结合卷积神经网络多层特征融合和 K-Means 聚类的服装图像检索方法

侯媛媛 何儒汉 李 敏 陈 佳
(武汉纺织大学湖北省服装信息化工程技术研究中心 武汉 430200)

摘 要 随着服装电子商务的蓬勃发展,海量的服装图像数据被累积,对服装图像"以图搜图"成为了当前的一个热点研究方向。服装图像有着丰富的整体语义信息和大量细节信息,要对其实现精准检索是一项挑战性难题。传统的基于人工语义标注的服装图像方法和以人工设计的颜色与纹理等内容特征进行服装图像检索的方法均存在较大局限性。文中利用卷积神经网络多层特征融合提取特征,然后使用 K-Means 聚类加快服装图像的检索,充分利用深度卷积神经网络在图像特征提取上的有效性和层次性,融合不同卷积层次特征的细节信息和抽象语义信息以提升检索的准确度,并利用 K-Means 加快检索速度。所提方法首先对服装图像数据集进行统一的尺寸处理,然后利用卷积神经网络进行训练和特征提取,抽取出服装图像从低到高的多层次特征,进而将多种层次的特征进行融合,最终使用 K-Means 聚类方法对提取的图像库特征进行有效检索。在 DeepFashion 子类数据集 Category and Attribute Prediction Benchmark 和 In-shop Clothes Retrieval Benchmark 上的实验结果表明,所提方法能有效增强服装图像的特征表达能力,提高了检索准确率和检索速度,优于其他主流方法。

关键词 服装图像检索,卷积神经网络,特征融合,K-Means 聚类
中图法分类号 TP183 **文献标识码** A

Clothing Image Retrieval Method Combining Convolutional Neural Network Multi-layer Feature Fusion and K-Means Clustering

HOU Yuan-yuan HE Ru-han LI Min CHEN Jia
(Engineering Research Center of Hubei Province for Clothing Information, Wuhan Textile University, Wuhan 430200, China)

Abstract The booming of clothing e-commerce has accumulated a large amount of clothing image data, and the "image search" of clothing images has become a hot research direction. Apparel images have rich overall semantic information and a large amount of detailed information, and achieving accurate retrieval is a challenging problem. Traditional methods of clothing image based on artificial semantic annotation and methods of image retrieval based on artificially designed content features such as color and texture have significant limitations. This paper proposed a clothing image retrieval method based on multi-layer feature fusion and K-Means clustering of convolutional neural networks, which makes full use of the effectiveness and hierarchy of deep convolutional neural network in image feature extraction, fuses the detailed information and abstract semantic information of different convolutional hierarchical features to improve retrieval accuracy, and uses K-Means to improve the retrieval speed. The proposed method firstly performs uniform size processing on the clothing image data set, then uses the convolutional neural network for training and feature extraction, extracts multi-level features of the clothing image from low to high, and then fuses various levels of features. Finally, the K-Means clustering method is used to efficiently retrieve large-scale image data. The experimental results on the DeepFashion sub-category data set Category and Attribute Prediction Benchmark and In-shop Clothes Retrieval Benchmark show that the proposed method can effectively enhance the feature expression ability of clothing images, and improve its retrieval accuracy and retrieval speed. The proposed method is supprior to other mainstream methods.

Keywords Clothing image retrieval, Convolution neural network, Feature fusion, K-Means clustering

---

本文受国家自然科学基金面上项目(61170093)资助。

侯媛媛(1995—),女,硕士生,主要研究方向为计算机视觉、深度学习、图像分析与处理;**何儒汉**(1974—),男,博士,教授,主要研究方向为机器学习、计算机视觉、多媒体检索、图像和视频处理,E-mail:heruhan@wtu.edu.cn(通信作者);**李 敏**(1978—),女,博士,副教授,主要研究方向为计算机视觉、模式识别等;**陈 佳**(1982—),女,博士,副教授,主要研究方向为数据库、数据挖掘、图像处理等。

1 引言

近年来,随着服装电子商务的蓬勃发展,服装图像数据呈现爆炸性增长。面对庞大的服装图像数据,“以图搜图”为用户提供了一种崭新的网上服饰搜索模式,百度识图、GazoPa搜图和 Google 识图均是图像检索领域中的成功案例。服装图像检索是图像检索的一个分支,图像检索这项研究开始于20 世纪 70 年代,可以将其划分成两大类:基于文本的图像检索(TBIR)和基于内容的图像检索(CBIR)。TBIR 即以文本描述的形式描述图像特征,譬如图像的尺寸、出品时间、风格等;CBIR 是对图像的色彩、纹理等图像视觉内容特征进行分析和检索的图像检索技术。现今,大多数商业图像检索系统主要采用 TBIR 方式。京东、淘宝等购物平台中的图片检索是基于文本的。TBIR 方法的优点是检索的精度高且检索速度快,但存在的缺点是:1)要对每一张图片事先进行关键字文本的标注,随着网上图片数据量的剧增,对其进行人工标注会耗费巨大的人力和物力;2)人工标注存在一定的主观性,即不同的人对同一幅图片往往会有不同的理解,这种主观性以及不确定性会影响检索的效果。而 CBIR 使用颜色特征以及纹理特征等对服装图像进行检索,如 HOG[1],SIFT[2],ORB[3]等在图像处理上都有很成功的应用;但由于其只能提取服装图片的浅层特征,服装图像检索的准确率仍然不高,同时特征描述子缺乏一定的学习能力,限制了图片内容的表达能力,难以适应大量不同规格的图片数据。

服装作为一种时尚、品味、个性展示的主要载体,其图像有着丰富的语义信息和大量细节信息,包括颜色、款式、样式、风格、类别等整体信息,以及纹理、材质、花形、领型、褶边等细节信息;同时,服装是典型的柔性、非刚性物体,具有高度可变形性,外形结构不规则、不完整,其图像对光照、视角、尺度等更敏感。因此,要实现快速、准确的服装图像检索,是一项极具挑战性的难题。

鉴于 TBIR 方法和传统的 CBIR 方法均存在较为明显的局限性,受深度卷积神经网络强大的特征提取能力和多特征融合与聚类思想的启发,本文提出一种新的服装图像检索模型结构,即基于卷积神经网络(Convolutional Neural Network,CNN)多层特征融合和 K-Means 聚类的服装图像检索方法。该方法在一个较大规模的服装图像数据集DeepFashion 上,综合利用卷积神经网络的浅层和深层多层特征来解决服装图像检索的特征提取问题,获得了较好的特征表达能力,有益于取得较好的检索准确度;经过深度学习提取的图像特征的维度都比较高,通过全连接层减少参数来降低维度,然后通过分类器得到最终的特征向量,最后再根据输出进行 K-Means 聚类,极大地加快了检索的速度。本文方法采用 CNN 和 K-Means 相结合的结构,避免了人工标注所带来的弊端和手工设计特征产生的特征表达能力不足的问题,实现了对大规模服装图像的快速、高效检索,在检索准确度和检索速度方面取得了较好的平衡,相比于其他主流方法,具有一定的优势。综合来看,本文所提方法有效利用了深度卷积神经网络强大的特征提取能力,并采用多特征融合的思路提升了检索准确度,同时利用聚类方法提高了检索速度。

本文第 2 节简要介绍一些图像检索近年来的相关工作;第 3 节详细介绍图像检索的方法;第 4 节给出实验结果,并进行相应的对比;最后总结全文。

2 相关工作

针对服装图像检索问题,研究者们提出了一系列的服装图像检索方法。贾巧丽等[4]综合颜色和款式,利用不变矩和离心率做基于形状特征的检索,然后再利用颜色直方图做二次检索。薛培培和邬延辉[5]融合了图片颜色特征和尺度不变特征(SIFT),实现了图像的检索,并在此基础上利用支持向量机加入相关反馈技术来提高图像的检索准确率。胡玉平等[6]消除服装图像背景的影响,针对 GrabCut 算法存在对图像局部像素值的变化敏感、时间开销大、边缘不准确等问题进行了改进。自 Hinton 等[7]提出了利用深度神经网络来从大量的训练数据集中提取特征,不同结构的神经网络已经被成功运用在各种图像处理中,通过深度的神经网络发现训练图片信息中隐藏的表征特征。Ciresan 等[8]用训练样本,(可以是相同的样本,也可以是不同的样本)去训练多个深度卷积神经网络(Deep Convolutional Neural Networks,DNN),用多个 DNN 进行分类,结果取平均。Cirshick 等[9]用 selective search 代替传统的滑动窗口,提取出 $2k$ 个候选 region proposal,对于每个 region,以 AlexNet 网络为网络基础模型,去掉其 softmax 层,训练出 L 个(L 为目标类数)SVM 分类器;利用 AlexNet 提取的特征作为输出,得到每个 region 在每个类别的得分,最终对每个类别用 NMS(Non-Maximum-Suppression)来舍弃掉一部分 region,得到 detection 的结果(对得到的结果做针对 boundingbox 的回归,用来修正预测的 boundingbox 的位置)。Lin 等[10]对预训练的模型进行微调,从而得到类哈希的图像特征,最后利用深度网络进行图像的检索。Krizhevsky 等[11]首先从卷积层的第七层提取特征来检索图像,并在 ImageNet 上得到了较好的实验结果。Kiapour 等[12]研究了跨场景的服装图像检索,他们将全连接层的输出作为特征表示,之后将余弦相似度作为特征距离的计算依据,得到检索结果。

最近几年,深度学习被广泛应用于图像处理等领域,众多成功案例表明:基于深度学习的图像特征提取不同于传统的浅层学习算法,其通过深层模型结构学习其隐藏的特征,同时通过逐层特征变换将原样本在原空间下的特征变换到另一个空间特征之下,使图片的分类和检索更加有效。深度学习的本质可以解释为:通过搭建许多隐层的机器学习模型,利用海量的训练数据,学习得到更加有用的特征,进一步提高分类和预测的准确性。卷积神经网络[13]是一种运用最为广泛的深度学习方法之一,其思想最早是在 1989 年由 LeCun 等人提出的,并且成功应用于英语手写字体识别中,同时在其他众多应用中均取得了良好效果,如被用于图像分类中[14]。在文献[15]中提出以 LeNet-L 为基础的图像检索模型,最后通过距离函数比较待检索图像与图像库的相似度。因为较低层的特征通常比较高层的特征简单,简单的图像在浅层所提取的特征能很好地表达出复杂的图像背景。因此,利用融合多阶段的特征可以进一步提升图像的分类能力。基于这一发现,文

献[16]提出了一种多阶段 CNN 特征融合方法，该方法仅设计了 3 个卷积层，其中利用了顶部两层的特征。文献[16]中 CNN 模型的深度相对较浅，因此可能无法完全探索多阶段的特征。

针对服装检索问题，目前的图像数据集规模比较小，而基于深度学习的方法则需要大规模的服装数据集进行训练和学习。同时，为了克服文献[16]网络层较浅的缺点，本文提出一种基于 GoogleNet[17] 的多层特征融合的网络模型。基础 GoogleNet[17] 模型超过 20 层，比文献[15-16]网络结构更深。另一方面，由于相邻层之间具有特征相似性且模型复杂性较高，因此不必融合所有中间层特征；并且本文利用一个较大规模的服装图像数据集 DeepFashion 多网络模型进行微调，提取服装图像的多层特征并予以综合，最后结合 K-Means 聚类算法完成服装图像的检索。

3　基于卷积神经网络多层特征融合和 K-Means 的服装图像检索

本文提出基于卷积神经网络和 K-Means 聚类算法相结合的服装图像检索模型，使用 DeepFashion 数据集中的 Category and Attribute Prediction Benchmark 中 50 种类别、1 000 种属性、13 万数量级的图片，其中包含 10 万多张训练图片和 3 万多张测试图片。采用卷积神经网络对不同种类的服装图片进行特征提取，得到特征库信息，最后结合 K-Means 聚类完成检索。

3.1　卷积神经网络特征训练模型

卷积神经网络是一种多层前馈网络，含有输入层、卷积层、池化层和输出层，每一层都由许多二维平面所组成，每个平面又由多个神经元组成。输入层能够直接接收二维视觉模式，譬如二维图像。卷积层也称为特征抽取层，每个神经元的输入来自前一层的局部感受域，从而提取出该局部的特征。卷积层的运算过程如下：

$$X_j^l = f\left(\sum_{i\in M_j^n} X_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

其中，X_j^l 表示第 l 层的第 j 个运算结果；$f(\cdot)$ 表示激活函数；M_j^n 表示第 n 层第 j 个输出特征图所对应的多个输入特征图的索引集合；b_j^l 表示偏置项，被所有输入特征图共用；k_{ij}^l 表示第 l 层中一个 $i\times j$ 大小的卷积核。

池化层为特征映射层，每个特征映射是一个平面，平面上所有神经元的权值相等。池化层可以实现平移、旋转、尺度等不变性，不仅能保存主要的特征信息，而且能够减少参数（降维，效果类似于 PCA）和计算量，防止过拟合，提高模型的泛化能力。池化层的计算过程如下：

$$X_j^l = f(\beta_j^l down(X_j^{l-1}) + b_j^l) \tag{2}$$

其中，$down(\ )$ 表示一个下采样函数，该操作能对输入图像不同块中的所有像素进行求和，从而使得输出的图像在两个维度上都能缩小到原图的 $\frac{1}{n}$。每个输出的 map 都拥有属于自己的一个乘性偏置 β 和加性偏置 b。

3.1.1　多层特征融合的网络模型

在卷积神经网络的多层结构中，不同层特征映射的意义不尽相同，在浅层中提取的一般为服装图片的颜色、纹理特征等局部的和具体的信息，而在深层提取的特征代表了图片总体的和抽象的特征。图 1 为一张图片在一个简单卷积网络中各层特征的可视化图片，从图中可看出浅层网络提取的是纹理、细节特征，包含的特征更多；深层网络提取的是轮廓、形状，相对而言，层数越深，提取的特征越具有代表性，图像的分辨率是越来越小的。

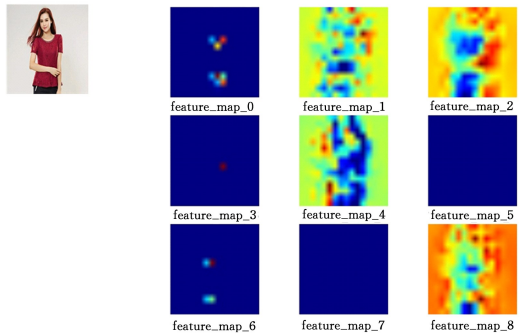

图 1　卷积层中各层提取的特征图

近年来，计算机性能大幅提升，训练的数据集随之快速发展，依次出现了 AlexNet[13]，VGG-Net[18]，GoogleNet[18]，ResNet[19] 等性能优秀的卷积神经网络架构。本文的网络模型基于 GoogleNet[17]，其网络层数更深，且出现了一个新颖的 Inception 组件，这是一种网中网（Network In Network）的结构，即原来的结点也是一个网络，图 2 所示为本文 Inception 组件结构。Inception 模块由一些小尺寸的卷积核（如 1×1，3×3 和 5×5）组成，有利于限制参数的规模和模型的复杂性。

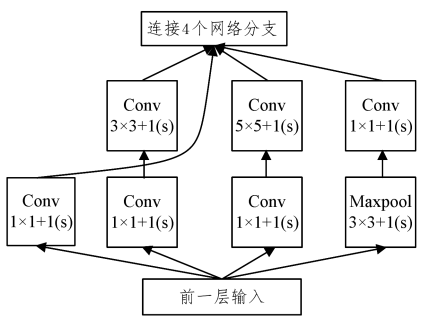

图 2　Inception 组件结构图

基于特征融合思想，本文提出了一种基于 GoogleNet 的多层特征融合方案，充分利用了高层次特征的丰富语义信息和低层次特征的细节信息，并形成优势互补，有利于捕获服装的整体风格特征和款式细节特征。在 GoogleNet 中有 9 个 Inception 组件，为了克服梯度消失和过度拟合的问题，将这 9 个 Inception 模块分为 3 组。第一步，输入层输入一张 224×224×3 的图片信息，分别经过 3 个卷积层 Conv1，Conv2，Conv3（卷积核大小分别为 7×7，1×1，3×3），输出体积为 28×28×192 的信息。第二步，经过第一个 Inception 组得到 14×14×512 的信息，在此处进行特征提取，在提取特征之前经过 average-pooling 在全局进行平均池化操作。相较于 max-pooling，average-pooling 在减少维度的同时，更多地保留了图片的背景信息，有利于信息传递到下一个模块进行特征提取。由于特征融合会使网络变大，增大计算难度，因此通过卷积核为 1×1 的卷积做降维；同时，为了获取相同大小的维度而更有利地融合，进行了全连接，得到一个 1 024 维的特征

向量作为特征融合的第一个向量。第三步，将第二步中 14×14×512 的信息作为输入，做与第二步相似的操作提取出 1024 维特征向量。第四步，将第二个 Inception 组的输出作为输入，得到 7×7×1 024 的输出，然后经过平均池化得到 1×1×1024 的信息，此处维度足够小，不用通过卷积，直接经过全连接层得到一维向量。第五步，将三部分分别提取到的 1024 维特征向量线性地连接，以形成 3072 维特征向量，最后经过 softmax 层得到一个 50 维（数据集中有 50 种类别的服装）的特征向量。图 3 所示为网络模型。

在该网络模型中，融合了三层网络特征，分别是 3 个 inception 组的输出。为了保证不同组之间在特征映射面上的大小一致性，对第一组 inception 模块和第二组 inception 模块的输出均做了平均池化、1×1 卷积和全连接操作，对第三组 inception 模块的输出做了平均池化和全连接操作。

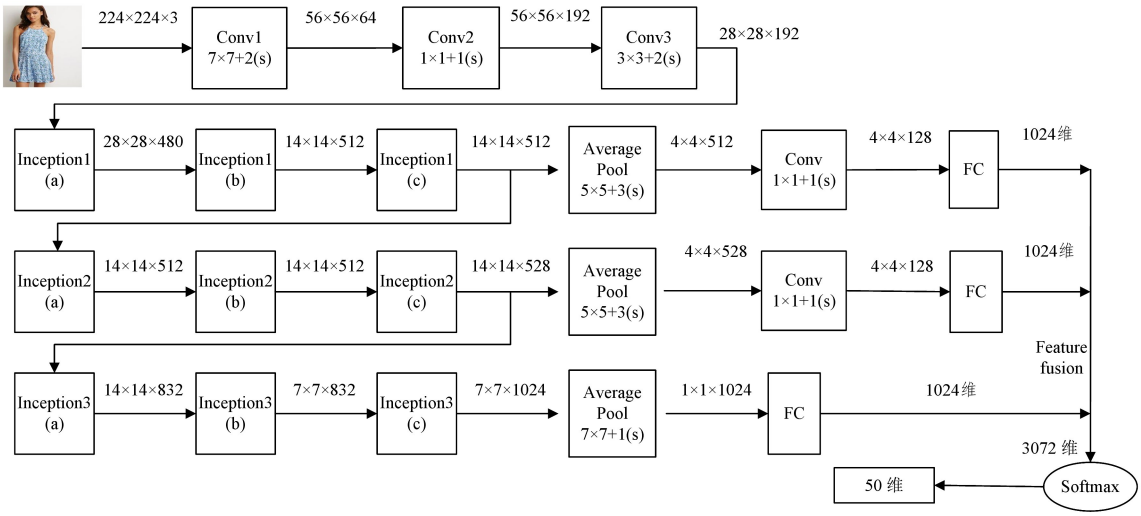

图 3 多层特征融合的网络模型

3.1.2 训练时前向传播和后向传播

1）前向传播阶段：从训练样本中选取一个样本（X，Y），X 从输入层经过逐层变换传送到输出层，计算相应的实际输出 $Y\_$。

2）后向传播阶段：计算出实际输出 $Y\_$ 与期望输出 Y 之间的误差。对于模型的训练，我们设置损失函数 $Loss$，它由两部分别组成，即：

$$Loss=CrossEntropyLoss+TripletLoss \tag{3}$$

其中第一部分为交叉熵损失函数：

$$C=-\frac{1}{n}\sum_x[Y\ln a+(1-Y)\ln(1-Y\_)] \tag{4}$$

其中，x 表示样本，n 表示样本的总数，Y 为期望的输出，$Y\_$ 为神经元的实际输出。

第二部分使用了 triplet 结构[20]，其最早出现在 Google 做人脸识别的一个网络模型——faceNet 中，并广泛用于人脸识别中。triplet 是一个三元组（*Anchor*，*Positive*，*Negative*），其构成部分为：训练数据集中随机选取一个样本 *Anchor*；随机选取一个与 *Anchor* 同属一类的样本 *Positive*；随机选取一个与 *Anchor* 不同类的样本 *Negative*。*triplet loss* 的作用是通过学习，让 *Anchor* 和 *Positive* 的特征表达之间的距离足够小，而让 *Anchor* 和 *Negative* 的特征表达之间的距离尽可能的大。Triplet loss 的公式为：

$$L=\sum_i^N[\|f(x_i^a)-f(x_i^p)\|_2^2-\|f(x_i^a)-f(x_i^n)\|_2^2+\alpha]_+ \tag{5}$$

式(5)中的距离公式是欧氏距离，+表示[]中的数值小于零时损失值为零，若大于零则取这个大于零的数值。从目标函数中可以解释：当 *Anchor* 与 *Negative* 之间的距离小于 *Anchor* 与 *Positive* 之间的距离加 α 时，[]中的结果大于零，就会产生损失；当 *Anchor* 与 *Negative* 之间的距离大于 *Anchor* 与 *Positive* 之间的距离加 α 时，L 则为零。

我们通过调整两部分损失函数的权重来使得模型拥有更好的预测结果，最后将训练的模型存储下来。

3.2 特征向量的相似性度量

提取图片的特征向量之后，我们要将检索图片的特征向量与训练出的数据库中图片的特征向量进行相似度度量。相似性度量，即综合评定两个事物之间相近程度的一种度量。两个事物越接近，它们的相似性度量也就越大；两个事物越疏远，它们的相似性度量也就越小。不同相似性度量对于算法的结果，有时差异很大。本文采用的是余弦相似度，即计算两个向量的余弦值作为其相似性的表征，计算公式为：

$$d(x,y)=1-\cos(x,y)=1-\frac{x^{\mathrm{T}}y}{|x||y|} \tag{6}$$

其中，x 和 y 表示两个 n 维的特征向量，即 $x=\{x_1,x_2,\cdots,x_n\}$，$y=\{y_1,y_2,\cdots,y_n\}$；$d(x,y)$ 表示两个向量的距离。选择出距离最小的 10 张图片作为一个返回结果。

3.3 特征聚类

对服装图像进行检索时，是根据特征的匹配来输出结果的，如果数据量较少且特征维数较少，则可以一一计算待检索图片的特征向量与特征库中每个向量之间的距离，然后将距离值最小的几张图片作为结果返回。但是，我们选取的数据库中服装图片多达十几万张，使用这种计算方法会大大降低检索的速率。针对这种问题，我们使用 K-Means 聚类来进行检索，以缩小检索图片的范围，加快检索速率。

服装图片的聚类，是根据特征向量的相似度来实现的，属于无监督学习。其主要目的是将相似的特征向量分到同一个类簇中，并计算出每个类簇的聚类中心。在检索时，首先与所

有的类簇中心进行一一对比，选取距离最小的一个类簇，然后再与该类簇的所有数据进行一一比较。这种方法可以极大地缩小检索范围，提高了检索速率。在 K-Means 聚类算法中，值聚类个数 K 的选取以及聚类中心的选取对聚类结果都会产生很大的影响。因为我们使用的数据集有 50 个类别，所以本实验中 $K=50$，具体聚类步骤如下。

1）从训练的特征库中任意选择 K 个特征作为初始聚类中心；

2）计算特征库中其他特征对象到每个聚类中心的距离，并将这些特征划分到距离最近的那个类簇；

3）计算出每个类簇的平均值，并将其作为该类簇新的聚类中心；

4）重复步骤 2）和步骤 3），直到新的聚类中心与原聚类中心的距离小于一个指定的阈值。

3.4　实验总体步骤

1）对服装图片数据集做预处理之后，将大量的图片输入到本文的网络模型中进行监督学习，调整权值和偏置，得到一个训练模型并保存下来。

2）将数据集输入训练的数据集，分别提取中间 3 个 Inception 模块输出的特征并进行融合，然后通过降维减少参数与计算时间，最后通过分类器得到最终的特征向量。

3）将提取的特征作为 K-Means 算法的输入进行聚类，将相似的特征向量分配到同一类簇，并将类簇中心作为主要信息。

4）输入待查询的服装图片，将图片进行预处理，通过本文网络结构提取到待检索图片的融合特征。

5）将输入图片的特征向量与各个类簇中心的特征向量进行相似度度量，选择距离最小的中心，然后再与该类簇的所有特征一一比较，以缩小检索的范围，提高检索速率。最后，将与目标图片特征向量距离最近的 10 张图片作为结果返回。

与传统的特征提取方法相比，CNN 能够获取更深层次的信息特征，且融合了浅层的局部特征，这些特征信息大大提升了服装检索的准确率和速率。

4　实验结果与分析

4.1　实验数据准备

为了验证所设计模型的效果，本文选用了 DeepFashion 的两个个子集 Category and Attribute Prediction Benchmark 和 In-shop Clothes Retrieval Benchmark 作为数据集。DeepFashion 是香港中文大学开放的一个 large-scale 数据集，包含不同角度、不同场景、买家秀等 80 万张图片，每张图片也有非常丰富的标注信息，包括 50 种类别、1000 种属性、Bbox、特征点，还有约 30 万的不同姿势/不同场景的图片 pairs。DeepFashion 有 4 个子集：Category and Attribute Prediction Benchmark；In-shop Clothes Retrieval Benchmark；Consumer-to-shop Clothes Retrieval Benchmark；Fashion Landmark Detection Benchmark。Category and Attribute Prediction Benchmark 子集是用来做分类和属性预测的，共有 50 种分类标记，1000 种属性标记，包含 13 万张图像，分为 100 983 张训练图片和 38 726 张测试图片。每张图像都有 1 个类别标注、1000 个属性标注、Bbox 边框和 landmarks。服装数据集的分类如图 4 所示。

图 4　类别和属性预测基准

4.2　实验设置与测评指标

在检索策略中，通常采用平均查准率均值（Mean Average Precision，mAP）和召回率（Recall）两个评定标准来评价检索效果的好坏。准确率是检索结果中相似图片数量占全部检索出的图像数量的百分比，召回率是检索出的相似图片占数据集所有相似图片的百分比，定义如下：

$$mAP=\frac{\text{图像检索的平均查准率}}{\text{检索次数}}\times 100\% \tag{7}$$

$$Recall=\frac{\text{检索出的相关图像数}}{\text{全部相关图像数}}\times 100\% \tag{8}$$

4.3　实验结果

为了验证本文提出的方法在服装检索上的性能优越性，进行了一系列对比实验。选择了在图像检索方面常用的一些方法进行实验对比，所涉及的特征主要包括 HSV（Hue，Saturation，Value）颜色特征、方向梯度直方图（Histogram of Oriented Gradient，HOG）形状特征。HSV[21] 模型是由色度 H、饱和度 S、亮度 V 3 个分量组成的，通过划分 H，S，V 的区间来形成一定特征纬度的直方图向量。HOG[22] 特征是用来进行物体检测的特征描述子，用于计算机视觉和图像处理中，其特征是统计和计算图像局部区域的梯度方向直方图所得。HOG 特征与 SVM 分类器结合已经被广泛应用于图像识别中。对与 HSV 特征和 HOG 特征，使用 SVM[23] 作为分类器。CNN 方法是直接使用本文所提到的网络模型进行特征提取并融合多层特征，然后检索时与特征库的特征进行一一比对（FusionCNN）。FusionCNN＋K-Means 方法中的 CNN 模型是本文所提到的模型，然后在检索时使用聚类的方法加快检索速率。根据实验检索返回的前 10 张图片来计算准确率和召回率，实验结果如表 1 所列。

表 1　检索结果中前 10 张图片的 mAP 和 $Recall$ 比较

算法模型	Original		Watermark		Rotate		Crop		Mirror	
	mAP	*Recall*	*mAP*	*Recall*	*mAP*	*Recall*	*mAP*	*Recall*	*mAP*	*Recall*
HSV＋SVM	0.125	0.0508	0.193	0.0458	0.112	0.0488	0.146	0.0544	0.114	0.0491
HOG＋SVM	0.264	0.0725	0.278	0.0591	0.281	0.0658	0.262	0.0566	0.252	0.0644
FusionCNN	0.625	0.1289	0.615	0.0961	0.684	0.1069	0.424	0.0663	0.685	0.1227
FusionCNN＋K-Means	0.845	0.1398	0.885	0.1227	0.712	0.1113	0.525	0.0820	0.823	0.1286

从表 1 中可以得出，在服装数据集上，传统的检索方法中 HOG 的效果明显比 HSV 好，而利用深度学习中的卷积神经网络检索的准确率和召回率远远高于传统的方法，这证明了传统的特征提取只能提取图片的浅层信息，不能获取图片更深层次的信息，在卷积神经网络中进行多层特征融合比直接提取深层特征的检索的准确率和召回率高。这是因为浅层中提取服装图片的颜色特征等局部的和具体的信息，在深层提取的特征代表了图片总体的和抽象的特征，进行多层特征融合所得的特征向量具有更强的表达能力和区分性。因此，本文提出的模型最终能得到较好的检索结果。

本文为了说明使用 K-Means 聚类对检索速率产生的影响，将进行 3 组对比实验：使用传统的 CNN 模型 GoogleNet 进行特征提取，然后直接进行检索(GoogleCNN)；使用本文提到的模型进行特征提取，然后直接将提取的向量一一与特征库中的向量进行对比(FusionCNN)；使用本文的模型提取特征，并使用 K-Means 聚类进行检索(FusionCNN＋K-Means)。*mAP* 及检索时间的对比如表 2 所列。FusionCNN 的 mAP 最高，其次是 FusionCNN＋K-Means 方法，*mAP* 最低的是 GoogleCNN 方法。GoogleCNN 与 FusionCNN＋K-Means 的检索速度相似，而 FusionCNN＋K-Means 的检索速度明显快于 FusionCNN。融合多层网络特征导致向量参数增加而增加计量，使得检索速率降低，但是使用 K-Means 进行聚类检索提高了检索速率，所以 GoogleCNN 与 FusionCNN＋K-Means 的检索速率相近，且聚类中心数 K 为 50，相较于不使用聚类检索时速度提高了将近 50 倍。

表 2 不同方法的 *mAP* 及检索时间的对比

算法模型	*mAP*	*Time*/s
GoogleCNN	0.625	0.254
FusionCNN	0.862	13.752
FusionCNN＋K-Means	0.845	0.235

为了说明提取 GoogleNet 不同网络层作为最终的特征向量对检索结果的影响，在不同的迭代次数下训练出模型，并使用这些模型提出不同网络层的特征，比较 mAP，如图 5 所示。从图中可以看出，只提取 Inception Module(1)局部特征作为特征表示的检索性能最差，Inception Module(2)提取的特征检索性能稍好，只将 Inception Module(3)作为特征表示比前两者好稍，将三者融合之后的检索效果最好。同时发现：训练不同的迭代次数对实验的检索效果也有影响，训练次数太少不能达到很好的效果，训练次数过多会产生过拟合，检索效果也不理想。

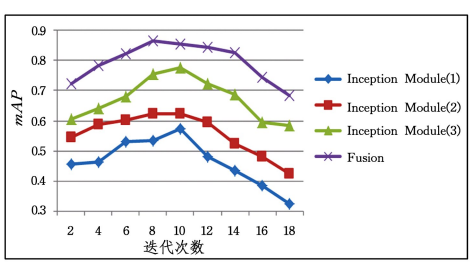

图 5 不同迭代次数训练所得模型提取不同网络层特征的检索 *mAP*

为了说明增加 triplet 结构训练模型对检索结果会产生一定的影响，比较了两种结构在不同聚类数训练方式下检索的 *mAP*，结果如图 6 所示。从图中可以看出，加入 triplet 结构训练网络模型检索的效果明显好于不加 triplet 结构的效果。

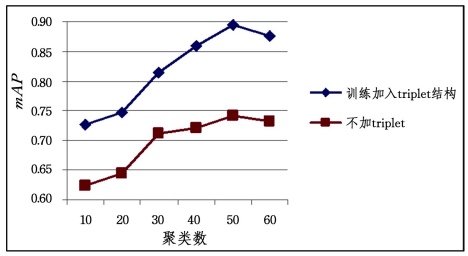

图 6 训练 triplet 结构在不同聚类数下对检索结果的影响

图 7 给出了卷积网络浅层局部特征和深层特征融合后提取特征，然后进行聚类检索的一个结果例图。每行第一张图片为输入图片，为一张检索的原图和其预处理后的图片；后面 10 张图片为图片库中与之最相似的 10 张图片。

图 7 服装图片检索示例图

结束语 本文提出的基于卷积神经网络多层特征融合的服装图像检索模型，通过 CNN 网络模型提取服装图片的浅层局部特征和深层内容特征，将其融合后作为特征表示，然后利用 K-Means 对特征进行聚类检索。实验表明，本文所提算法的 *mAP* 和召回率远远高于传统服装图像特征提取算法的结果，检索速度更快而且效果更好。后续可优化网络模型，在其中加入残差学习[19]的思想，使提取的图片特征更精准；也可考虑加入网页，使得能直接在网页中进行图像检索，从而设计出一个基于 Web 的服装图像检索系统。

参 考 文 献

[1] ALBIOL A，MONZO D，MARTIN A，et al. Face recognition using HOG-EBGM [J]. Pattern Recognition Letters，2008，29(10)：1537-1543.

[2] LO T W R，SIEBERT J P. Local feature extraction and matching on range images：2.5D SIFT[J]. Computer Vision & Image Understanding，2009，113(12)：1235-1250.

[3] ZHOU L，GENG Z，ZHANG J，et al. ORB feature based web pornographic image recognition [J]. Neurocomputing，2016，173(P3)：511-517.

[4] 贾巧丽，王娟，孔兵. 基于形状特征和颜色的服装图像检索[J]. 现代计算机(专业版)，2011(7)：30-32.

[5] 薛培培，邬延辉. 基于图像内容和支持向量机的服装图像检索方法研究[J]. 移动通信，2016(2)：79-82.

[6] 胡玉平，肖行，罗东俊. 基于 GrabCut 改进算法的服装图像检索方法[J]. 计算机科学，2016，43(S2)：242-246.

[7] HINTON G，OSINDERO S. A fast learning algorithm for deep belief nets [J]. Neural Computation，2006，18(7)：1527-1554.

[8] CIRESAN D，MEIER U，SCHMIDHUBER J. Multi-column Deep Neural Networks for Image Classification [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D C，USA：IEEE Press，2012：3642-3649.

[9] CIRSHICK R，DONAHUE J，DARRELL T，et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D C，USA：IEEE Press，2014：580-587.

[10] LIN K，YANG H F，LIU K H，et al. Rapid clothing retrieval via deep learning of binary codes and hierarchical search[C]// Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. ACM，2015：498-502.

[11] KRIZHECSY A，SUTSKEVER I，HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems，2012，25(2)：1097-1105.

[12] KIAPOU M H，HAN X，LAZEBNIK S，et al. Where to Buy It：Matching Street Clothing Photos in Online Shopes[C]// 2015 IEEE International Conference on Computer Vision (ICCV). Santiago：IEEE，2015：3343-3351.

[13] FUKUSHIMA K. Neocognitron：a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position [J]. Biological Cybernetics，1980，36(4)：193-202.

[14] 王利华，邹俊忠，张见，等. 基于深度卷积神经网络的快速图像分类算法[J]. 计算机工程与应用，2017，53(13)：181-188.

[15] 刘海龙，李宝安，吕学强，等. 基于深度卷积神经网络的图像检索算法研究[J]. 计算机应用研究，2017，34(12)：3816-3819.

[16] YIM J，JU J，JUNG H，et al. Image Classification Using Convolutional Neural Networks With Multi-stage Feature [M]// Robot Intelligence Technology and Applications 3. Springer International Publishing，2015.

[17] SZEGEDY C，LIU W，JIA Y，et al. Going deeper with convolutions[C]// 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston：IEEE，2015.

[18] SIMONYAN K，ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science，2014，1(2)：3.

[19] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). IEEE Computer Society，2016：770-778.

[20] 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，2015：815-823.

[21] SURAL S，QIAN G，PRAMANIK S. Segmentation and histogram generation using the HSV color space for image retrieval [C]// International Conference on Image Processing. IEEE，2002：589-592.

[22] DALAL N，TRIGGS B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE，2005：886-893.

[23] CHANG C C，LIN C J. LIBSVM：A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology，2011，2(3)：1-27.

(上接第 214 页)

参 考 文 献

[1] KREVELEN R V，POELMAN R. A Survey of Augmented Reality Technologies，Applications and Limitations[J]. The International Journal of Virtual Reality，2010，9(2)：1-20.

[2] 刘万奎，刘越. 用于增强现实的光照估计研究综述[J]. 计算机辅助设计与图形学学报，2016，28(2)：197-207.

[3] LIU Y L，QIN X Y，XU S H，et al. Light source estimation of outdoor scenes for mixed reality[J]. The Visual Computer，2009，25(5-7)：637-646.

[4] 张锐，钟凡，彭群生，等. 室外场景光照估计的基图像分解算法[J]. 计算机辅助设计与图形学学报，2013，25(4)：442-449.

[5] XING G Y，LIU Y L，QIN X Y，et al. A practical approach for real-time illumination estimation of outdoor videos[J]. Computers & Graphics，2012，36(7)：857-865.

[6] XING G Y，ZHOU X H，PENG Q S，et al. Lighting simulation of augmented outdoor scene based on a legacy photograph[J]. Computer Graphics Forum，2013，32(7)：101-111.

[7] ZHANG R，ZHONG F，LIN L L，et al. Basis image decomposition of outdoor time-lapse videos[J]. The Visual Computer，2013，29(11)：1197-1210.

[8] MARTIN S，POTO ČNIK B. Combinational illumination estimation method based on image-specific PCA filters and support vector regression[J]. Machine Vision & Applications，2017(12)：1-9.

[9] LI B，XIONG W，HU W，et al. Multi-Cue Illumination Estimation via a Tree-Structured Group Joint Sparse Representation [J]. International Journal of Computer Vision，2016，117(1)：21-47.

[10] 付文晓，张锐，林丽丽，等. 雾天室外场景光照参数估计[J]. 软件学报，2014，25(S2)：268-277.

[11] DEBEVEC P E，MALIK J. Recovering high dynamic range radiance maps from photographs[C]// Conference on Computer Graphics and Interactive Techniques. ACM Press/Addison-Wesley Publishing Co，1997：369-378.

[12] YAMAKAWA M，SUGITA Y. Image enhancement using Retinex and image fusion techniques[J]. Electronics & Communications in Japan，2018，101(8)：360-368.

[13] 唐正，刘宏哲，袁家政. 单一光照颜色恒常性计算研究进展[J]. 计算机科学，2016，43(11)：12-18.

[14] 史榕. 自动白平衡算法的研究与实现[J]. 信息技术，2012(3)：85-88.

[15] 武红玉. 阈值分割算法在图像处理中的应用[J]. 科技信息，2012(27)：201-202.

[16] 张锐，韩慧健，梁秀霞，等. 基于色度一致性的室外场景光照参数估计[J]. 计算机科学，2018，45(3)：58-62.