Journal of System Simulation

Volume 28 | Issue 10 Article 11

8-13-2020

Multi-View Feature Learning Based on User Contributed Tag

Tian Feng
1. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;;

Fuhua Shang
1. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;;

Zhuoxuan Liu
1. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;;

Xukun Shen
2. The State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China;

Follow this and additional works at: https://dc-china-simulation.researchcommons.org/journal

 Part of the Artificial Intelligence and Robotics Commons, Computer Engineering Commons, Numerical Analysis and Scientific Computing Commons, Operations Research, Systems Engineering and Industrial Engineering Commons, and the Systems Science Commons

This Paper is brought to you for free and open access by Journal of System Simulation. It has been accepted for inclusion in Journal of System Simulation by an authorized editor of Journal of System Simulation.

Multi-View Feature Learning Based on User Contributed Tag

Abstract

Abstract: *A multi-view feature learning method based on user contributed tag was proposed*. Bag-of-words representation for content feature and textual feature was learned. A multi-view feature learning framework was proposed to explicitly model the relevance between multimedia object and tags by learning a linear mapping from textual representation to visual representation. The learned feature encoded the information conveyed by original feature, and inner products of leaned features were preserved with a high probability with visual features and textual features. The complexity of the method is linear with respect to the size of dataset. Furthermore, the method can be extended to deal with more than two views. The performance of the proposed method indicts it's superiority over other representative method.

Keywords

multi-view feature, multi-view learning, user contributed tag, feature learning

Recommended Citation

Tian Feng, Shang Fuhua, Liu Zhuoxuan, Shen Xukun. Multi-View Feature Learning Based on User Contributed Tag[J]. Journal of System Simulation, 2016, 28(10): 2362-2368.

This paper is available in Journal of System Simulation: https://dc-china-simulation.researchcommons.org/journal/vol28/iss10/11

第 28 卷第 10 期 系统仿真学报© Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

基于用户生成标签的多视角特征学习方法

田枫[1]，尚福华[1]，刘卓炫[1]，沈旭昆[2]
（1. 东北石油大学计算机与信息技术学院，大庆 163318；2. 北京航空航天大学虚拟现实技术与系统国家重点实验室，北京 100191）

摘要：*提出了一种基于用户生成标签的多视角特征学习方法。*采用词袋模型分别得到媒体的内容特征表示和标签特征表示；通过媒体词汇和文本词汇的相关性建模，学习文本特征空间和内容特征空间的映射模型。在此基础上，给出了优化前后的特征表示具备近似等距映射保持的理论依据。该方法相对数据集规模具备线性时间复杂度，适用于大规模数据集，具备多视角特征融合能力。基准数据集上测试表明，优化后的特征表示较特征拼接和相关成分分析等方法鉴别力更强。

关键词：多视角特征；多视角学习；用户生成标签；特征学习
中图分类号：TP391 文献标识码：A 文章编号：1004-731X (2016) 10-2362-07

Multi-View Feature Learning Based on User Contributed Tag

Tian Feng[1], *Shang Fuhua*[1], *Liu Zhuoxuan*[1], *Shen Xukun*[2]
(1. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China;
2. The State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China)

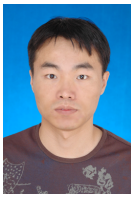
Abstract: *A multi-view feature learning method based on user contributed tag was proposed.* Bag-of-words representation for content feature and textual feature was learned. A multi-view feature learning framework was proposed to explicitly model the relevance between multimedia object and tags by learning a linear mapping from textual representation to visual representation. The learned feature encoded the information conveyed by original feature, and inner products of leaned features were preserved with a high probability with visual features and textual features. The complexity of the method is linear with respect to the size of dataset. Furthermore, the method can be extended to deal with more than two views. The performance of the proposed method indicts it's superiority over other representative method.

Keywords: multi-view feature; multi-view learning; user contributed tag; feature learning

引言

多媒体共享社区等社交网络为代表的新型信息生成和发布渠道的涌现，使得网络多媒体数据急剧增长并累积。如何实现多媒体数据的语义分析和识别一直是研究热点，特征表示是高层语义概念的基本载体，是解决语义鸿沟的关键。针对特定媒体对象，不同底层特征可以表达不同的属性。如图像的颜色特征、纹理特征、形状描述符、形状无关矩等视觉特征构成多个视角；音频的时域特征、频域特征、压缩域特征等底层听觉特征构成多个视角；视频的镜头运动、人脸特征、听觉特征构成多个视角；这些多视角特征共同描述一个媒体对象的内容，彼此是关联和互补的。

在数据挖掘、跨媒体检索等相关领域，一些研



收稿日期：2016-04-28 修回日期：2016-07-14；
基金项目：国家自然科学基金(61502094，61402099)；黑龙江省自然科学基金(F2016002，F2015020)；黑龙江省教育科学规划重点课题(GJB1215019)；
作者简介：田枫(1980-)，男，黑龙江，博士，副教授，研究方向为多媒体理解；尚福华(通讯作者 1962-)，男，黑龙江，博士，教授，研究方向为机器学习。

Published by Journal of System Simulation, 2016

Journal of System Simulation, Vol. 28 [2016], Iss. 10, Art. 11

究者从各自的角度提出了融合异构媒体特征提高应用性能的方法。总体而言有三个思路。一是特征选择，即选择具有区分能力的特征子集[1-10]；二是集成学习，比如异构测度学习[11-20]，可以直接计算异构特征的相似度或者排序值，而无需给出明确的表示，再如多核学习，对于每一个视角选择一个合适的核，然后通过核学习得到一个组合核表示；再者就是将原始异构特征投影到新的特征空间，投影后，异构特征的相关性等价于同一空间内的相似度计算[21-29]。

依据目标准则选择特征子集可以获得具有更强判别力的特征组合[1]。利用多核学习进行特征组合可以取得很好的效果[2-4]。为了解决核权重难以选择的难题，研究人员相继提出了依据最大边缘准则学习权重的监督多核学习[3]和利用未标注数据的半监督多核学习[4]。大多数特征选择方法依赖于已知的明确的语义类别[5]，因此研究人员提出了利用显著性、信息熵等准则的无监督特征选择方法[6-7]。利用组稀疏效应为不同语义标签选择具有判别力的组群特征进行标注在小规模概念集合上是有效的[8-10]。但是，特征选择本质上需要求解组合优化问题，如文献[7]所提出方法的计算复杂度为 $O(n^2)$，n 为样本规模，复杂度过高。

异构测度学习直接计算异构特征的相似度或者排序值，而无需给出明确的表示[11-20]。测度学习的一个问题是，其所需要的三元组(多媒体对象，同类样本，异类样本)不易自动获取，特别是异类样本。利用多视角共现信息获取同类样本是最常用的做法，如果不同媒体对象在一个多媒体文档中共现，它们就具有相同的语义[11-13]。这类方法严重依赖于文档中内容的匹配。异类样本更不宜获取，大多数异类样本的获取都是通过简单标签匹配[14]或者用户反馈[15]。一个思路是通过标签语义关系获得异构相似度[16-17]。文献[16]中，依据两个媒体对象的类别归属获得异构相似度。文献[17]中，研究人员为样本对建立“must-link”和“cannot-link”链接约束，通过文本相关性和内容相关性的组合传播获得异构相似度。文献[18-19]中应用测度学习进行图像标注，也只适用于单一媒体类别。异构媒体间的度量是个研究难点，大多数方法计算代价高昂，如文献[20]中所提出的方法时空复杂度为 $O(n^2)$，n 为样本规模。

另一个异构媒体多视角相关性分析的思路是，将原始异构特征投影到新的特征空间，投影后，异构特征的相关性等价于同一空间内的相似度计算[21-29]。典型相关性分析(CCA)[21]是经典的分析方法，其核心思想是：寻找两个变量场对应的两组基，使得两个变量场在基上投影后的结果最大程度地保持与原变量之间的相关性一致。其视角相关性保持特征已经在理论上得到证明。由于典型相关性分析是建立在两个不同变量场所对应的矩阵基础上，因此也适用于图像，音频与文本等异构媒体特征，利用线性映射，可以学习到一个子空间，最大化异构特征相关性。CCA 被成功应用在融合音视频进行事件检测[22]，说话者人脸识别[23]。但是由于其得到的相关因子是特征的线性组合，结果可解释性差。而且求解过程需要进行广义特征值分解，计算代价高。为了解决这一问题，不同研究人员提出了不同的改进方法，如利用核典型相关性分析(KCCA)融合文本和图像两个模态数据进行聚类[24]。一些研究人员着眼于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。文献[25-26]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稀疏相关性分析问题，将音频语义与词汇进行关联，抽取出最适合的少量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[27]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。但与其他方法不同，该算法

https://dc-china-simulation.researchcommons.org/journal/vol28/iss10/11

第 28 卷第 10 期 系统仿真学报 Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

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[28]针对一类原始-对偶数据问题给出了稀疏相关性分析算法，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。将 CCA 和稀疏特征选择结合，通过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，需要求解稀疏广义特征值分解问题。文献[29]利用邻近算法也无法规避迭代求解的高昂代价。

综上所述，多媒体底层内容特征是高层语义概念的基本载体。但是特征组合与选择、测度学习等方法并不适用于大规模语义概念集合，而“类典型相关性分析”方法和张量互补代表矩阵分解都不能有效回避广义特征值，或者高阶歧义值分解带来的计算复杂问题。图像共享社区 Flickr、视频共享社区 Youtube、三维模型共享社区 Google3DWarehouse、音频共享社区 Findsound 的快速发展引发了社会多媒体数据(Social multimedia)的爆炸式增长。这些社会化多媒体对象具备用户定义的多个文本标签描述，描述了对象、地点、景物、行为等媒体内容，涵盖了丰富的语义概念。

用户生成标签对于多媒体内容的理解具有重要的意义。举例来说，同样代表“体育运动”的游泳和足球，画面颜色分布以蓝(游泳池水)和绿(足球场草地)为主，但是文本特征却可能表达更多的相似性，从而弥补某一视角特征的弱相关性。因此，从网络上社会化多媒体所附带的用户生成标签出发，本文提出了一种基于媒体内容标签的多视角特征学习方法。通过学习从文本特征空间到媒体内容空间的映射，直接建模媒体和文本的相关性，进而得到优化后的特征表示。而且，本文方法相对于数据集 N 的时间复杂度为 $O(N)$。两个基准数据集上的语义标注实验结果表明，优化后的特征表示能够取得比特征拼接、合典型相关性分析等方法更好的效果。第 2 节介绍了多视角特征学习方法，给出了算法复杂度的分析过程，并对方法的等距映射性给出了证明；第 3 节对实验过程做了介绍，并对实验结果进行了相应分析；最后，对本文工作进行了总结和展望。

1 多视角特征学习方法(MVFL)

本文所提出的多视角特征学习方法(Multi-View Feature Learning based on User Contributed Tag，MVFL)，其目标是充分利用媒体所附带的语义标签，融合媒体内容特征表示，进而给出其多视角特征的一致性描述。令 $\boldsymbol{X}\in \mathrm{R}^{N\times M_1}$，$\boldsymbol{Y}\in \mathrm{R}^{N\times M_2}$ 分别表示内容特征矩阵与文本特征矩阵，其中 N 为样本数量，M_1 为内容特征的维度，M_2 为标签数量。特征矩阵 X 与 Y 均采用词袋特征模型(BoW)表示。例如，针对图像数据集，可以提取其 patch 的 Sift 特征；对于视频数据集，可以提取其关键帧的直方图特征；对于音频数据集，提取 MFCC 频率倒谱系数生成听觉词袋；对于三维模型，提取其投影图的词袋特征。本文将构成码书的基元统一定义为媒体词汇，这些基元构成媒体的字典表示。令 $\boldsymbol{U}\in \mathrm{R}^{M_1\times M_2}$ 为特征空间映射矩阵，优化目标如下：

$$\min_{\boldsymbol{O},\boldsymbol{U}} \|\boldsymbol{O}-\boldsymbol{X}\|_F^2 \quad s.t.\ \boldsymbol{O}=\boldsymbol{Y}\boldsymbol{U} \tag{1}$$

其中：$\boldsymbol{O}$ 为优化后的特征矩阵。需要说明的是，如果 $\boldsymbol{X}=\boldsymbol{Y}\boldsymbol{U}$，则 $\boldsymbol{X}_{\cdot i}=\sum_{j=1}^{M_2}\boldsymbol{Y}_{\cdot j}\boldsymbol{U}_{ji}$，这意味着媒体词汇和文本词汇具备线性重构关系，$\boldsymbol{U}_{ji}$ 值越高，意味着相应的两个词汇越相关。利用这种相关性，我们可以得到融合媒体词汇(如视觉词汇)和文本词汇(用户定义标签)两个视角的特征表示。为避免过拟合并保证解稳定性，式(1)改写为

$$\min_{\boldsymbol{O},\boldsymbol{U}} \|\boldsymbol{O}-\boldsymbol{X}\|_F^2+\lambda\|\boldsymbol{U}\|_F^2 \quad s.t.\ \boldsymbol{O}=\boldsymbol{Y}\boldsymbol{U} \tag{2}$$

对上式添加如下约束，以保持特征变换前后特征向量的局部结构

$$\sum_{i,j}\|\boldsymbol{U}_{\cdot i}-\boldsymbol{U}_{\cdot j}\|_2^2\boldsymbol{A}_{ij}=\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{U}^{\mathrm{T}}) \tag{3}$$

Published by Journal of System Simulation, 2016

Journal of System Simulation, Vol. 28 [2016], Iss. 10, Art. 11

其中：$\boldsymbol{A}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$，$A_{ij}$ 为第 i 个视觉特征和第 j 个视觉特征的相似度，$\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{G}^{1/2}\boldsymbol{A}\boldsymbol{G}^{-1/2}$ 为规范化拉普拉斯矩阵($\boldsymbol{L}\in\mathrm{R}^{M_1\times M_2}$)。综合式(2)和式(3)，得到如下优化目标：

$$\begin{aligned}&\min_{\boldsymbol{O},\boldsymbol{U}}\left\|\boldsymbol{O}-\boldsymbol{X}\right\|_F^2+\lambda\left\|\boldsymbol{U}\right\|_F^2+\gamma\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{U}^{\mathrm{T}})\\&s.t.\ \ \boldsymbol{O}=\boldsymbol{Y}\boldsymbol{U}\end{aligned}\tag{4}$$

式(4)是一个受限条件下的优化问题，简化目标表达式为

$$\min_{\boldsymbol{U}}\left\|\boldsymbol{Y}\boldsymbol{U}-\boldsymbol{X}\right\|_F^2+\lambda\left\|\boldsymbol{U}\right\|_F^2+\gamma\mathrm{tr}(\boldsymbol{U}\boldsymbol{L}\boldsymbol{U}^{\mathrm{T}})\tag{5}$$

式(5)是一个凸优化问题。对式(5)的映射矩阵 $\boldsymbol{U}$ 求导，并令其为 0，得到

$$(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}+\lambda\boldsymbol{I}_{M_2})\boldsymbol{U}+\gamma\boldsymbol{U}\boldsymbol{L}-\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}=0\tag{6}$$

其中：$\boldsymbol{I}_{M_2}$ 为 M_2 阶单位阵。上式是控制系统分析中常用的 Sylvester 方程，该类方程是广义的 Lyapunov 方程，解的标准形式如下：

$$\boldsymbol{S}_1\boldsymbol{U}+\boldsymbol{U}\boldsymbol{S}_2+\boldsymbol{S}_3=0\tag{7}$$

其中：$\boldsymbol{S}_1=\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}+\lambda\boldsymbol{I}_{M_2}$，$\boldsymbol{S}_2=\gamma\boldsymbol{L}$，$\boldsymbol{S}_3=-\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}$。式(7)可表示为

$$(\boldsymbol{I}_{M_1}\otimes\boldsymbol{S}_1+\boldsymbol{S}_2^{\mathrm{T}}\otimes\boldsymbol{I}_{M_2})\cdot vec(\boldsymbol{U})=-\mathrm{vec}(\boldsymbol{S}_3)\tag{8}$$

其中：$\otimes$ 表示两个矩阵的 Kronecker 积，vec($\boldsymbol{U}$) 为 $\boldsymbol{U}$ 的列向量化表示。由式(8)可得 $\boldsymbol{U}$ 的列向量化表示

$$\mathrm{vec}(\boldsymbol{U})=-(\boldsymbol{I}_{M_1}\otimes\boldsymbol{S}_1+\boldsymbol{S}_2^{\mathrm{T}}\otimes\boldsymbol{I}_{M_2})^{-1}\cdot\mathrm{vec}(\boldsymbol{S}_3)\tag{9}$$

优化后的特征 $\boldsymbol{O}=\boldsymbol{Y}\boldsymbol{U}$，该特征融合了文本特征 $\boldsymbol{X}$ 和内容特征 $\boldsymbol{Y}$。由于式(6)需要计算 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 和 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{X}$，其复杂度为 $O(NM_1M_2)$，而计算拉普拉斯矩阵 L 所需复杂度为 $O(NM_1{}^2)$，求解式(7)的复杂度为 $O(\max(M_1,M_2)^3)$。可知，MVFL 的复杂度为 $O(NM_1M_2+NM_1{}^2+\max(M_1,M_2)^3)$，其中，$M_1$ 和 M_2 为常量，因此，MVFL 对于数据规模 N 是线性的，适用于大规模数据集合。

MVFL 的学习目标是使得优化后的特征 $\boldsymbol{O}$ 融合内容特征 $\boldsymbol{X}$ 和标签特征 $\boldsymbol{Y}$。由式(1)可知，优化后的特征 $\boldsymbol{O}$ 融合了特征 $\boldsymbol{X}$，而且具备变换前后的等距映射性质。优化后的特征 $\boldsymbol{O}$ 与特征 $\boldsymbol{Y}$ 同样具备等距映射性质。该结论可依据随机投影理论中的等距映射定理[30]证明。等距映射定理如下：

对任意 $p,q\in\mathrm{R}^{M_2}$ ($\|p\|\leqslant1$，$\|q\|\leqslant1$)，存在映射 $f(x)=\frac{1}{\sqrt{M_1}}\boldsymbol{U}\boldsymbol{x}$ ($\boldsymbol{U}\in\mathrm{R}^{M_1\times M_2}$)，其中，$\boldsymbol{U}$ 为高斯随机矩阵或者稀疏随机矩阵，映射前后的数据点保持局部内积不变：

$$\Pr(\left|p\cdot q-f(p)\cdot f(q)\right|\geqslant\varepsilon)\leqslant4\varepsilon^{-(\varepsilon^2-\varepsilon^3)\mathrm{k}/4}\tag{10}$$

等距映射定理表明，高维空间的点集合经过稀疏随机矩阵投影到低维空间后，可以很高的概率保持了两点之间的距离变化不大。下文证明映射矩阵 $\boldsymbol{U}$ 为稀疏随机矩阵，分析如下。由于内容特征 $\boldsymbol{X}$ 和标签特征 $\boldsymbol{Y}$ 是归一化特征，满足定理条件，可令 $p=\boldsymbol{Y}_{i.}$，$q=\boldsymbol{Y}_{j.}$。由于列向量 $\boldsymbol{X}_{.i}$ 与 $\boldsymbol{Y}_{.j}$ 满足标准正态分布，可得

$$\boldsymbol{Y}\boldsymbol{U}_{.i}=\sum_{j=1}^{M_2}\boldsymbol{U}_{ji}\boldsymbol{Y}_{.j}\sim\mathrm{N}(0,\sum_{j=1}^{M_2}\boldsymbol{U}_{ji}^2)\tag{11}$$

另一方面，由式(1)可知

$$\boldsymbol{O}_{.i}=\boldsymbol{Y}\boldsymbol{U}_{.i}\sim\mathrm{N}(0,1)\tag{12}$$

由式(11)和(12)，可知 $\boldsymbol{U}$ 中元素满足 $\sum_{j=1}^{M_2}U_{ji}^2=1$，为列随机矩阵。同时，式(1)的稀疏性约束保证了映射矩阵 $\boldsymbol{U}$ 稀疏矩阵，因此，$\boldsymbol{U}$ 为稀疏随机投影矩阵。因此，由等距映射定理可知，$\boldsymbol{U}$ 为等距映射矩阵。令 $f(p)=O_{i.}$，$f(q)=O_{j.}$，k=1 000，ε=0.1，则等距映射定理保证了下式成立：

$$\Pr(\left|\boldsymbol{Y}_{i.}\cdot\boldsymbol{Y}_{j.}-\boldsymbol{O}_{i.}\cdot\boldsymbol{O}_{j.}\right|\geqslant0.1)\leqslant0.1\tag{13}$$

式(13)说明，优化后的特征 O 的空间与文本特征 Y 同样具备局部结构保持。

相比于类 CCA 方法，MVFLT 求解过程避免了广义特征值分解等复杂计算，其对数据规模呈线性复杂度，使得其可以应用于大规模数据。为了融合多视角特征，MVFLT 优化目标可拓展为

$$\begin{aligned}&\min_{\boldsymbol{O},\boldsymbol{U}_1,\boldsymbol{U}_2}\left\|\boldsymbol{O}-\boldsymbol{X}\right\|_F^2+R(\boldsymbol{U}_1)+\mu_1\left\|\boldsymbol{O}-\boldsymbol{Y}_1\boldsymbol{U}_1\right\|_F^2+\\&R(\boldsymbol{U}_2)+\mu_2\left\|\boldsymbol{O}-\boldsymbol{Y}_2\boldsymbol{U}_2\right\|_F^2\end{aligned}\tag{14}$$

对上述问题求解，可以得到融合更多视角的优化特征表示。

https://dc-china-simulation.researchcommons.org/journal/vol28/iss10/11 4

第 28 卷第 10 期 系统仿真学报 Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

2 实验结果与分析

为了验证 MVFL 特征学习方法的有效性，本文在 2 个基准数据集上进行图像分类测试。Corel 5K 数据集[31]是图像语义分析常用的经典数据集，该数据集包含 50 个目录，每个目录包含 100 幅图像，共计 5 000 幅图像。取 4 500 幅图像作为训练集，500 幅图像作为测试集，平均每幅图像有 3.5 个标签。此外，我们在来自 Flickr 图像共享社区的 NUS-WIDE-Object 数据集[32]上也进行了测试，实验中，我们从 31 个目录中获取了 23 953 幅图像，取 14 270 幅图像作为训练集，9 683 幅图像作为测试集。

对 Corel 5K 数据集，每幅图像提取其 128 维 SIFT 特征[33]，具体的，以特征点为中心取 16×16 的邻域作为采样窗口，最后获得 4×4×8 的 128 维特征描述子，采用 K-MEANS 聚类生成 2 000 个视觉词汇构成的视觉词典，每幅图像采用 BoW 模型生成 2 000 维的视觉特征。对于 NUS-WIDE-Object 数据集，我们直接使用该数据集自带的 500 维 BoW 特征。两个数据集的标签文本特征也采用 BoW 模型生成。

为了测试不同特征表示方法的鉴别力，实验中对上述数据集抽取了多种特征表示集合，包括：视觉特征构成特征集合 V(Visual Feature)；标签文本特征构成特征集合 T(Textual Feature)；将视觉特征与标签文本特征做拼接，并利用 PCA 降维[34]构成的特征集合 PCA(V+T)；对视觉特征与标签文本特征做相关成分分析[21]构成的特征集合 CCA(V+T)；本文所提出的特征集合 MVFL(V+T)。利用线性 SVM[35]对上述特征集合做图像分类测试，实验结果如下：

从表 1 可以看到，文本特征和视觉特征的组合较单一视角具有更高的准确率。在多视角特征融合方法中，采用 MVFL 优化后的特征相比于 PCA(V+T)和 CCA(V+T)具有更高的分类准确率。这是因为 PCA(V+T)只是对特征进行了拼接，虽然降维得到的子空间中保持了视觉特征和文本特征的组成结构，但是没有挖掘多视角特征的潜在关联。CCA(V+T)通过线性映射实现两个视角的特征学习，相比于 PCA(V+T)具有更高的准确率。MVFL 对图像视觉词汇和文本标签词汇的相关性建模，通过文本特征空间和图像特征空间的映射模型，得到的优化特征优于典型相关分析方法(CCA)。实验中，MVFL 的优化目标(式(5))中两个待定参数(λ 和 γ)采用 5 折交叉验证方式确定。

表 1 评测集上的分类准确率 /%

特征	Corel 5K 数据集	NUS-WIDE-Object 数据集
V(Visual Feature)	46.2	31.0
T(Textual Feature)	68.6	68.6
PCA(V+T)	70.4	69.2
CCA(V+T)	70.6	69.5
MVFL(V+T)	71.8	74.7

表 2 记录了 MVFL 和 CCA 的运行时间，可以观察到，由于 CCA 的求解过程需要进行广义特征值分解，其运行时间远大于 MVFL，进一步分析可知，MVFL 对数据集规模的增长呈线性时间复杂度。

表 2 运行时间对比 /s

特征	Corel 5K 数据集	NUS-WIDE-Object 数据集
CCA(V+T)	29.8	48.9
MVFL(V+T)	19.4	12.7

3 结论

从网络上社会化多媒体所附带的用户生成标签出发，本文提出了一种基于文本标签的多视角特征学习方法 MVFL。通过学习从标签文本特征到媒体内容空间的映射，直接建模媒体对象和文本的相关性，得到优化后的特征表示。同时，本文方法相对于数据集规模的时间复杂度为。在两个评测数据集上的图像分类实验结果表明，MVFL 获得的特征表示得比特征拼接、典型相关分析等方法获得的特征鉴别力更强。

Published by Journal of System Simulation, 2016

参考文献:

[1] Guyon I, Elisseeff A. An introduction to variable and feature selection [J]. The Journal of Machine Learning Research (S1532-4435), 2003, 3(9): 1157-1182.

[2] Lanckriet G R G, Cristianini N, Bartlett P, et al. Learning the kernel matrix with semidefinite programming [J]. The Journal of Machine Learning Research (S1532-4435), 2002, 5(1): 323-330.

[3] Vedaldi A, Gulshan V, Varma M, et al. Multiple kernels for object detection [C]// Computer Vision, 2009 IEEE 12th International Conference on. USA: IEEE, 2009: 606-613.

[4] Wang S, Jiang S, Huang Q, et al. S3MKL: scalable semi-supervised multiple kernel learning for image data mining [C]// Proceedings of the international conference on Multimedia. USA: ACM, 2010: 163-172.

[5] Xu Z, King I, Lyu M R T, et al. Discriminative semi-supervised feature selection via manifold regularization [J]. Neural Networks, IEEE Transactions on (S1045-9227), 2010, 21(7): 1033-1047.

[6] Xing E P, Karp R M. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts [J]. Bioinformatics (S1367-4803), 2001, 17(S1): 306-315.

[7] Cai D, Zhang C, He X. Unsupervised feature selection for multi-cluster data [C]// Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery And Data Mining. USA: ACM, 2010: 333-342.

[8] Zhang S T, Huang J Z, Huang Y C, et al. Automatic image annotation using group sparsity [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, USA: IEEE Computer Society, 2010: 3312-3319.

[9] Wu F, Han Y H, Tian Q, et al. Multi-label Boosting for image annotation by structural grouping sparsity [C]// Proceedings of the ACM Multimedia International Conference. New York, USA: ACM, 2010: 15-24.

[10] Yang Y, Yang Y, Huang Z, et al. Tag localization with spatial correlations and joint group sparsity [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE Computer Society, 2011: 881-888.

[11] Yang Y, Xu D, Nie F, et al. Ranking with local regression and global alignment for cross media retrieval [C]// Proceedings of the 17th ACM international conference on Multimedia. USA: ACM, 2009: 175-184.

[12] Zhuang Y T, Yang Y, Wu F. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval [J]. Multimedia, IEEE Transactions on (S1520-9210), 2008, 10(2): 221-229.

[13] Yang Y, Zhuang Y T, Wu F, et al. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval [J]. Multimedia, IEEE Transactions on (S1520-9210), 2008, 10(3): 437-446.

[14] Wu P, Hoi S C H, Xia H, et al. Online multimodal deep similarity learning with application to image retrieval [C]// Proceedings of the 21st ACM International Conference on Multimedia. USA: ACM, 2013: 153-162.

[15] Xia H, Wu P, Hoi S C H. Online multi-modal distance learning for scalable multimedia retrieval [C]// Proceedings of the sixth ACM international conference on Web search and data mining. USA: ACM, 2013: 455-464.

[16] Zhai X, Peng Y, Xiao J. Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval [M]. Germany: Springer Berlin Heidelberg, 2012: 312-322.

[17] Zhai X, Peng Y, Xiao J. Cross-modality correlation propagation for cross-media retrieval [C]// Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. USA: IEEE, 2012: 2337-2340.

[18] Guillaumin M, Mensink T, Verbeek J, et al. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation [C]// Proceedings of Intel Conference on Computer Vision. Piscataway, NJ, USA: Institute of Electrical and Electronics Engineers Inc, 2009: 309-316.

[19] Verbeek J, Guillaumin M, Mensink T, et al. Image annotation with TagProp on the MIRFLICKR set [C]// Proceedings of the ACM SIGMM Intel Conference on Multimedia Information Retrieval. New York, USA: ACM, 2010: 537-546.

[20] Li Z, Liu J, Jiang Y, et al. Low rank metric learning for social image retrieval [C]// Proceedings of the 20th ACM international conference on Multimedia. USA: ACM, 2012: 853-856.

[21] Hotelling H. Relations Between Two Sets of Variates [M]. Breakthroughs in Statistics. USA: Springer New York, 1992: 321-377.

[22] Kidron E, Schechner Y Y, Elad M. Pixels that sound [C]// Computer Vision and Pattern Recognition, IEEE

第 28 卷第 10 期 系统仿真学报 Vol. 28 No. 10
2016 年 10 月 Journal of System Simulation Oct., 2016

Computer Society Conference on. USA: IEEE, 2005, 1: 88-95.

[23] Bredin H, Chollet G. Audio-visual speech synchrony measure for talking-face identity verification [C]// Acoustics, Speech and Signal Processing, IEEE International Conference on. USA: IEEE, 2007, 2: 233-236.

[24] Blaschko M B, Lampert C H. Correlational spectral clustering [C]// Computer Vision and Pattern Recognition, IEEE Conference on. USA: IEEE, 2008: 1-8.

[25] Torres D A, Turnbull D, Barrington L, et al. Identifying Words that are Musically Meaningful [C]// International Conference on Music Information Retrieval. Sydney: Austrian Computer Society, 2007, 7: 405-410.

[26] Torres D A, Turnbull D, Sriperumbudur B K, et al. Finding musically meaningful words by sparse CCA [C]// Neural Information Processing Systems (NIPS) Workshop on Music, the Brain and Cognition, Germany: Springer Berlin Heidelberg, 2007: 1-8.

[27] Rai P, Daume H. Multi-label prediction via sparse infinite CCA [C]// Advances in Neural Information Processing Systems, Germany: Springer Berlin Heidelberg, 2009: 1518-1526.

[28] Hardoon D R, Shawe-Taylor J. Sparse canonical correlation analysis [J]. Machine Learning (S0885-6125), 2011, 83(3): 331-353.

[29] 庄凌, 庄越挺, 吴江琴, 等. 一种基于稀疏典型性相关分析的图像检索方法 [J]. Journal of Software (S1000-9825), 2012, 23(5): 1295-1304.

[30] Vempala S S. The random projection method [M]. American: American Mathematical Society, 2005: 312-322.

[31] Duygulu P, Barnard K, Freitas JFG, et al. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary [C]// Proceedings of the European Conference on Computer Vision. Berlin, Germany: Springer Press, 2002: 97-112.

[32] Chua T S, Tang J H, Hong R H. NUS-WIDE: A real-world web image database from national university of singapore [C]// Proceedings of ACM Conference on Image and Video Retrieval. New York, USA: ACM, 2009: 1-9.

[33] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision (S0920-5691), 2004, 60(2): 91-110.

[34] Hotelling H. Analysis of a complex of statistical variables into principal components [J]. Journal of Educational Psychology (S2044-8279), 1933, 24(6): 417-441.

[35] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification [J]. The Journal of Machine Learning Research (S1532-4435), 2008, 9(9): 1871-1874.

(上接第 2361 页)

[6] 吴海彬, 林宜. 基于改进 Voronoi 图的移动机器人在线路径规划 [J]. 中国工程机械学报, 2007, 5(1): 117-121.

[7] 闫浩文, 郭仁忠. 用 Voronoi 图描述空间方向关系的理论依据 [J]. 武汉大学学报信息科学版, 2002, 27(3): 306-310.

[8] Hoff III K, Culver T, Keyser J, et al. Fast computation of generalized Voronoi diagrams using graphics hardware [C]// Proceedings of the 26th annual conference on computer graphics and interactive techniques (SIGGRAPH '99). New York, USA: ACM Press, 1999: 277-286.

[9] 方巍. 多层动态 Voronoi 图的研究与应用 [D]. 北京: 中国科学院研究生院, 2012.

[10] Sud A, Andersen E, Curtis S, et al. Real-Time path planning in dynamic virtual environments using multiagent navigation graphs [J]. IEEE Transactions on Visualization & Computer Graphics (S1077-2626), 2008, 14(3): 526-538.

[11] 卢瑾, 杨东勇. 基于双重遗传算法机制的路径规划 [J]. 系统仿真学报, 2008, 20(8): 2048-2051.