594-611.
- [29] GRIFFIN G, HOLUB A, PERONA P. The Caltech-256. (2006-11-15) [2020-01-20]. http://www.vision.caltech.edu/Image_Datasets/Caltech256/.
- [30] QUATTONI A, TORRALBA A. Recognizing indoor scenes [C] // **2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009**. Miami: IEEE Computer Society, 2009: 5206537.
- [31] VEDALDI A, FULKERSON B. VLFeat-An open and portable library of computer vision algorithms [C] // **MM'10 - Proceedings of the ACM Multimedia 2010 International Conference**. New York: Association for Computing Machinery, 2010: 1469-1472.
- [32] VEDALDI A, LENC K. MatConvNet: Convolutional neural networks for MATLAB [C] // **MM 2015 - Proceedings of the 2015 ACM Multimedia Conference**. Brisbane: Association for Computing Machinery, 2015: 689-692.
- [33] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C] // **Proceedings - 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006**. New York: IEEE Computer Society, 2006: 1641019.
- [34] GONG Yunchao, WANG Liwei, GUO Ruiqi, *et al.* Multi-scale orderless pooling of deep convolutional activation features [C] // **Computer Vision, ECCV 2014 - 13th European Conference, Proceedings**. Zurich: Springer Verlag, 2014: 392-407.
- [35] YOO D, PARK S, LEE J Y, *et al.* Multi-scale pyramid pooling for deep convolutional representation [C] // **2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2015**. Boston: IEEE Computer Society, 2015: 7301274.

Local Fisher vector encoding method based on k -dense neighborhood algorithm

JI Zhihang^{1,2}, HU Xiaopeng¹, YANG Bo¹, TIAN Yunyun¹, WANG Fan^{*1}

(1. School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China;

2. School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China)

Abstract: For the image classification methods based on the bag-of-visual-words model, Fisher vector (FV) encoding is one of the popular image representation approaches. In this method, gradient information of the likelihood functions, which is achieved by fitting each feature with all Gaussian sub-models, is used to build image representation. However, in this encoding procedure, each feature is mapped to all of the Gaussian sub-models and encoded by them, and the inherent differences between these sub-models have not been considered. These drawbacks limit the representative ability of the Fisher vector. To solve these problems, a local Fisher vector encoding approach based on k -dense neighborhood(KDN) algorithm is proposed, which introduces the local constraint and utilizes the difference between the topological structures of the Gaussian sub-models. Experiments are conducted on several benchmark datasets, and the results demonstrate the effectiveness of the proposed method in improving the accuracy of the classification.

Key words: bag-of-visual-words model; image classification; Fisher vector encoding; k -dense neighborhood algorithm