DOI:10.11918/j.issn.0367-6234.201803047

基于深度特征聚类的海量人脸图像检索

李振东[1,2], 钟 勇[1,2], 张博言[1,2], 曹冬平[1,2]

(1.中国科学院 成都计算机应用研究所, 成都 610041;2. 中国科学院大学,北京 100049)

摘 要: 针对海量人脸图像数据库检索时长的问题,提出了一种基于深度特征聚类的海量人脸图像检索算法. 该算法首先使用人脸图像训练集对深度卷积神经网络模型进行人脸图像分类训练,在此基础上采用三元组损失方法对已训练好的人脸图像分类网络模型进行微调,使得网络能够更加有效地提取人脸图像的高层语义特征,构建更具有表征性的人脸图像深度特征. 其次采用 K-means 聚类算法对提取的人脸图像深度特征进行聚类,使得同一个人的人脸图像能够划分到同一簇中,然后在相应的簇中进行人脸图像的深度特征相似度匹配执行人脸图像检索任务. 为了进一步提高系统的检索性能,提出人脸图像深度特征融合的查询扩展方法,对待检索的人脸图像深度特征进行融合再次执行检索任务得到最终的检索结果. 通过在两个人脸检索数据集(Celebrities Face Set 和 Labeled Faces in the Wild dataset)上进行详尽实验验证,结果表明,该算法能极大地缩小海量人脸图像数据库的检索范围,在保证一定准确率的前提下有效地提高了人脸图像检索的速度.

关键词: 图像检索;卷积神经网络;特征提取;K-means 聚类

中图分类号: TP391.41 **文献标志码**: A **文章编号**: 0367-6234(2018)11-0101-09

Massive face image retrieval based on depth feature clustering

LI Zhendong[1,2], ZHONG Yong[1,2], ZHANG Boyan[1,2], CAO Dongping[1,2]

(1. Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu, 610041, China;
2. University of Chinese Academy of Science, Beijing 100049, China)

Abstract: A massive face image retrieval method based on depth feature clustering is proposed to overcome the long time retrieving in a huge scale face image database. Firstly, the convolutional neural network model is trained for face classification by face image training set. Based on it, the triplet loss method is used to fine-tune the model so that the network can be more efficient to extract high-level semantic features and construct a more representative depth features of face image. Secondly, the K-means clustering algorithm is used to cluster the extracted depth features, so that face images of the same person can be divided into the same cluster, and then similarity matching of face images is performed in the corresponding clusters to perform the retrieval task. In order to further improve the performance of system retrieval, the face image feature fusion query expansion method is proposed to fuse the depth features of the face image to be retrieved. Through exhaustive experimental verification on two face retrieval datasets (Celebrities Face Set and Labeled Faces in the Wild dataset), the results show that the proposed method can significantly reduce the retrieval range of massive face image database, thus effectively increasing the face image retrieval speed while ensuring similar retrieval accuracy.

Keywords: image retrieval; convolutional neural network; feature extraction; K-means clustering

人脸图像检索是计算机视觉领域中的一个基本研究问题. 最初,图像是用关键字和基于文本的检索系统进行人工注释的. 但是,随着待检索对象数量的不断增加形成了海量人脸图像数据库,人工手动注释的检索系统变得不可行. 因此,提高海量人脸图像数据库的检索速度和准确率是大型人脸图像数据库检索必须解决的问题.

Liu 等提出建立图像库的层次索引结构[1-2],即建立合适的层次特征索引结构,将整个数据库分成多个类,检索时只在一个或少数几个类内进行,从而提高检索速度. 刘小华等提出基于 L-K 均值层次聚类算法[3],把大型人脸图像数据库划分成一些子类数据集,只在一个或几个子类中进行检索,但是该算法会随着类内数据量和检索节点的增加导致检索时间显著增加. 杨之光等提出基于聚类的家庭数码相册人脸图像检索算法[4],利用归一化分割在每个时间段内分别对人脸图像进行聚类,并采用连续 AdaBoost 算法得到人脸识别分类器度量人脸之间的相似度,该算法在小的家庭相册人脸检索准确率和时效方面得到了一定的提升,但不能用于海量人脸图像检索的实用性和可靠性.

收稿日期: 2018-03-08
基金项目: 四川省科技厅重点研发项目(2017SZ0010); 四川省科技支撑计划项目(2016JZ0035)
作者简介: 李振东(1990—),男,博士研究生;
钟 勇(1966—),男,研究员,博士生导师
通信作者: 钟 勇, zhongyong@casit.com.cn

近年来,深度卷积神经网络(Convolutional Neural Network, CNN)在图像分类[5-6]、人脸检测[7-8]以及人脸识别[9-10]等计算机视觉任务中实现了最优的性能. 深度卷积神经网络经过大量数据训练后的用于学习特征表示的模型可泛化使用,Yandex 等[11-14]采用了基于图像分类任务预先训练的 CNN 模型提取深度卷积特征进行图像检索的解决方案,实现了在流行的检索基准上达到最好的性能. 此外,卷积层的特征对应于接收图像特定区域特征的自然解释,被认为是类比于图像浅的纹理特征如 HOG[15]、LBP[16]等特征. HOG 特征将图像分成小的连通区域,称作细胞单元,然后采集细胞单元中各像素点的梯度或边缘方向直方图,把这些直方图组合起来构成特征描述符. 但是 HOG 特征很难处理遮挡、物体形变及方向改变等问题,由于梯度的性质,HOG 特征对噪点相当敏感,而且由于描述子生成过程冗长导致速度慢,实时性差. 与 HOG 特征相比,LBP 是一个简单但非常有效的纹理运算符,它将各个像素与其附近的像素进行比较,并把结果保存为二进制数. LBP 最重要的属性是对诸如光照变化等造成的灰度变化的鲁棒性,并且计算简单,对图像进行实时分析. 但 LBP 对方向信息敏感,容易导致纹理信息丢失.

针对以上问题,提出一种基于深度特征聚类的海量人脸图像快速检索方法. 该方法首先通过深度卷积神经网络提取人脸图像的深度卷积特征,构建人脸图像检索的特征向量集,并利用 K-means 聚类算法[17]对人脸特征向量集进行聚类,使得同一人的人脸图像能够正确聚为一类;然后提取待检索人脸图像的深度特征构建特征向量,将构建的待检索人脸特征向量与距离类中心最近的类内人脸图像特征进行相似度度量匹配得到初始的检索结果列表;在此基础上,采用特征融合的查询扩展方法进一步提高系统的检索性能.

1 人脸图像检索模型设计

设计深度卷积神经网络并使用人脸图像数据集对网络模型进行人脸分类训练,使得网络能够有效地提取人脸图像的深度特征执行人脸图像检索任务,算法模型结构见图 1. 主要分为两个部分,第一部分为用于提取人脸图像深度特征的深度卷积神经网络,含有 13 层卷积层和 3 层全连接层的深度卷积神经网络结构;第二部分为将数据库中保存的人脸图像特征向量进行 K-means 聚类使得人脸图像划分为不同的类 $\{C_1, C_2, \cdots, C_k\}$ 及对应的聚类中心 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k\}$. 对于给定一张待检索的人脸图像 I',该算法执行检索图像 I' 的流程见图 2.

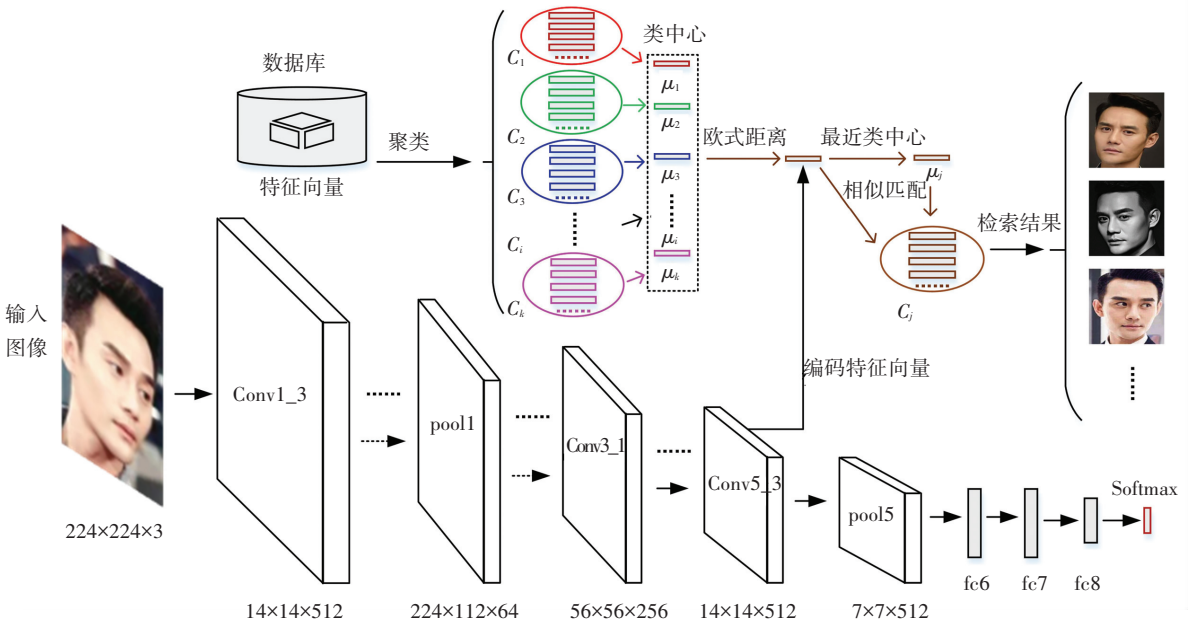


图 1 算法模型结构

Fig.1 Algorithm model structure

1.1 网络模型人脸分类训练

考虑深度神经网络 f 识别 N 个不同人的人脸问题,将网络输出设置为 N 路分类问题. 深度卷积神经网络通过包含 N 个线性预测器(每个标识一个类别)的全连接层(图 1 中 fc8 层)将每个训练图像 I_t, $t = 1, \cdots, T$ 与分数向量 $\boldsymbol{s}_t = \boldsymbol{W}\boldsymbol{f}(I_t) + \boldsymbol{b}$ 相关联,其中 $\boldsymbol{W} \in \mathbf{R}^{N \times D}$ 为全连接层 fc8 层的权重,$\boldsymbol{b} \in \mathbf{R}^N$ 为网络偏置项. 通过计算 softmax 对数损失 $L(f)$,将这些得分与真实标注类别 $c_t \in \{1, \cdots, N\}$ 进行比较,softmax 对数损失 $L(f)$ 计算如下:

$$L(f) = -\sum_t (\log e^{\langle c_t, x_t\rangle} - \log(\sum_{q=1,\cdots,N} e^{\langle c_q, x_t\rangle})) . \quad (1)$$

式中 $\boldsymbol{x}_t = \boldsymbol{f}(I_t) \in \mathbf{R}^D$ 表示类别分数向量.

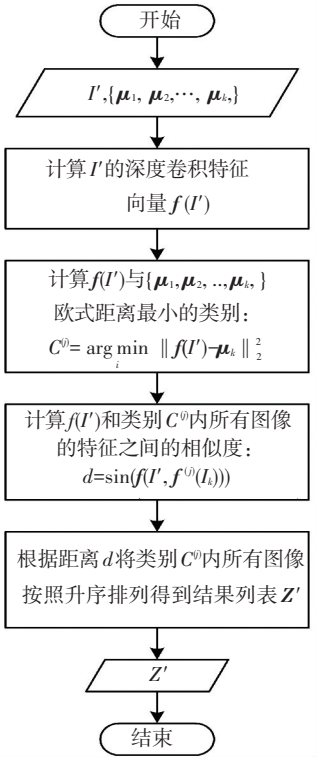


图 2　人脸图像检索流程

Fig.2　Face image retrieval process

网络模型训练学习之后，使用欧式距离比较分数向量 $\boldsymbol{x}_t = \boldsymbol{f}(I_t)$ 进行人脸验证. 然而，通过使用三元组损失方案微调网络模型在欧式空间中进行验证，可显著提高得分. 虽然后者可以获得良好的整体性能，但是本节将网络首先作为分类器进行预训练会使得后续微调训练变得更加容易和快速.

1.2　三元组损失微调网络

三元组损失微调网络模型学习能够在最终应用中表现良好的分数向量，即通过在欧式空间中比较人脸特征描述符进行身份验证. 与许多度量学习方法相似，三元组损失用于学习具有独特性和紧凑性的投影，同时实现降维.

三元组损失微调方法与文献[9-10]相似，一个三元组(a, p, n) 包含一个人脸图像的训练样本 a 以及一个与 a 相同类别的正样本 p 和一个与 a 不同类别的负样本 n 的三元组. 预先训练的 CNN 的输出 $\boldsymbol{f}(I_t) \in \mathbf{R}^D$ 经过 l_2 正则化，并使用仿射投影

$$\boldsymbol{x}_t = \boldsymbol{W}' \frac{\boldsymbol{f}(I_t)}{\| \boldsymbol{f}(I_t) \|_2}, \ \boldsymbol{W}' \in \mathbf{R}^{L\times D}. \quad (2)$$

将 $\boldsymbol{f}(I_t)$ 投影到更低维 $L(L < D)$ 维空间. 与线性预测器不同的是，L 不等于类别的数量，但它是描述符嵌入的大小(实验中设置 L = 1 024). 投影 $\boldsymbol{W}'$ 被训练最小化三元组损失：

$$L(\boldsymbol{W}') = \sum_{(a,p,n)\in T} \max\{0, \alpha + \| \boldsymbol{x}_a - \boldsymbol{x}_p \|_2^2 - \| \boldsymbol{x}_a - \boldsymbol{x}_n \|_2^2\}, \quad (3)$$

其中 $\boldsymbol{x}_i$ 的定义如式(2)，$\alpha \geqslant 0$ 是表示学习边界的固定标量，T 是训练三元组的集合. 三元组损失的目的就是通过学习，让训练样本 a 和 p 特征表达之间的距离尽可能小，而 a 和 n 的特征表达之间的距离尽可能大，并且要让 a 和 n 之间的距离和 a 和 p 之间的距离之间有一个最小的间隔，即 α，公式化的表示为

$$\alpha + \| \boldsymbol{x}_a - \boldsymbol{x}_p \|_2^2 < \| \boldsymbol{x}_a - \boldsymbol{x}_n \|_2^2. \quad (4)$$

当 a 和 n 之间的距离小于 a 和 p 之间的距离加 α 时，就会产生损失；当 a 和 n 之间的距离大于等于 a 和 p 之间的距离加 α 时，损失为零，如式(3).

2　人脸图像检索方法

2.1　人脸图像特征提取

对于给定像素大小为 $W_I \times H_I$ 的输入图像 I 的卷积层激活响应值是一个三维张量 $W \times H \times K$，K 表示输出特征通道数，即多维滤波器. 空间像素 $W \times H$ 是由网络结构和输入图像像素大小决定的. 图 1 所示网络结构中，每个卷积层都利用前一层的输出作为本层的输入，定义为

$$\boldsymbol{h}_{W,b}^{m(l)}(x) = f(b^{m(l)} + \sum_n \boldsymbol{W}^{mn(l)} * \boldsymbol{h}_{W,b}^{n(l-1)}) . \quad (5)$$

式中：$\boldsymbol{h}^{m(l)}$ 和 $\boldsymbol{h}^{n(l-1)}$ 分别为 l 层的第 m 个输出通道和 $l-1$ 层的第 n 个输入通道，$\boldsymbol{W}^{mn(l)}$ 和 $b^{m(l)}$ 为相应的卷积核滤波器和偏置项，$*$ 为卷积运算符.

由于从卷积层直接提取的图像特征存在特征数据量大、数据维度高导致结果容易出现过拟合及计算量大等问题，因此在相应的卷积层后采用最大特征值池化(max-pooling)方法对卷积层的不同位置进行特征聚合统计：

$$\boldsymbol{h}_{(i,j)}^{m(l)} = \max_{\forall (p,q)\in\Omega_{(i,j)}} \{\boldsymbol{h}_{(p,q)}^{m(l)}\} . \quad (6)$$

式中：$\Omega_{(i,j)}$ 为索引为 (i, j) 特定区域，(p, q) 为在 Ω 上具体位置的索引.

在网络的最后一层卷积层(Conv5_3)后同样执行最大特征值池化操作(pool5)得到尺寸大小为 7×7×512 维的人脸图像特征响应值 $\boldsymbol{f}(I) = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_i, \cdots, \boldsymbol{X}_{512}]$，其中 $\boldsymbol{X}_i$ 表示尺寸大小为 14×14 的特征激活响应值，512 表示输出特征通道数，即多维滤波器. 然后，对人脸图像 I 的每个维度的特征激活响应值进行总和池化求平方根(Sum Pooling Followed By Square Root, SPSR)的特征聚合操作，构建人脸图像 I 的 SPSR 深度卷积特征向量：

$$\boldsymbol{f}_{\text{sum}}(I) = [f_1, f_2, \cdots, f_i, \cdots, f_{512}]. \tag{7}$$

$$f_i = \sqrt{\sum \boldsymbol{X}_i} = \sqrt{\sum_{y=1}^{14}\sum_{x=1}^{14} f_i(x, y)}. \tag{8}$$

式中 $f_i(x, y)$ 为第 i 个特征通道激活响应值的空间位置坐标 (x, y). 最后将 $\boldsymbol{f}_{\text{sum}}(I)$ 进行 l_2 正则化:

$$\boldsymbol{f}(I) = \frac{\boldsymbol{f}_{\text{sum}}(I)}{\| \boldsymbol{f}_{\text{sum}}(I) \|_2}. \tag{9}$$

$\boldsymbol{f}(I)$ 即为人脸图像 I 的 SPSR 深度卷积特征向量用于人脸图像检索任务.

2.2 人脸图像深度特征聚类

从网络的最后一层卷积层 Con5_3 层提取待检索人脸图像 I' 的深度卷积特征向量 $\boldsymbol{f}(I')$ 和数据库中保存的人脸图像集 $\boldsymbol{Z} = \{I_1, \cdots, I_i, \cdots, I_n\}$ 所对应的深度特征向量 $\{\boldsymbol{f}(I_1), \cdots, \boldsymbol{f}(I_i), \cdots, \boldsymbol{f}(I_n)\}$, $\boldsymbol{f}(I_i) \in \mathbf{R}^{512}$, 其中 i 为数据库中第 i 张人脸图像, n 的大小代表了人脸数据库中数据集的大小.

给定人脸图像数据集对应的深度特征向量,利用 K-means 聚类算法通过迭代寻找 k 个聚类的划分方案,针对聚类所得簇划分 $C = \{C_1, C_2, \cdots, C_k\}$ 最小化误差平方和(各个类畸变程度之和)

$$E = \sum_{i=1}^{k}\sum_{\boldsymbol{f}(I)\in C_i} \| \boldsymbol{f}(I) - \boldsymbol{\mu}_i \|_2^2. \tag{10}$$

式中 $\boldsymbol{\mu}_i = \dfrac{1}{|C_i|}\sum_{\boldsymbol{f}(I)\in C_i} \boldsymbol{f}(I)$ 是簇 C_i 的均值向量.

实验中采用 Elbow Method[18] 估计最优聚类数量 k 值. 当 k 小于真实聚类数时,由于 k 的增大会大幅增加每个簇的聚合程度,故 E 的下降幅度会很大,而当 k 到达真实聚类数时,再增加 k 所得到的聚合程度回报会迅速变小,所以 E 的下降幅度会骤减,然后随着 k 值的继续增大而趋于平缓,也就是说 E 和 k 的关系图是一个手肘的形状,而这个肘部对应的 k 值就是数据的最佳聚类数. 当最优聚类数量 k 值确定后具体聚类实现过程如下:

1) 随机选取 k 个深度特征向量作为初始聚类中心,分别为 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k \in \mathbf{R}^{512}$

2) 计算每一个样本 i 的所属类别 $\boldsymbol{C}^{(i)}$

$$\boldsymbol{C}^{(i)} = \underset{j}{\text{argmin}} \| \boldsymbol{f}(I_i) - \boldsymbol{\mu}_j \|_2^2. \tag{11}$$

即样本到类别中心欧式距离最小的类别;

3) 更新每一类的中心 $\boldsymbol{\mu}_j$

$$\boldsymbol{\mu}_j = \frac{\sum_{i=1}^{m} \boldsymbol{f}(I_i)\,|_{C^{(i)}=j}}{\sum_{i=1}^{m} 1\,|_{C^{(i)}=j}}. \tag{12}$$

4) 重复步骤 2)、3) 至式(10)所示误差 E 收敛.

通过上述方法将人脸图像集对应的特征向量划分为不同的簇 $C = \{C_1, C_2, \cdots, C_k\}$, 对应的聚类中心为 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k\}$.

2.3 人脸图像类内检索

得到人脸图像集对应的特征向量簇划分 $C = \{C_1, C_2, \cdots, C_k\}$ 和对应的聚类中心 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k\}$ 后,计算待检索人脸图像 I' 的深度特征向量 $\boldsymbol{f}(I')$ 和每个聚类中心 $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_k\}$ 之间的欧式距离,寻找欧式距离最小的类别:

$$C^{(j)} = \underset{i}{\text{argmin}} \| \boldsymbol{f}(I') - \boldsymbol{\mu}_i \|_2^2. \tag{13}$$

得到距离最近的类 $C^{(j)}$ 后,计算待检索图像 I' 的深度特征向量 $\boldsymbol{f}(I')$ 和类 $C^{(j)}$ 内所有的人脸图像 $\boldsymbol{Z}' = \{I_1^{(j)}, \cdots, I_k^{(j)}, \cdots, I_m^{(j)}\}$ 的深度特征向量 $\{\boldsymbol{f}^{(j)}(I_1), \cdots, \boldsymbol{f}^{(j)}(I_k), \cdots, \boldsymbol{f}^{(j)}(I_m)\}$ $(m < n)$ 之间的余弦相似度:

$$d = \text{sim}(\boldsymbol{f}(I'), \boldsymbol{f}^{(j)}(I_k)) = \frac{\sum \boldsymbol{f}(I') \times \boldsymbol{f}^{(j)}(I_k)}{\sqrt{\sum (\boldsymbol{f}(I'))^2} \times \sqrt{\sum (\boldsymbol{f}^{(j)}(I_k))^2}} = \frac{\sum_{i=1}^{512} (x_i' \times x_i)}{\sqrt{\sum_{i=1}^{512} (x_i')^2} \times \sqrt{\sum_{i=1}^{512} (x_i)^2}}. \tag{14}$$

根据特征向量之间的相似度度量计算结果的大小进行排序便可得到人脸图像检索初始排序结果列表 $\boldsymbol{Z}' = \{I_1'^{(j)}, \cdots, I_k'^{(j)}, \cdots, I_m'^{(j)}\}$ 及对应的深度特征向量 $\{\boldsymbol{f}^{(j)}(I_1'), \cdots, \boldsymbol{f}^{(j)}(I_k'), \cdots, \boldsymbol{f}^{(j)}(I_m')\}$. 排序结果列表 $\boldsymbol{Z}'$ 中人脸图像 $I_1'^{(j)}$ 为在数据库中保存的人脸图像数据集中与待检索人脸图像 I' 最相似、检索结果最匹配的人脸图像. 相反,人脸图像 $I_m'^{(j)}$ 则表示与待检索人脸图像 I' 最不相似、检索结果最不匹配的人脸图像.

2.4 人脸特征融合查询扩展

通过上述方法得到人脸图像检索的初始排序结果列表 $\boldsymbol{Z}' = \{I_1'^{(j)}, \cdots, I_k'^{(j)}, \cdots, I_m'^{(j)}\}$ 及对应的深度特征 $\{\boldsymbol{f}^{(j)}(I_1'), \cdots, \boldsymbol{f}^{(j)}(I_k'), \cdots, \boldsymbol{f}^{(j)}(I_m')\}$. 其中,待检索人脸图像的深度卷积特征向量 $\boldsymbol{f}(I')$ 与检索结果列表 $\boldsymbol{Z}'$ 中的人脸图像所对应的深度特征向量之间满足距离函数 d 的关系如下: $d(\boldsymbol{f}(I'), \boldsymbol{f}^{(j)}(I_i')) \leqslant d(\boldsymbol{f}(I'), \boldsymbol{f}^{(j)}(I_{i+1}'))$, $1 \leqslant i < m$. 得到初始人脸图像检索结果后,采用初始检索排序结果前 q $(1 \leqslant q \leqslant 10)$ 个人脸图像的卷积特征向量的均值进行查询扩展(Query Expansion, QE)以提高人脸检索的结果,实验中 $q = 10$:

$$\boldsymbol{f}(I') = \frac{1}{q+1}(\boldsymbol{f}(I') + \sum_{i=1}^{q} \boldsymbol{f}^{(j)}(I_i')). \tag{15}$$

通过人脸图像卷积特征向量的均值查询扩展方法将均值特征向量作为待检索人脸图像的特征向量 $\boldsymbol{f}(I')$ 再次执行检索任务得到最终的有序排序结果

列表 $\boldsymbol{Z}' = \{I_1'^{(j)}, \cdots, I_k'^{(j)}, \cdots, I_m'^{(j)}\}$.

通过查询扩展方法对待检索人脸图像的卷积特征向量进行融合加强，能够进一步提高检索的结果.

3　实验及分析

实验在以下平台实现：CPU 为 Intel(R) Core(TM) i5-6600 CPU @ 3.30 GHz，内存为 16 G，GPU 为 Nvidia Titan X 平台，操作系统为 Ubuntu 14.04.4 LTS，程序是基于 Caffe 环境采用 python 语言实现.

3.1　实验数据集

实验采用以下人脸图像数据集进行性能评估：

1) LabeledFaces in the Wild dataset (LFW)[19]：该数据集包含 5 749 个不同人的 13 233 张人脸图像，是人脸图像识别的标准验证基准.

2) 明星人脸数据集(Celebrities Face Set, CFS)：在线收集了 40 位明星人脸图像数据集，每人 300 张左右的人脸图像共12 063 张人脸图像. 图像数据集收集过程如下：

a) 确定明星名单：为了确保能够收集到足够多的人脸图像建立数据库，首先选取能够在图像搜索引擎搜索到大量正面人脸图像的明星名单；

b) 人脸图像收集：使用百度图像搜索引擎收集每个明星 300 张左右的人脸图像；

c) 数据集提纯过滤：去除同一人错误和严重遮挡的人脸图像，将保留下的人脸图像统一裁剪为像素大小 224×224 的尺寸. 最终收集的明星人脸图像数据集示例见图 3.

图 3　人脸图像示例

Fig.3　Examples of face image

3.2　实验参数设置

训练网络模型学习 N 路人脸分类器遵循文献[5-6]的步骤和修改，目标是找到最小化 softmax 层之后的平均预测对数损失的网络参数. 网络的配置，通过随机梯度下降和 0.9 的动量系数优化，权重衰减系数为 5×10^{-4}，在 fc6 和 fc7 的两个全连接层之后以 0.5 的比率让节点输出置 0，学习率初始设置为 10^{-2}，在验证集精度停止增加时减少到10^{-3}，滤波器权重通过零均值和标准偏差为10^{-2}的高斯分布随机采样初始化，网络偏置项初始化为零.

3.3　人脸图像特征可视化

使用图 1 中所示深度卷积神经网络提取人脸图像的深度卷积特征并进行可视化见图 4. 图 4 中第一列为原始人脸图像，后面五列依次为卷积层 Conv1_2 层、Conv2_2 层、Conv3_3 层、Conv4_3 层和 Conv5_3 层提取的人脸图像特征. 从图中可看出底层特征可明显分辨出人脸轮廓纹理等信息，而高层语义特征更加抽象稀疏，更具表征性.

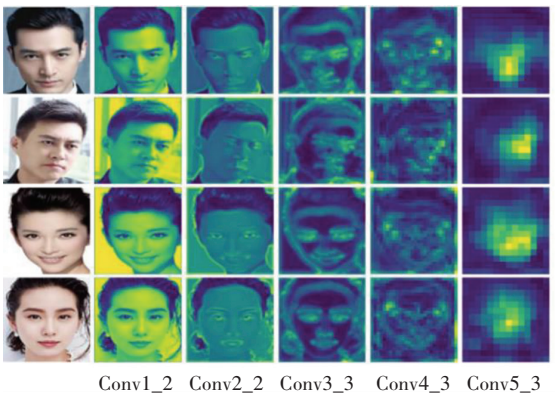


图 4　人脸图像特征可视化

Fig.4　Visualization of face image features

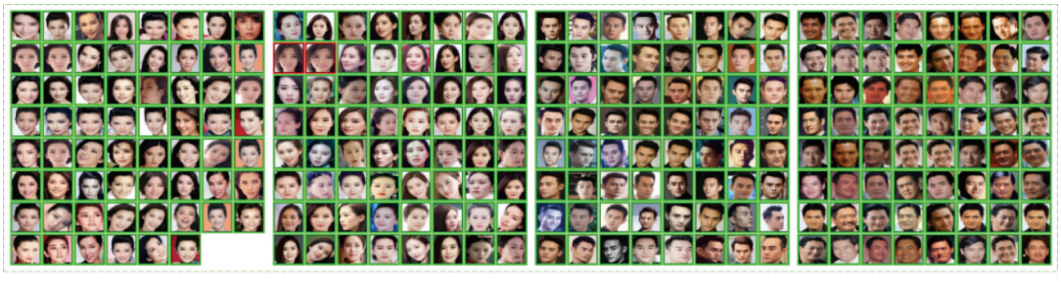
为验证图 1 所示 CNN 提取的人脸图像深度特征能够很好地表征人脸图像，图 5 显示了 7 位明星的部分人脸图像分别利用 LBP 特征、HOG 特征、CNN 提取的 Conv3_3 层特征、Conv4_3 层特征和 Conv5_3 层特征分别进行 K-means 聚类将对应的图像划分为 7 个簇后的聚类结果(图中只列出了每种特征聚类结果的其中 4 个簇的划分结果). 图中边界框为绿色的图像表示聚类正确的图像，边界框为红色的图像表示错误聚类的图像. 从图 5 中可看出采用图像浅层纹理特征的 LBP 特征、HOG 特征、CNN 提取的 Conv3_3 层特征进行 K-means 聚类并不能将相应的图像正确的划分为一类. 而采用 CNN 提取的图像高层语义特征 Conv4_3 层以及 Conv5_3 层特征进行 K-means 聚类能够将相应的图像正确的划分为一类，从图 5 中(d)、(e)两组聚类结果可看出通过使用 Conv5_3 层的图像高层语义特征进行 K-means 聚类得到的图像聚类结果最理想，能够将同一个人的人脸图像正确划分为一类. 经过实验验证，当聚类数目小于明星人数的时候聚类结果会更加可靠的将同一人的图像划分到同一个类中.

(a) LBP 特征聚类图像结果

(b) HOG 特征聚类图像结果

(c) Conv3_3 层特征聚类图像结果

(d) Conv4_3 层特征聚类图像结果



(e) Conv5_3 层特征聚类图像结果

图 5 不同特征对部分图像聚类结果对比

Fig.5 Comparison of clustering results of partial images by different features

3.4 人脸图像检索结果验证

使用人脸分类训练和三元组损失微调后的 CNN 在 CFS 和 LFW 两个数据集上进行了人脸深度特征提取并进行 K-means 聚类,采用 Elbow Method

估计最优聚类数量 k 值，见图 6. 图中横坐标为聚类数 k，纵坐标为所有样本聚类误差平方和 E. 根据 Elbow Method 选取肘部对应的 k 值为最佳聚类数，图 6 中肘部对应 k 值为 2，因此最佳聚类数 k 应该选取 2.

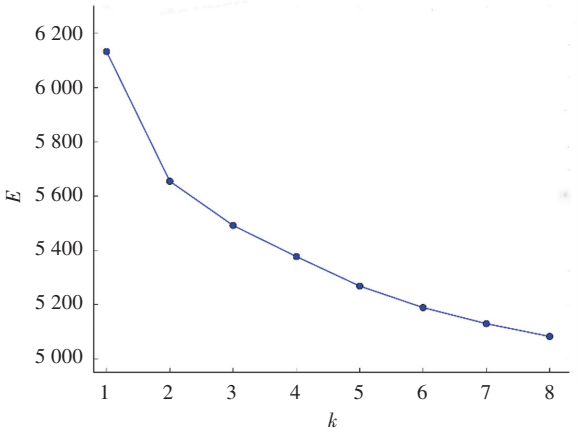


图 6　Elbow Method 估计 k 值

Fig.6　Elbow Method estimates the value of k

通过 Elbow Method 确定最佳的聚类数 k 后将图像划分为不同的类进行人脸检索性能验证，为验证通过采用 Elbow Method 确定最佳 k 值的有效性，实验中分别选取不同的 k 值进行聚类分析，见表 1. 从表 1 中实验结果数据可看出，在相同的条件下当图像集划分为 2 个簇或 3 个簇时，使得检索范围变小，能够在更小的数据集中快速检索正确的结果. 在不使用查询扩展方法和使用查询扩展方法两种情况下都能够显著降低检索时间，可节省约一半的检索时间，同时平均检索准确率达到最高，检索结果最佳. 从表中实验结果可以看出，当聚类数 k 值为 2 时，在不使用查询扩展方法时平均检索准确率达到 95.12%，使用查询扩展方法后将平均检索准确率提高到 95.75%. 从实验结果得出，通过 Elbow Method 确定最佳的聚类数 k 值后，对深度特征进行 K-means 聚类将海量人脸图像划分为不同的簇后，然后在相应的最近簇内进行相似人脸图像检索具有一定的可靠性和可行性，能够显著降低检索时间.

表 1　不同情况下人脸检索结果

Tab.1　Face retrieve results in different situations

数据集	聚类数目	查询扩展	速度/秒	检索结果前 100 平均准确率/%	检索结果前 200 平均准确率/%	平均准确率/%
明星人脸数据集+ LFW	1	No	0.331	98.16	94.12	92.11
		Yes	0.3299	99.34	97.13	95.7
	2	No	0.138	98.16	94.12	92.12
		Yes	0.138	99.34	97.14	95.75
	3	No	0.135	98.16	94.12	92.12
		Yes	0.136	99.34	97.14	95.75
	4	No	0.132	98.16	94.2	92.03
		Yes	0.131	99.34	97.17	95.66
	5	No	0.131	98.16	94.2	92.03
		Yes	0.143	99.34	97.17	95.66
	6	No	0.13	98.16	94.2	92.03
		Yes	0.13	99.34	97.17	95.66
	7	No	0.13	98.16	94.2	92.03
		Yes	0.13	99.34	97.17	95.66
	8	No	0.13	98.16	94.2	92.03
		Yes	0.135	99.34	97.18	95.66

图 7 中列出部分人脸图像检索排序前 10 张的检索结果示例图，其中每行的第一列人脸图像包围边框为蓝色的表示待检索的人脸图像，第一列后的包围边框为绿色的人脸图像表示相对应的被正确检索的结果，包围边框为红色的人脸图像表示被错误检索的结果. 从图中直观的看出排在最前面的人脸图像检索结果与待检索的人脸图像更为相似，包括人脸的朝向及面部表情.

3.5　实验结果及时效性比较

将文中提出的人脸图像检索方法与文献中的检索方法进行比较，见表 2.

表 2 中显示了算法在两个人脸图像数据集上执行检索结果的平均准确率和检索每张人脸图像所用平均时间的时效性对比.

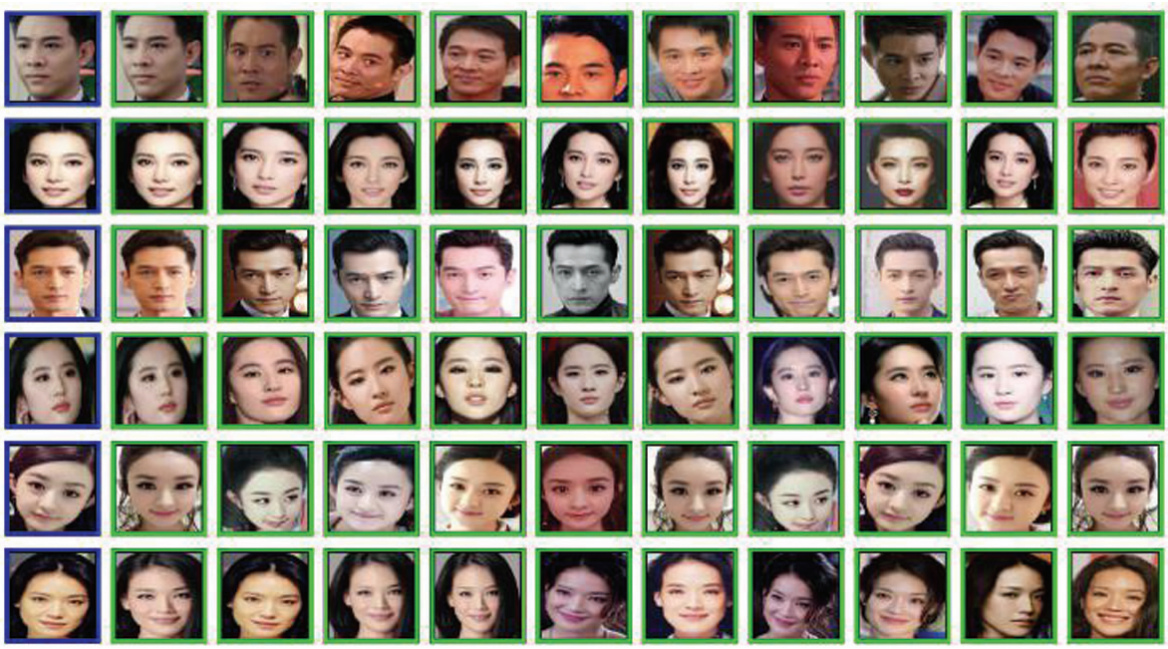


图 7 部分人脸图像检索结果

Fig.7 Part of face image retrieve results

表 2 结果及时效性比较

Tab.2 Results and timeliness comparison

方法	检索平均准确率	速度/秒
L-K one node[3]	77.5%	0.33
L-K two node[3]	86.0%	0.67
L-K three node[3]	91.0%	1.05
L-K four node[3]	93.5%	1.32
L-K five node[3]	95.1%	1.65
DCNNFR	92.12%	0.138
DCNNFR+QE	95.75%	0.138

从表 2 中看出,刘小华等提出的用于人脸图像检索的 L-K 均值层次聚类方法[3]在类内数据量和检索节点增加时所得到的人脸检索结果平均准确率明显提高,在没有时间限制的情况下最好的检索准确率达到 95.1%,但是该方法随着检索准确率的提升,时间消耗显著增加,是以时间开销为代价换取检索平均准确率. 因此,该方法在现实应用场景中在保持一定准确率的检索结果时不能达到快速实时的检索速度. 而本文提出的方法通过训练好的深度卷积神经网络用于人脸检索(Deep Convolutional Neural Network Face Retrieval, DCNNFR)模型平均提取和构建每张人脸图像的深度特征只需要 0.02s 的时间开销,从表中看出该方法在保证一定检索准确率的同时能够达到近实时的时效性,在人脸图像集聚类最好的情况下检索每张图像平均用时 0.138s,检索过程简单快速无需人为参与设计特征的提取和检索过滤.

4 结 论

本文提出了一种基于深度特征聚类的海量人脸图像检索方法. 通过使用人脸图像训练集对深度卷积神经网络模型进行人脸分类训练,在此基础上采用三元组损失方法对已训练好的人脸分类网络模型进行微调,使得网络能够更加有效地提取人脸图像深度特征,构建更具有表征的高层语义特征. 采用 K-means 聚类算法对提取的深度特征进行聚类,使得对应的人脸图像集划分为不同的簇,然后在相应的簇中进行人脸图像特征相似度匹配执行检索任务. 最后通过查询扩展方法对待检索人脸图像深度特征进行融合进一步提高检索性能. 实验结果证明,该方法能够根据深度特征快速地实现人脸图像的簇划分,极大地缩小海量人脸图像检索范围,在保证一定准确率的同时有效地提高了人脸图像检索速度.

参考文献

[1] LIU C. Gabor-based kernel PCA with fractional power polynomial models for face recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2004, 26(5):572. DOI: 10.1109/TPAMI.2004.1273927

[2] ZHENG X, CAI D, HE X, et al. Locality preserving clustering for image database[C]// ACM International Conference on Multimedia. New York: ACM, 2004:885. DOI: 10.1145/ 1027527.1027731

[3] 刘小华, 周春光, 张利彪, 等. 海量人脸数据库的快速检索[J]. 吉林大学学报(工), 2010, 40(1):183. DOI: 10.13229/j.cnki.jdxbgxb2010.01.036

LIU Xiaohua, ZHOU Chunguang, ZHANG Libiao, et al. Method of quick searching in a huge scale face database[J]. Journal of Jilin

University (Engineering and Technology Edition), 2010, 40(1): 183. DOI: 10.13229/j.cnki.jdxbgxb 2010.01.036

[4] 杨之光，艾海舟. 基于聚类的人脸图像检索及相关反馈[J]. 自动化学报，2008，34(9):1033. DOI: 10.3724 /SP.J.1004.2008. 01033

YANG Zhiguang, AI Haizhou. Cluster-based Face Image Retrieval and Its Relevance Feedback[J]. Acta Automatica Sinica, 2008, 34 (9):1033. DOI: 10.3724/SP.J.1004.2008. 01033

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C] //Advances in neural information processing systems. Long Beach: Curran Associates Inc.2012: 1097. DOI: 10.1145 /3065386

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv Preprint ArXiv:1409. 1556, 2014

[7] ZHANG K, ZHANG Z, LI Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE Signal Processing Letters, 2016, 23(10): 1499. DOI: 10.1109/ LSP.2016.2603342

[8] LI H, LIN Z, SHEN X, et al. A convolutional neural network cascade for face detection[C]// Computer Vision and Pattern Recognition. Boston: IEEE, 2015:5325. DOI: 10.1109/CVPR. 2015. 7299170

[9] PARKHI O M, VEDALDI A, ZISSERMAN A. Deep Face Recognition[C]// British Machine Vision Conference. Newcastle: BMVC, 2015: 41.1. DOI: 10.5244/C.29.41

[10] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston: IEEE, 2015: 815. DOI: 10.1109/CVPR.2015.7298682

[11] RAZAYIAN A S, AZIZPOUR H, SULLIVAN J, et al. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus: IEEE Computer Society, 2014:512. DOI: 10.1109/CVPRW.2014.131

[12] YANDEX A B, LEMPITSKY V. Aggregating Local Deep Features for Image Retrieval[C]// IEEE International Confer-ence on Computer Vision. Santiago: IEEE, 2016:1269. DOI: 10.1109 /ICCV. 2015.150

[13] TOLIAS G, SICRE R, JEGOU H. Particular object retrieval with integral max-pooling of CNN activations[C]// Proc International Conference on Learning Representations (ICLR), Lille, France, 2016:1. DOI: arXiv ID: 1511.05879

[14] KALANTIDIS Y, MELLINA C, OSINDERO S. Cross-Dimensional Weighting for Aggregated Deep Convolutional Features[C]// European Conference on Computer Vision. Switzerland: Springer, Cham, 2016:685. DOI: https://doi.org/10.1007/ 978-3-319-46604-0_48

[15] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision & Pattern Recognition. San Diego: IEEE Computer Society, 2005:886. DOI: 10.1109/CVPR.2005.177

[16] OJALA T, Pietikäinen M, MÄENPÄENPÄÄ T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2000, 24(7):971. DOI: 10.1109/ TPAMI. 2002.1017623

[17] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1):48. DOI: 10.3724/SP.J.1001.2008.00048

SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research [J]. Journal of Software, 2008, 19(1):48. DOI:10. 3724/SP.J. 1001.2008.00048

[18] ROBERT L. THORNDIKE (December 1953). "Who Belongs in the Family?". Psychometrika. 18 (4): 267. DOI: 10. 1007/ BF02289263

[19] HUANG G B, MATTAR M, BERG T, et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments[J]. Technical Report 07-49, University of Massachusetts, Amherst, 2007. DOI: 10.1.1.122.8268

（编辑　苗秀芝）