软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(5):1295−1304 [doi: 10.3724/SP.J.1001.2012.04032]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

一种基于稀疏典型性相关分析的图像检索方法*

庄 凌+, 庄越挺, 吴江琴, 叶振超, 吴 飞

(浙江大学 计算机科学与技术学院,浙江 杭州 310027)

Image Retrieval Approach Based on Sparse Canonical Correlation Analysis

ZHUANG Ling+, ZHUANG Yue-Ting, WU Jiang-Qin, YE Zhen-Chao, WU Fei

(College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: E-mail: zhuangling2000@yahoo.com.cn

Zhuang L, Zhuang YT, Wu JQ, Ye ZC, Wu F. Image retrieval approach based on sparse canonical correlation analysis. *Journal of Software*, 2012,23(5):1295−1304. http://www.jos.org.cn/1000-9825/4032.htm

Abstract: A key issue of semantic-based image retrieval is how to bridge the semantic gap between the low-level feature of image and high-level semantics, which can be expressed by means of free text effectively. The cross-modal relationship between the text and image is studied by a modeling semantic correlation between text and image. Based on the model, an approach to image retrieval is proposed so that images are retrieved according to meaning of the query text rather than query keywords. First, an algorithm for solving sparse canonical correlation analysis (CCA) is designed in this paper. Then a semantic space is learned by way of latent semantic analysis from text corpus, and images are represented by bag of visual words. After that, a semantic correlation space, by which the map between visual words of image and the high-level semantics is made explicit, can be constructed. The proposed method solves CCA in a sparse framework in order to make the result more interpretable and stable. The experimental result demonstrates that Sparse CCA outperform CCA in the context, and also substantiates the feasibility of the proposed approach to image retrieval.

Key words: image retrieval; text; semantics; sparse canonical correlation analysis; visual word

摘 要: 图像语义检索的一个关键问题就是要找到图像底层特征与语义之间的关联,由于文本是表达语义的一种有效手段,因此提出通过研究文本与图像两种模态之间关系来构建反映两者间潜在语义关联的有效模型的思路.基于该模型,可使用自然语言形式(文本语句)来表达检索意图,最终检索到相关图像.该模型基于稀疏典型性相关分析(sparse canonical correlation analysis,简称 sparse CCA),按照如下步骤训练得到:首先利用隐语义分析方法构造文本语义空间,然后以视觉词袋(bag of visual words)来表达文本所对应的图像,最后通过 Sparse CCA 算法找到一个语义相关空间,以实现文本语义与图像视觉单词间的映射.使用稀疏的相关性分析方法可以提高模型可解释性和保证检索结果稳定性.实验结果验证了 Sparse CCA 方法的有效性,同时也证实了所提出的图像语义检索方法的可行性.

关键词: 图像检索;文本;语义;稀疏典型性相关分析;视觉单词

中图法分类号: TP391 **文献标识码**: A

* 基金项目: 国家自然科学基金(90920303, 61070068); 中央高校基本科研业务费专项资金(KYJD09015)
收稿时间: 2010-10-11; 定稿时间: 2011-04-02

图像检索一直是近几年来计算机视觉和信息检索领域的研究热点.起源于20世纪90年代的基于内容图像检索(content-based image retrieval,简称CBIR)是早期实现图像检索的一种主流方法,这一方法根据图像视觉特征(如局部特征或全局特征)的相似性计算来达到图像搜索目的.但是,基于内容的图像检索面临语义鸿沟挑战[1,2],即图像底层特征难以反映图像所蕴含的对象、事件和场景等丰富语义.

为了突破CBIR这个固有的瓶颈问题,研究者提出了各种语义图像检索方法,如在图像检索过程中利用相关反馈机制[3−5];Wang等人提出一种距离测度学习方法,在图像视觉特征空间中学习出一种距离测度,可以近似地反映图像间的语义距离[6];QBSE方法通过计算图像间概念概率分布的距离来衡量图像间所表达的语义近似度[7];PAMIR则根据训练集中检索词与图像之间的对应关系学习得到一个排序(rank)模型[8],用于衡量检索词与图像之间的语义相似度;M-OntoMat-Annotizer使用语义Web、本体理论,通过构造概念网络在图像底层特征与语义之间建立一定的关联[9].

基于语义的图像检索也可以通过图像标注来实现.由于为海量的图像进行手工标注是不现实的,因此大量的研究关注于如何使用机器学习的方法来实现图像语义的半自动标注或自动标注.一种方法是定义若干语义类别,将图像标注问题转化为分类问题,可以为每个语义类别训练各自的分类器[10],或训练一个多标签分类器[11];通过建立反映图像底层特征与语义概念之间关联程度的模型也可以进行图像标注,Mori等人使用共生模型[12],CMRM[13],DCMRM[14]等方法则通过学习得到图像视觉特征与语义概念之间的联合概率分布,此即两模态生成模型(bi-modal generative model);最初应用于文本分析的各种隐语义分析方法如LDA等,也被进一步扩展,用于图像语义标注[15−17];Guillaumin等人在距离测度学习的基础上,通过加权最近邻模型得到图像标注[18];Liu等人提出CML方法,利用语义上下文信息进行距离测度学习,并将该方法用于图像语义标注[19].在Web2.0时代,互联网上的图像大多伴随有文本或标签(tag),如截止到2010年7月,Flickr有50亿张图像,对这些图像进行标注的标签词条过亿,涵盖千万种概念.因此,如何利用互联网从图像伴随文本或者标签中选择最佳单词来标注图像语义[20−25],又成为图像检索领域的一个新的热点问题.

无论是使用关键词还是本体标注,目前基本上都是通过若干关键词或概念来表达图像的语义.这样的方式所表达的语义是有限的、不完整的,也忽略了单词之间在语言学(linguistic)上的联系.为了提供更完整的语义表示能力,Rasiwasia等人用概念概率分布表示图像所表达的语义[7].Cheng等人[26]则在构建用于表示图像语义的层次类别结构时考虑了单词间的共生关系.这些努力都试图找到一种更合理的模型来表示图像所蕴含的语义.另外,目前的图像语义标注结果未能体现用户更易于通过自然语言来表达检索意图的事实,现有的主要检索方式还是采用图像或关键字的形式,即由用户输入一幅图片或一组关键字,系统检索出相关的图像.为了克服这一局限性,Zhu等人做了有意义的探索,实现了一个原型系统,用户可以输入一段文本来检索图像.原型系统通过自然言语处理技术提取用户输入的短句中的关键字,然后检索相关的图像组合后提交给用户[27].虽然通过这种方式用户可以表述出更丰富的语义来描述所要检索的图像,但是由于系统最终通过抽取短句中的关键字来检索图像,所以对语义是有损的,本质上还是类似于传统的基于关键词的检索.我们考虑到文本语句是日常生活中人们表达语义最直接的一种载体,具有表达丰富语义的能力,如果能够直接在文本语义空间与图像的底层特征空间之间建立有效的映射关系,那么不仅可以更好地表达图像所蕴含的语义,而且可以很自然地支持以文本语句的形式进行图像检索,这是一个值得关注与研究的问题.

本文在隐性语义索引(latent semantic indexing,简称LSI)和视觉单词(visual word)的基础上,提出了挖掘文本语义空间与图像特征空间这两个异构空间潜在关联,从而构造出语义与视觉单词间映射的思路.由于从图像中提取的特征众多、维数很高,而对于给定图像,其蕴含的语义一般可以通过有限视觉特征来表达,因此本文设计了一种稀疏典型性相关算法(sparse canonical correlation analysis,简称sparse CCA),引入稀疏表达机制来学习视觉特征和语义空间之间的关联.在此基础上,直接采用基于文本语句的图像检索方式,把用户以自然语言表达的检索意图(即以文本语句所表达的检索请求)投影到语义空间中,并通过文本语义和图像特征之间的映射关系进行图像检索.这实际上是将用一种媒体形式(文本语句)表示的语义转换成用另一种媒体形式来表达(图像),所以从本质上说,是一种跨媒体的检索方式.

1 基于邻近算子的稀疏相关性分析算法

1.1 典型相关性分析

典型相关分析(canonical correlation analysis,简称 CCA)是研究维度分别为 n 和 m 的两组随机向量 $x=(x_1,x_2,\ldots,x_n)^T$ 与 $y=(y_1,y_2,\ldots,y_m)^T$ 之间相关关系的一种统计分析方法.它试图找出一对向量 a,b,使得由此构造的线性组合 a^Tx,b^Ty 的相关系数达到最大.可以用下面的数学模型严格地定义它:

设随机向量$(x_1,\ldots,x_n,y_1,\ldots,y_m)^T$ 的协方差阵为 $C=\begin{pmatrix}C_{xx} & C_{xy}\\ C_{yx} & C_{yy}\end{pmatrix}$, 则 $corr(a^Tx,b^Ty)=\dfrac{a^TC_{xy}b}{\sqrt{a^TC_{xx}a}\sqrt{b^TC_{yy}b}}$, 典型相关分析就是要求如下的优化问题的解:

$$\max_{a,b} a^TC_{xy}b,\ \text{s.t. } a^TC_{xx}a=1, b^TC_{yy}b=1 \tag{1}$$

如果将随机向量 x,y 看成是从两个不同的角度去描述同一对象而得到的两组特征,那么相关性分析在统计意义上建立了一个对象的这两组特征间的内在联系.因此,可以用 CCA 对表述相同语义的两种不同模态的媒体数据之间进行潜在语义关联.它可以将两个异构空间上的数据同时变换到一个相关语义空间中,并直接在相关语义空间中度量两种不同模态数据间的距离.文献[28]用 CCA 方法建立起来的单词和图像间的关联进行了图像自动标注,文献[29]则用 CCA 建立了视觉和听觉特征这两种不同模态数据间的相关性映射.

1.2 稀疏典型性相关分析

虽然 CCA 有着广泛的应用,但是和其他的一些统计分析算法(如主成分分析,即 PCA)一样,CCA 算法得到的相关因子是所有特征(变量)的线性组合,这样的结果可解释性差.本文希望通过对训练集进行学习,从原始高维特征集合中选取对表现相关性最有意义特征子集,得到这些特征的线性组合,从而形成具有极大语义相关性的稀疏表达.这不仅可以使结果具有可解释性,也可以剔除噪声变量在相关性分析中的影响,提高模型的稳定性,有效防止出现过拟合(或过学习)的情况.

为了解决这一问题,不同研究者从各自的角度提出了若干 Sparse CCA 方法.比如,文献[30,31]直接在 CCA 的目标函数后添加了惩罚项,扩展了稀疏特征值问题的求解算法用于求解 Sparse CCA 问题,在音乐中所表达的语义与词汇间进行关联,从而抽取出最合适的少量的词汇对音乐进行标注.但该算法将两组特征合并成一个特征集进行特征选择,不能分别为两组特征建立各自的稀疏表达.文献[32]基于 CCA 的概率解释,提出一个算法可自动选择隐变量的维度.但与其他稀疏方法不同,该算法不能选择对表现相关性最有意义的特征,且将特征向量投影到低维空间后,还要在低维空间中学习相关任务参数;文献[33]针对一类原始-对偶数据问题给出了 Sparse CCA 算法,即进行相关性分析时,一方的数据来自原始输入空间,另一方是变换到核空间中的数据.该算法虽然通过数学方法规避了 1-范数求导的问题,但增加了算法的复杂性.

与上述 Sparse CCA 模型不同,为了实现文本语义与视觉单词之间的映射,本文使用原始文本语义矩阵和原始视觉单词矩阵,建立对视觉单词进行单侧稀疏的相关分析模型.针对该模型,本文在文献[33]中提出的 Sparse CCA 算法框架的基础上引入邻近算子(proximity operator),给出了一种更简洁、更易于实现的 Sparse CCA 算法.

设 X,Y 分别是随机向量 x,y 的样本矩阵,其中的每一行是一个样本观测值,且已做了归一化处理,则可以用样本协方差矩阵代替随机向量的总体协方差阵求得相关因子,这时,优化问题(1)可转换成

$$\max_{a,b} a^TX^TYb,\ \text{s.t. } a^TX^TXa=1, b^TY^TYb=1 \tag{2}$$

引理 1. 设 a,b 是优化问题(2)的解,当且仅当存在ζ,γ,使得$\zeta a,\gamma b$ 是优化问题(3)的解:

$$\min_{a,b} \|Xa-Yb\|^2,\ \text{s.t. } \|Yb\|_2^2=1 \tag{3}$$

文献[33]给出了该引理的严格证明.该引理说明优化问题(2)与问题(3)是等价的.

为了能够找到尽可能少的特征子集(即稀疏特征子集),使它们之间的线性组合具有尽可能大的相关性,而不是将过完备(over-complete)特征进行组合,可以在优化问题(3)的目标函数上添加相应的惩罚项,将传统的典

型性相关问题转变成 Sparse CCA 问题,即通过求下面的优化问题来进行相关性学习:

$$\min_{a,b} \| Xa - Yb \|^2 + \mu_1 \| a \|_1 + \mu_2 \| b \|_1, \text{ s.t. } \| Yb \|_2^2 = 1 \tag{4}$$

其中,μ_1,μ_2 是两个用于控制解的稀疏程度的参数.该优化问题在形式上与 L1-范式正则化(即 least absolution shrinkage and selection operator,简称 Lasso)是相似的,但是由于这里 a,b 同时都是优化变量,所以它刻画的问题和 Lasso 就有本质上的区别,同时也需要设计相应的算法来求得问题(4)的解.

1.2.1 邻近算法

本文用邻近算法(proximal algorithm)解 Sparse CCA 问题.邻近算法可用于解形如公式(5)的一类优化问题.

$$\min_{\alpha} f_1(\alpha) + f_2(\alpha) \tag{5}$$

问题的求解依赖邻近算子.邻近算子由 Moreau(1965)引入[34],Combettes 等人对邻近算子和邻近算法进行了详细的分析[35].邻近算法使用邻近算子,通过求不动点问题得到公式(5)的解.机器学习中的一些问题可以归结为形如公式(5)的优化问题并用该算法求解,如 Kowalski 等人就将此算法应用于多核学习问题[36].

定义 1(邻近算子). 设 $\phi: R^n \to R$ 是一个凸的下半连续函数,与 ϕ 相关的邻近算子 $prox_\phi: R^n \to R^n$ 定义为

$$prox_\phi(u) = \arg\min_{\alpha \in R^n} \frac{1}{2} \| u - \alpha \|_2^2 + \phi(\alpha).$$

当优化问题(5)中的 f_1 是凸的下半连续函数,f_2 可微且 ∇f_2 满足 β-Lipschitz 条件,则该优化问题可用下面的算法求最优解[36]:

算法 1. 邻近算法.

initialize: 设置系数 $\gamma < 2/\beta$,为迭代向量 α 赋初始值 $\alpha^{(0)}$

repeat

$$\alpha^{(s+1)} = prox_{\gamma f_1}\left(\alpha^{(s)} - \gamma \nabla_\alpha f_2(\alpha^{(s)})\right)$$

until convergence

1.2.2 基于邻近算子的 Sparse CCA 问题求解

对于 Sparse CCA,本文用上面的邻近算法来求解.但是该算法是用于求解形如公式(5)这样的无约束优化问题的,而本文的 Sparse CCA 问题(4)是带约束条件的:$\| Yb \|_2^2 = 1$,因此,首先要对原问题进行一些必要的变换,使得邻近算法适用于 Sparse CCA 问题的求解.可将问题的约束条件改为 $\|b\|_\infty = 1$,这对于问题的解是没有影响的,因为只相差一个系数.

如果 $\|b\|_\infty = 1$,那么向量 b 的分量中必有一个为 1,其他的小于或等于 1.

我们不妨先固定 $b_k = 1$,对于某个 $1 \leqslant k \leqslant m$,令 $\tilde{b}^{(k)} = (b_1,...,b_{k-1},b_{k+1},...,b_m)^T$,$\tilde{Y}^{(k)} = (y_1,...,y_{k-1},y_{k+1},...,y_m)$,即剔除向量 b 的第 k 个分量和样本矩阵 Y 的第 k 列.这样,使用文献[33]中的算法框架,可以用下面的算法来求解 Sparse CCA 问题:

算法 2. Sparse CCA 算法框架.

Repeat $k=1,\ldots,m$

$$\min_{a,b} \| Xa - \tilde{Y}^{(k)} \tilde{b}^{(k)} - y_k \|^2 + \mu_1 \| a \|_1 + \mu_2 \| \tilde{b}^{(k)} \|_1, \text{ s.t. } \| \tilde{b}^{(k)} \|_\infty \leqslant 1 \tag{6}$$

End

在这 m 个优化问题的最优解中选择最优者,即为原始问题(4)的最优解.

对于问题(6)中的目标函数,有下面的性质:

引理 2. 设 $u_1, a \in R^n, u_2, b \in R^m, u = (u_1^T, u_2^T)^T \in R^{n+m}$,则与 $\mu_1\|a\|_1 + \mu_2\|b\|_1$ 对应的邻近算子定义为

$$prox_{\mu_1\|a\|_1 + \mu_2\|b\|_1}(u) = \arg\min_{a,b} \frac{1}{2} \| u_1 - a \|_2^2 + \frac{1}{2} \| u_2 - b \|_2^2 + \mu_1 \| a \|_1 + \mu_2 \| b \|_1 .$$

可以求得各个分量的值为 $\hat{a}_k = sign(u_{1,k}) \| u_{1,k} | - \mu_1 |_+, \hat{b}_k = sign(u_{2,k}) \| u_{2,k} | - \mu_2 |_+$.

对于目标函数中的第 1 项$\| Xa-\tilde{Y}^{(k)}\tilde{b}^{(k)}-y_k \|^2$,令$H^{(k)}=(X,-\tilde{Y}^{(k)})$, 则

$$\| Xa-\tilde{Y}^{(k)}\tilde{b}^{(k)}-y_k \|^2=(a^T,\tilde{b}^{(k)T})H^{(k)T}H^{(k)}(a^T,\tilde{b}^{(k)T})^T-2(a^T,\tilde{b}^{(k)T})H^{(k)T}y_k+y_k{}^Ty_k.$$

因此,$\| Xa-\tilde{Y}^{(k)}\tilde{b}^{(k)}-y_k \|^2$ 可微,且可以计算其梯度为

$$\nabla \| Xa-\tilde{Y}^{(k)}\tilde{b}^{(k)}-y_k \|^2=2H^{(k)T}H^{(k)}(a^T,\tilde{b}^{(k)T})^T-2H^{(k)T}y_k \tag{7}$$

取$\beta^{(k)}=2\left\|H^{(k)T}H^{(k)}\right\|$, 于是求解算法解优化问题(6)的算法如下:

算法 3. 使用邻近算法求解问题(6).

计算$H^{(k)},\beta^{(k)}$,设置$\gamma<2/\beta^{(k)}$,为迭代向量赋初值$a_{(0)},\tilde{b}_{(0)}^{(k)}$

Repeat

$$\left(a_{(j+1)}{}^T,\tilde{b}_{(j+1)}^{(k)T}\right)^T=prox_{\mu_1\|a\|_1+\mu_2\|b\|_1}\left(\left(a_{(j)}{}^T,\tilde{b}_{(j)}^{(k)T}\right)^T-2\gamma H^{(k)T}H^{(k)}\left(a_{(j)}{}^T,\tilde{b}_{(j)}^{(k)T}\right)^T-2\gamma H^{(k)T}y_k\right)$$

if $\tilde{b}_{(j+1),i}^{(k)T}>1$, set $\tilde{b}_{(j+1),i}^{(k)T}=1$,

else if $\tilde{b}_{(j+1),i}^{(k)T}<-1$, set $\tilde{b}_{(j+1),i}^{(k)T}=-1,\ \forall i,1\leqslant i\leqslant m,i\neq k$

Until convergence

通过对原始样本矩阵 X,Y 进行缩并(deflation)的方法,可以让 Sparse CCA 抽取多组特征[33,37],从而可以学习得到一个多维度的相关空间.具体地,对于每次求得向量 a_i 后,将 X_i 的每一列投影到 X_ia_i 的正交补空间得到新的缩并后的样本矩阵,即

$$X_{i+1}=X_i(I-a_ip_i^T) \tag{8}$$

其中, $p_i=\dfrac{X_i^TX_ia_i}{a_i^TX_i^TX_ia_i}$.对 Y_i 也进行同样的缩并过程.经过缩并后,再用上面的 Sparse CCA 算法对 X_{i+1},Y_{i+1} 学习出一组新的特征组合系数 a_{i+1},b_{i+1}.

为了建立图像视觉特征与文本语义空间之间的潜在关联,本文采用了 Sparse CCA 模型(4)进行语义相关性分析,用 X 表示文本的文本语义矩阵,Y 表示图像视觉单词矩阵,从原始高维特征集合中选取对表现相关性最有意义的特征子集,得到具有极大语义相关性的稀疏表达,具体过程将在第 2 节中详细加以介绍.

2 基于文本语句的图像检索

2.1 隐性语义索引

隐性语义索引(LSI)是对文本进行语义分析的一种经典方法.它是建立在文本的向量表示基础上,在对一个训练文本集合(或语料库)进行综合考量后,构造出一个语义空间.假定在一个文档集里有 d 个文档以及 t 个用户所关心的索引项(可以是一个词或一个短语).设矩阵 D 为 $t\times d$ 矩阵,表示 t 个索引项和 d 个文档之间的关系,其中每一个数值表示该行所代表的那个索引项在该列所代表的那个文档中出现的次数(document frequency).矩阵 D 称为单词-文档(term-document)矩阵.LSI 方法建立在对单词-文档矩阵进行奇异值分解的基础上.

$$D=U\Lambda V^T=\sum\lambda_iu_iv_i^T,$$

其中,U,V 是两个正交矩阵;u_i,v_i 分别是 U,V 的第 i 列;Λ为由 D 的奇异值构成的对角阵,其中的奇异值λ_i 按降序排列.设$(U)_k$ 为 U 的前 k 列向量构成的矩阵,$(\Lambda V^T)_k$ 为ΛV^T 的前 k 行组成的矩阵,由于当 $i>k$ 时,$\lambda_i<<\lambda_1$,则 $D_k=(U)_k(\Lambda V^T)_k$ 是 D 的一个近似表示,由$(U)_k$ 中的 k 个列向量构成的一个 k 维子空间,即为语义空间Ω,$(\Lambda V^T)_k$ 就是 D 在这个 k 维子空间中的投影.其中的一列表示一篇文档在这个语义空间中的投影,可用来作为这篇文档所表达的语义在语义空间Ω中的向量表示.所以,LSI 实际上是通过单词-文档矩阵的奇异值分解确定了一个子空间Ω,对于其他任何文档,我们可以将文档向量投影到该子空间中,由于这个子空间的维度 k 远远小于单词个数,故将文档向量投影到语义空间上,在一定意义上来说是对文档向量进行了维度约减.

2.2 图像的词袋表示

图像的视觉词袋(bag of visual words,简称 BoW)表示最初是在 Sivic[38],Li 等人[39]的论文中提出和使用的,目前已经成为 Web 图像搜索、图像语义建模等领域的主流方法之一.图像的视觉词袋表示是建立在一个码本(codebook)基础上的,对从训练集中每个图像上的所有特征点提取的特征向量进行聚类,将每个聚类中心定义成一个视觉单词,所有的视觉单词构成了一个码本.对将要处理的图像,将其所有特征点映射到视觉单词上就形成了图像的词袋表示.无论从形式上还是语义上,图像的词袋表示和文本的向量模型都是类似的,于是很自然地就可以将文本的分类、隐语义分析等技术用于图像处理上.同时,通过词袋表示可使文本和图像有相对统一的表达,这样,文本与相关图像之间的对应关系就类似于多语言文本间的语义关联.基于这种考虑,本文就试图通过建立图像的视觉单词和文本语义之间的映射关系找到文本与图像之间的内在关联,从而试图通过文本来检索关联的图像.

2.3 基于文本语句的图像检索方法

利用上一节中提出的 Sparse CCA 算法,本文提出一种基于文本语句的图像检索方法.对于给定的一段文本语句,可以根据这段文本的语义从图像库中找出与之相关的图像.整个过程由训练和检索两个过程组成:训练数据是像(图像,文本)这样的二元组构成的集合,训练过程利用 Sparse CCA 算法,根据训练集上文本与图像间的对应关系,学习出一个语义相关空间,以及文本语义空间和图像底层特征空间到语义相关空间的变换;检索阶段可以将给定的检索文本投影到语义相关空间中,然后在语义相关空间中寻找与检索文本邻近的点,这些点就是在语义上与检索文本相关的图像在语义相关空间上的投影.下面具体加以说明.

训练过程:

(1) 将训练集中的文本进行隐语义索引,构造出文本语句的语义子空间,同时计算出训练集中文本的文本语义矩阵 X.

(2) 对训练集中的图像进行底层图像特征提取,并形成图像的视觉单词和码本.根据码本可以得到训练集中每幅图像的词袋表示 y_i,并合成图像视觉单词矩阵 $Y=(y_1,\ldots,y_i,\ldots,y_s)^T$.

(3) 对样本矩阵 X 与 Y 使用 Sparse CCA,建立文本语义与图像特征间的潜在相关性.由于文档已经投影到一个低维的语义空间中,所以在进行语义相关空间学习时,我们只需进行单侧稀疏,即在 Sparse CCA 问题(4)中设定 $\mu_1=0$,由此计算得到 $\{a_i,b_i\}_1^l$,我们就可以通过 $\{a_i,b_i\}_1^l$ 将文本和图像同时映射到同一个 l 维的相关空间中.

检索过程:

(1) 用学习出来的 $\{b_i\}_1^l$ 将图像库中的图像投影到 l 维的语义相关空间中,每张图像变换成该空间中的一个点,构成了图像点的集合 $\{p_i\}_1^s$.

(2) 对于用户通过文本语句所表达的检索意图,先将其文本语句向量投影到语义子空间中,得到一个文档语义向量,然后通过 $\{a_i\}_1^l$ 再将其映射到语义相关空间中的一个点 d.

(3) 在语义相关空间中,在 $\{p_i\}_1^s$ 中查找与 d 距离最相近的 j 个点,将对应的图像提交给用户.

3 实验结果

为了验证本文提出的 Sparse CCA 算法的有效性和基于相关性算法的文本的图像检索方法的可行性,我们在 Window XP 下用 matlab 实现了一个原型系统.原型系统的训练数据和测试数据来自 IAPR TC-12 数据集[40].该数据集包含各种类别的图像 20 000 余幅,每幅图像都用英文和德文给出了关于其语义内容的一段描述(实验只使用了其中的英文文本).本文从中挑选了大海、山、草原、动物、街景、高楼和运动等 15 类图像,在其中选择了英文描述比较详细的图像共 1 100 张构成训练集.另外,又从这 15 类图像中选取了 700 张作为测试集.

在通过 LSI 构造文本语义空间的训练中,本文所选取的文本语义空间的维度,使得这个语义空间至少能够表达 75%以上的在文本集合中所体现的语义,故通过下面的公式来确定维度 k 的值(这里,λ_i 是按降序排列的单

词-文档矩阵的奇异值):

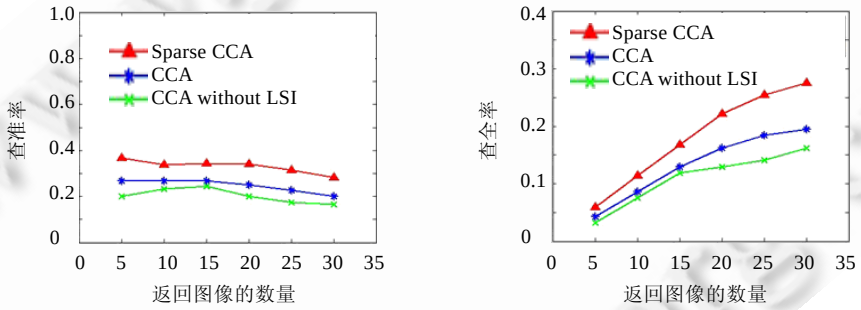
$$\min_k k,\ \text{s.t.}\ \frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i}>75\% \tag{9}$$

对于图像集合,基本上按照文献[39]中的方法生成视觉单词和码本,本文选择使用 SIFT 特征作为图像的局部特征.SIFT 特征算法是由 Lowe[41,42]提出并完善的.它是一种提取图像局部特征的算法.算法首先在 DoG 尺度空间(difference-of-Gaussian scale space)中检测局部极值点作为图像的特征点,然后计算在特征点局部邻域内的梯度方向直方图,形成 m 维的特征向量作为描述子.由于 SIFT 特征对旋转、尺度缩放、亮度变化保持不变性,对视角变化、仿射变换、噪声也具有一定程度的稳定性,因此已被广泛应用于图像匹配等领域.

在实验中,我们首先对训练集中的所有图像提取 SIFT 特征,然后对提取的 SIFT 特征使用 k-means 方法进行聚类,获得的聚类中心作为视觉单词.由于我们使用了稀疏 CCA 算法学习语义相关空间,所以我们生成了 1 000 个视觉单词,这样可以让 Sparse CCA 算法在更大的范围内学习并选择合适的视觉单词.

在 Sparse CCA 模型(4)中,稀疏因子μ_2的选取至关重要,同时也是比较困难的一个问题.它的大小与算法最终选取的视觉单词的数量成反比,也将直接影响实验结果的好坏.在实验中,本文采用文献[33]所使用的办法,最终选择μ_2=0.003.

在实验中,我们设计了 3 种情况下的对比实验:(1) 不进行隐性语义分析,直接在文本向量和图像的视觉单词间进行典型性相关分析(CCA without LSI);(2) 使用 LSI 对文本进性隐性语义分析,在文本语义空间与图像视觉单词间进行典型性相关分析,用的是传统 CCA 方法(CCA);(3) 类似于第 2 种情况,但是用稀疏 CCA 进行相关性分析(Sparse CCA).图 1 列出了这 3 种情况下得到的文本检索的结果,其中,查准率和查全率采用与基于内容的图像检索在性能检测时相同的方法来计算[29].



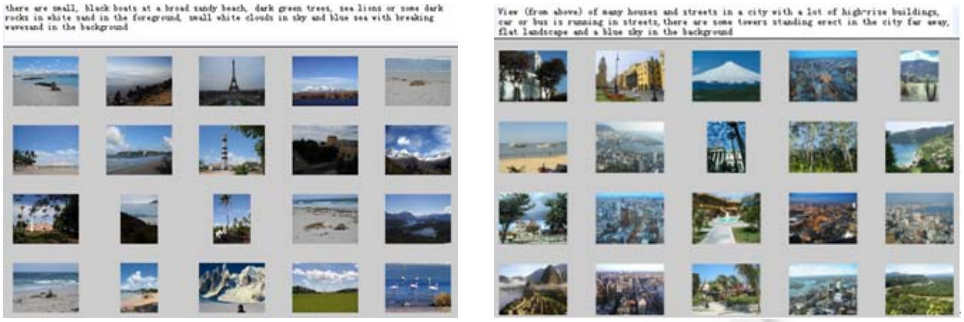
(a) 查准率 (b) 查全率

Fig.1 Comparison of experimental results

图 1 实验结果比较

上面的实验结果是在对每个类别进行文本查询的基础上计算平均值得到的.从实验结果的比对可以看出,CCA without LSI 得到的结果在 3 种情况下是最差的,但与 CCA 相比,结果比较相近;在使用稀疏 CCA 以后,无论是查准率还是查全率都比传统的 CCA 方法有了较大的提高.可见,通过稀疏学习,Sparse CCA 有效地选择一些有意义的特征,提高了模型的稳定性,并在对测试数据的检索时得到了验证.

图 2 是文本语句检索的例子,实例 1 给出的文本语句检索请求是关于沙滩海景的,用户检索意图是希望所返回的检索图像中包含大海、沙滩和天空等对象.从返回的结果看,有 9 幅图像与用户检索意图相吻合,其他图像虽然与大海无关,但有几幅也是以大片的蓝天和白云作为背景.实例 2 给出的文本语句检索请求是有关城市全景图的,用户检索意图是希望所返回图像中有高楼、街道,蓝天等,从返回的结果看,有 7 幅图像与用户所表达的检索意图相符.



(a) 实例 1 (b) 实例 2

Fig.2 Two examples of text retrieval

图 2 文本检索的两个例子

4 结束语

传统的基于内容的图像检索和传统的语义图像检索方法都不能很好地刻画图像所蕴含丰富的语义信息,检索时也不能很好地利用文本所具有的表达完整语义的能力.其核心问题就是如何能够有效地在文本语义空间与图像特征空间之间建立有效的映射关系.这是一个值得研究的问题.

本文设计了一种 Sparse CCA 算法,提出并实现了一种能够基于文本语句进行图像检索的方法,使得用户可以直接用一小段文本来描述他的检索意图,检索系统可以根据查询文本中蕴含的语义检索到相关的图像.这种方法是以 Sparse CCA 算法为基础,该算法能从训练样本中学习出一部分有意义的特征的线性组合构造出一个语义相关空间,将查询文本与图像同时映射到该空间中,并有效地刻画出它们之间的语义相关关系.

进一步的研究还包括如何进一步提高检索精度,如是否可以在图像的视觉词袋表示基础上进行语义分析,然后再使用 Sparse CCA 进行图像文本的语义关联.

References:

[1] Smeulders AWM, Worring M, Santini S, Gupta A, Jain R. Content-Based image retrieval at the end of the early years. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(12):1349−1380. [doi: 10.1109/34.895972]

[2] Wu F, Zhuang YT. Cross media analysis and retrieval on the Web: Theory and algorithm. Journal of Computer-Aided Design and Computer Graphics, 2010,22(1):1−9 (in Chinese with English abstract).

[3] Yoon J, Jayant N. Relevance feedback for semantics based Image retrieval. In: Proc. of the 2001 Int'l Conf. on Image Processing. Thessaloniki, 2001. 42−45. [doi: 10.1109/ICIP.2001.958948]

[4] Ferecatu M, Boujemaa N, Crucianu M. Semantic interactive image retrieval combining visual and conceptual content description. Multimedia Systems, 2008,13(5-6):309−322. [doi: 10.1007/s00530-007-0094-9]

[5] He XF, King O, Ma WY, Li MJ, Zhang HJ. Learning a semantic space from user's relevance feedback for image retrieval. IEEE Trans. on Circuits and Systems for Video Technology, 2003,13(1):39−48. [doi: 10.1109/TCSVT.2002.808087]

[6] Wang CH, Zhang L, Zhang HJ. Learning to reduce the semantic gap in Web image retrieval and annotation. In: Myaeng SY, Oard DW, Sebastiani F, Chua TS, Leong MK, eds. Proc. of the 31st Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Singapore: ACM Press, 2008. 355−362. [doi: 10.1145/1390334.1390396]

[7] Rasiwasia N, Moreno PJ, Vasconcelos N. Bridging the gap: Query by semantic example. IEEE Trans. on Multimedia, 2007,9(5): 923−938. [doi: 10.1109/TMM.2007.900138]

[8] Grangier D, Bengio S. A discriminative kernel-based approach to rank images from text queries. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), 2008,30(8):1371−1384. [doi: 10.1109/TPAMI.2007.70791]

[9] Petridis K, Anastasopoulos D, Saathoff C, Timmermann N, Kompatsiaris Y, Staab S. M-OntoMat-Annotizer: Image annotation linking ontologies and multimedia low-level features. In: Gabrys B, Howlett RJ, Jain LC, eds. Proc. of the 10th Int'l Conf. on Knowledge-Based Intelligent Information and Engineering Systems. Bournemouth: Springer-Verlag, 2006. 633−640. [doi: 10.1007/11893011_80]

[10] Vogel J, Schiele B. Natural scene retrieval based on a semantic modeling step. In: Enser P, *et al*., eds. Proc. of the Image and Video Retrieval. Berlin: Springer-Verlag, 2004. 207−215. [doi: 10.1007/978-3-540-27814-6_27]

[11] Wang M, Zhou XD, Chua TS. Automatic image annotation via local multi-label classification. In: Luo JB, Guan L, Hanjalic A, Kankanhalli MS, Lee I, eds. Proc. of the 7th ACM Int'l Conf. on Image and Video Retrieval. Niagara Falls: ACM Press, 2008. 17−26. [doi: 10.1145/1386352.1386359]

[12] Mori Y, Takahashi H, Oka R. Image-to-Word transformation based on dividing and vector quantizing images with words. In: Proc. of the 1st Int'l Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99). 1999.

[13] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Toronto: ACM Press, 2003. 119−126. [doi: 10.1145/860435.860459]

[14] Liu J, Wang B, Li MJ, Li ZW, Ma WY, Lu HQ, Ma SD. Dual cross-media relevance model for image annotation. In: Lienhart R, Prasad AR, Hanjalic A, Choi S, Bailey BP, Sebe N, eds. Proc. of the 15th Int'l Conf. on Multimedia 2007. Augsburg: ACM Press, 2007. 605−614. [doi: 10.1145/1291233.1291380]

[15] Blei DM, Jordan MI. Modeling annotated data. In: Proc. of the 26th Annual Int'l ACM SIGIR Conf. on Research and Development in Informaion Retrieval. Toronto: ACM Press, 2003. 127−134. [doi: 10.1145/860435.860460]

[16] Monay F, Gatica-Perez D. On image auto-annotation with latent space models. In: Rowe LA, Vin HM, Plagemann T, Shenoy PJ, Smith JR, eds. Proc. of the 11th ACM Int'l Conf. on Multimedia. Berkeley: ACM Press, 2003. 275−278. [doi: 10.1145/957013.957070]

[17] Gao S, Wang DH, Lee CH. Automatic image annotation through multi-topic text categorization. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Toulouse, 2006. II377−II380. [doi: 10.1109/ICASSP.2006.1660358]

[18] Guillaumin M, Mensink T, Verbeek J, Schmid C. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: Proc. of the IEEE Int'l Conf. on Computer Vision. Kyoto, 2009. 309−316. [doi: 10.1109/ICCV.2009.5459266]

[19] Liu ZT, Zhou XD, Xiang Y, Zheng YT. Learning contextual metrics for automatic image annotation. In: Qiu G, Lam KM, Kiya H, Xue XY, Kuo CCJ, Lew MS, eds. Proc. of the Pacific Rim Conf. on Multimedia. Shanghai: Springer-Verlag, 2010. 124−135. [doi: 10.1007/978-3-642-15702-8_12]

[20] Yang HC, Lee CH. Image semantics discovery from Web pages for semantic-based image retrieval using self-organizing maps. Expert Systems with Applications, 2008,34(1):266−279. [doi: 10.1016/j.eswa.2006.09.016]

[21] Feng YS, Lapata M. Automatic image annotation using auxiliary text information. In: Proc. of the 46th Annual Meeting of the Association for Computational Linguistics. Columbus, 2008. 272−280.

[22] Wang XJ, Zhang L, Jing F, Ma WY. AnnoSearch: Image auto-annotation by search. In: Andrew F, Camillo JT, Yann L, eds. Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. New York: IEEE Computer Society, 2006. 1483−1490. [doi: 10.1109/CVPR.2006.58]

[23] Fergus R, Li FF, Perona P, Zisserman A. Learning object categories from Google's image search. In: Proc. of the 10th IEEE Int'l Conf. on Computer Vision. Beijing, 2005. 1816−1823. [doi: 10.1109/ICCV.2005.142]

[24] Lindstaedt SN, Mörzinger R, Sorschag R, Pammer V, Thallinger G. Automatic image annotation using visual content and folksonomies. Multimedia Tools and Applications, 2009,42(1):97−113. [doi: 10.1007/s11042-008-0247-7]

[25] Hsieh LC, Hsu WH. Search-Based automatic image annotation via flickr photos using tag expansion. In: Proc. of the IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing. Dallas: IEEE, 2010. 2398−2401.

[26] Cheng PJ, Chien LF. Personalized image browsing and annotation on the Web using query taxonomy. In: Proc. of the Int'l Conf. on Digital Archive Technologies. Taipei, 2002. 131−140.

[27] Zhu XJ, Goldberg AB, Eldawy M, Dyer CR, Strock B. A text-to-picture synthesis system for augmenting communication. In: Proc. of the 22nd AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2007. 1590−1595.

[28] Hardoon DR, Saunders C, Szedmak S, Shawe-Taylor J. A correlation approach for automatic image annotation. In: Li X, Zaiane OR, Li Z, eds. Proc. of the Advanced Data Mining and Applications. Berlin: Springer-Verlag, 2006. 681−692. [doi: 10.1007/11811305_75]

[29] Zhang H, Wu F, Zhuang YT, Chen JX. Cross-Media retrieval method based on content correlations. Chinese Journal of Computers, 2008,31(5):820−826 (in Chinese with English abstract).

[30] Torres D, Turnbull D, Barrington L, Lanckriet G. Identifying words that are musically meaningful. In: Proc. of the 8th Int'l Conf. on Music Information Retrieval. Vienna, 2007. 405−410.

[31] Torres DA, Turnbull D, Sriperumbudur BK, Barrington L, Lanckriet GRG. Finding musically meaningful words by sparse CCA. In: Proc. of the Neural Information Processing Systems (NIPS) Workshop on Music, the Brain and Cognition. 2007. http://www.cse.ucsd.edu/~datorres/bibs/torres-NIPSWorkshop07.pdf

[32] Rai P, Daumé H III. Multi-Label prediction via sparse infinite CCA, In: Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A, eds. Proc. of the 23rd Annual Conf. on Neural Information Processing Systems. Vancouver, 2009. 1518−1526.

[33] Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. Machine Learning, 2011,83(3):331−353. [doi: 10.1007/s10994-010-5222-7]

[34] Moreau JJ. Proximité et dualité dans un espace hilbertien. Bulletin de la Société Mathématique de France, 1965,93:273−299.

[35] Combettes PL, Pesquet JC. A douglas-rachford splitting approach to nonsmooth convex variational signal recovery. IEEE Journal of Selected Topics in Signal Processing, 2007,1(4):564−574. [doi: 10.1109/JSTSP.2007.910264]

[36] Kowalski M, Szafranski M, Ralaivola L. Multiple indefinite kernel learning with mixed norm regularization. In: Danyluk AP, Bottou L, Littman ML, eds. Proc. of the 26th Annual Int'l Conf. on Machine Learning. Quebec: ACM Press, 2009. 545−552. [doi: 10.1145/1553374.1553445]

[37] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. New York: Cambridge University Press, 2004. 176−192.

[38] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos. In: Proc. of IEEE Int'l Conf. on Computer Vision (ICCV), Vol.2. 2003. 1470−1477.

[39] Li FF, Perona P. California institute of technology: A Bayesian hierarchical model for learning natural scene categories. In: Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 524−531. [doi: 10.1109/CVPR.2005.16]

[40] Grubinger M, Clough P, Müller H, Thomas D. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. In: Proc. of the Int'l Conf. on Language Resources and Evaluation. 2006. http://thomas.deselaers.de/publications/papers/grubinger_lrec06.pdf

[41] Lowe DG. Object recognition from local scale invariant features. In: Proc. of the 7th Int'l Conf. on Computer Vision. Corfu: IEEE, 1999. 1150−1157. [doi: 10.1109/ICCV.1999.790410]

[42] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004,60(2):91−110. [doi: 10.1023/B:VISI.0000029664.99615.94]

附中文参考文献:

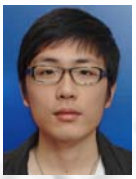
[2] 吴飞,庄越挺.互联网跨媒体分析与检索:理论与算法.计算机辅助设计与图像图形学学报,2010,22(1):1−9.

[29] 张鸿,吴飞,庄越挺,陈建勋.一种基于内容相关性的跨媒体检索方法.计算机学报,2008,31(5):820−826.

庄凌(1970－),男,浙江镇海人,博士,高级工程师,主要研究领域为多媒体分析与检索,多媒体数字图书馆,机器学习.

庄越挺(1965－),男,博士,教授,博士生导师,主要研究领域为多媒体数据库,人工智能,多媒体检索,视频动画.

吴江琴(1965－),女,博士,副教授,主要研究领域为多媒体数字图书馆,数据挖掘,模式识别.



叶振超(1986－)男,硕士,主要研究领域为多媒体数字图书馆.

吴飞(1973－),男,博士,教授,CCF 高级会员,主要研究领域为多媒体分析与检索,机器学习.