跨模态社交图像聚类

赵其鲁 李宗民
(中国石油大学(华东)计算机与通信工程学院 山东 青岛 266580)

摘 要 社交图像包含两种模态的信息:视觉信息和社交标签信息.绝大部分跨模态学习领域的研究者,将其精力集中在多模态信息的共享特征空间学习上,从而往往忽略了各模态信息所独有的特征.在该文中将探究如何利用二者的共享信息以及独有信息进行跨模态的图像聚类.该文将共享特征空间的学习看作一个共轭词典学习问题(Coupled Dictionary Learning,CDL),通过一个 $L_{1,\infty}$ 范数的正则项使各模态的词典稀疏化,这种结构化的稀疏性限制会使各模态独有的特征得以保留.除此之外,该文还提出了一个简单的语义相似度度量框架.借助一个包含丰富语义关系的信息库 WordNet,该文通过度量标签间的概念距离(conceptual distance)与释义相似度(gloss similarity),为标签添加一定的语义关系,以度量样本间的语义相似度.通过实验证明该文"共享&独有"模式的跨模态学习的方法,相比其它只利用共享特征的方法,在聚类任务上表现更为出色.

关键词 跨模态学习;共轭词典学习;WordNet;图像聚类;社交图像;语义相似度度量
中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2018.00098

《计算机学报》

Cross-Modal Social Image Clustering

ZHAO Qi-Lu LI Zong-Min
(*Computer Applications Technology Department, China University of Petroleum (East of China), Qingdao, Shandong* 266580)

Abstract With the growth of industrial demands, cross modal learning has gradually attracted more and more attention. Due to the popularity of social media websites, people can tag social images according to their social or cultural backgrounds, personal expertise and perception. With the exponential growth of tagged social images, it has become increasingly attractive to develop new algorithms for achieving more effective organization and summarization of large-scale social images. In general, social images contain two modalities of information: visual information and keyword information. Combining them may lead to a comprehensive description of the social images. However, most researchers on cross-modal learning focus attention on the shared latent space learning, and ignore the private information of each modality. In this paper, we leverage a novel approach to find a latent space in which the information is correctly factorized into shared and private parts. First, we consider the latent space learning as the coupled dictionary learning problem, which can generates homogeneous dictionaries for different modalities by associating and jointly updating their shared coefficients. Second, we add structured sparseness constraints on the dictionaries to allow a latent dimension to be associated with a single modality. Specifically, for each modality's dictionary matrix we add a $L_{1,\infty}$ norm regularizer to encourage some dictionary entries to be zeroed-out. By imposing such structured sparseness constraints, some latent dimensions would be explained by one modality rather than by both the modalities only. We leverage an optimization method which optimizes the objective function with respect to the dictionary matrices

收稿日期:2016-07-03;在线出版日期:2017-06-12.本课题得到国家自然科学基金(61379106)、山东省中青年科学家奖励基金(BS2010DX037)、山东省自然科学基金(ZR2009GL014,ZR2013FM036,ZR2015FM011)、浙江大学 CAD&CG 国家重点实验室开放课题(A1315)资助.
赵其鲁,男,1987年生,博士研究生,主要研究方向为计算机视觉和机器学习.E-mail:kg19872006@163.com.**李宗民**(通信作者),男,1965年生,博士,教授,中国计算机学会(CCF)会员,主要研究领域为图像处理、模式识别和计算机图形学.E-mail:lizongmin@upc.edu.cn.

and the shared coefficients matrix alternately. For the sub-problems involving dictionary matrices, we leverage an efficient optimization algorithm based on the composite gradient mapping method which has been proved to converge very fast. For the sub-problem of the shared coefficients matrix, a multiplicative update algorithm is used. In addition, it's important to extract sufficient semantic relations from a limited number of social keywords. To this end, basing on an extra lexical database (such as WordNet) that contains sufficient semantic relationships, we propose a framework for semantic similarity measurement. First, a common sense determination algorithm is used to detect the common sense for each keyword. Then, we compute the semantic similarities between social keywords through the measurement of conceptual distance and gloss similarity between the common senses. Finally, the image-level semantic similarities are computed to describe the semantic relations among the social images, which construct the semantic feature matrix feeding to the cross-modal learning algorithms tested in this paper. In the experiments, two real-world datasets were employed for quantitatively testing the performance of "shared&private" approach (S&P) on social image clustering task. In order to show the effectiveness of S&P, we compared it with four baseline methods, including the Canonical Correlation Analysis algorithm widely used in many cross-modal learning tasks as a workhorse tool. Through the experiments, we demonstrate that the S&P approach achieves better performance than the baselines. Besides, we also investigated the influence of S&P's parameters on its performance by varying one parameter at a time while fixing the other. This investigation can guide the practical applications in industries.

Keywords　cross-modal learning; coupled dictionary learning; WordNet; image clustering; social images; semantic similarity measurement

《计算机学报》

1　引　言

近几年来，随着数据的积累和工业需求的增长，跨模态数据处理逐渐受到越来越多的关注，尤其是跨模态图像检索. 跨模态图像检索采取的检索方式是"文字检索图像"或者"图像检索文字"，有较广泛的工业应用需求，比如给出一段语言描述，找出一组能从视觉上很好解释这段描述的图像集合. 目前在跨模态图像检索领域主要有两种方法：一种是假设多模态数据之间存在一个共享的子空间，比如典型相关分析算法(CCA)[1]和多模态词典学习(multi-modal dictionary learning)[2-4]；另外一种方法是假设多模态数据之间存在一些共享的语义主题，比如 correspondence Latent Dirichlet Allocation(corr-LDA)[5]. 无论是哪种方法，其目的一致，就是由多模态数据学习出一种描述力较强、可泛化的跨模态特征.

相对于跨模态图像检索，跨模态图像聚类显然受到的关注较少. 从本质上讲，此二者差别不大，核心问题都是学习一个跨模态的特征. 部分跨模态检索方面的算法可以直接迁移到跨模态图像聚类任务中，比如 CCA 和 corr-LDA，而多模态词典学习则无法直接应用到跨模态聚类之中，因其侧重于不同模态特征之间的转换，而非跨模态特征的学习.

在跨模态学习领域，大多数研究人员将精力集中于多模态信息的共享特征空间学习上，往往忽略了各模态信息本身所特有的特征，这显然并不符合实际应用情况. 如何学习一个"共享&独有"模式的跨模态特征，是本文研究的核心内容.

首先，我们仍然从如何学习共享特征开始思考. 近几年来，多模态词典学习逐渐在跨模态学习领域成为主流(文献[2-4]和文献[6-8]，机器学习领域中，称跨模态学习为多角度学习，multi-view learning)，其主要思想可概括为：通过学习一系列对齐的词典来获取各个模态之间的相关关系，我们统称此类方法为共轭词典学习(coupled dictionary learning). 受此启发，我们借助稀疏编码技术(sparse coding)来同时学习各模态的词典，并使其共享相同的重构系数来进行词典对齐. 这么做背后的逻辑是：重构系数相当于信息在词空间的表达，重构系数相同则表达的信息相同，即共享信息. 各模态中的词典可以视为同一词空间在不同的模态中所具有的不同的表现形

式,即质同形不同.

在学习共享信息的同时,我们希望能够保留各模态独有的信息.为了降低算法的复杂度,我们希望共享信息与独有信息在同一过程中被学习.我们的目标是将独有信息的特征量化过程融入至上面所设计的共享信息的特征量化过程中.假设我们有 A 和 B 两种模态的数据,A 中存在某一独有的信息,则通过词典学习我们可以使用某个词及相应的重构系数来表达此信息,但是 B 中并不包含此信息,所以我们要设计一个方法使此重构系数对 B 中的信息重构无效.因为此信息存在于 A 中,故对应的重构系数不可能为零,所以只能使 B 中对应的词为零,也就是和传统的稀疏编码技术在重构系数上添加稀疏性不同,在本文中,我们通过在词典上添加结构化的稀疏性限制来学习各模态独有的特征.在词典上添加稀疏性已有先例,比如文献[6].依据文献[6,8]的经验,我们选择 $L_{1,\infty}$ 范数的正则项来使词典稀疏化.

社交图像指用户在社交类网站上所上传的图像,比如人人网、朋友圈和脸书(Facebook)等.用户在上传图像的同时,可以对图像进行标注,我们称其为社交标签(social tags,以下简称标签).所以,社交图像包含两种模态的信息:标签信息和视觉信息.标签中含有丰富的语义信息,对社交图像的处理与组织极为重要.相应地,如何从标签中提取足够的语义信息以描述社交图像间的语义相似度,是跨模态社交图像处理中非常重要的一个问题,也是本文研究的另外一个问题.

在之前的许多文献中,经常使用词包特征[9]表征标签中的信息,比如文献[3-4]等.词包特征统计了标签的出现频率,但缺乏标签间的语义关系信息,而且标签的数量较少,也限制了词空间的完备性.在文献[10-11]中,Latent Dirichlet Allocation(LDA)被用于提取文本的语义主题信息,有不错的效果.当应用于社交图像处理时,LDA 同词包特征类似,同样受制于标签的数量,使其很难从有限的标签中提取出足够的语义主题信息.

社交标签间具有丰富的语义关系,比如"城堡"和"建筑".城堡是建筑的一种,二者具有语义上的抽象-具体关系,而传统的词包特征会忽略这种关系.我们认为,在跨模态的社交图像处理中,丰富的语义关系能够带来丰富的相关性.为了获得丰富的语义关系,我们需要一个蕴含丰富语义关系的辅助数据库.WordNet[12]是一个巨大的电子词典,同一词义的单词组成一个单位,称之为 synset, synset 按照其释义间的语义关系组织成一个巨大的网络,这点使 WordNet 成为了一个良好的语义关系库.首先,synset 间的概念关系(抽象-具体、整体-部分等等)可以度量标签间的概念距离,比如,相对于"鸡蛋","建筑"同"城堡"在概念上更为相似,在 WordNet 的网络结构中,"建筑"与"城堡"间的节点数少于"建筑"同"鸡蛋"间的节点数,也就是距离上更为接近.其次,每个 synset 的释义,也为我们提供了非常好的语料资源,来度量两个标签的相似度.语义关系较近的单词,其释义中往往会有一定比例的重复用词,通过一定的方法来度量其重复度,便可计算出两个标签间的语义相似度.

用户在添加标签时,有可能使用多义词,也就是说,一个标签有可能对应多个 WordNet 中的 synset,但用户其实只是使用其中的一个来描述图像.如果我们使用标签所对应的全部 synset 来度量它与其他标签的相似度,那么准确性势必会下降.通过观察,我们发现用户往往倾向于使用单词最常见的词义来描述图像.鉴于此,我们设计了一个通用词义检测算法,为每个多义词标签,选出两个最为常见的词义,以度量标签间的语义相似度.

下面,我们总结一下这篇文章的两个主要贡献:

(1)我们使用一种"共享&独有"的学习方法来学习跨模态特征.

(2)我们提出了一种简单的语义相似度度量框架,对社交标签之间的语义关系进行挖掘.

2 相关工作

随着工业需求的增长,跨模态学习逐渐引起研究人员的注意.早期,很多工作使用一些较直接的方法来结合不同模态的特征.比如,Cai 等人在文献[13]中,通过一个页面分割算法将网页分割成很多块,块中的图像和文本信息具有一定的相关性,或者说"连接".页面间超链接将块与其他页面联系起来,形成另外一种连接.通过这些连接信息以及块的空间信息,使用简单的概率估计方法,可以计算出图像与图像间的连接权重图,再结合图像的视觉信息计算出的视觉图与文本信息计算出的文本图,使用一个三级的图像聚类算法,可以对搜索引擎的结果进行可视化.Wang 等人在文献[14]中同样关注网页图像的处理,并使用与文献[13]相同的页面分析技术,建立网页图像与其周围文本之间的联系.一个较

简单的相似度传播算法被用于调节样本间的相似度. Rege 等人[15]和 Gao 等人[16]中都注意到了一种名为偶图的工具[17]，他们在偶图的基础上进行扩展，得到一种三元偶图，并对相关计算方法进行了完善.

目前跨模态学习领域主要的研究方法为两类：以 Blei 和 Jordan[5]的 Correspondence LDA 为代表的主题学习方法和以典型相关分析[1]为代表的共享空间学习方法. 下面，我们对这两类方法进行阐述.

基于统计学习的方法在结合多模态特征方面变现不俗，出现了很多优秀的文章，比如文献[5,18-20]等等. Barnard 和 Forsyth 等人领导的视觉研究组，有一个名为“WORDS and PICTURES”的项目，研究主旨为：通过多种媒体数据间的沟通进行图像理解(Image understanding as multi-media translation). 在文献[18,20]中，Barnard 使用一个层次生成模型来结合文本特征和视觉特征. 这个模型最早由 Hofmann[19]提出，是一种基于词频统计的概率模型，主要用于文档聚类. Kobus 将图像进行分割，每一块图像区域可类比于一个单词(文章中称之为“blobs”)，采用与文本单词不同的分布，他实现了对 Hofmann 模型的扩展. 在文献[5]中，Blei 基于其著名的 LDA 模型，提出了 Gaussian-Multinomial LDA 及其扩展 Correspondence LDA. GM-LDA 同样将图像视为区域的集合，并认为区域类似于单词，隐含了某些主题，简言之，Blei 将一幅图像等同于一篇包含若干主题的文档. Blei 和 Kobus 的方法不仅在思想上有相似之处，具体算法上也有相似之处，比如将图像分割成区域以类比单词. 二人在文献[21]中进行了合作，将这些方法进行了对比. Blei 的方法通过主题学习，生成一个主题向量，可以视之为结合了文本与视觉信息的一种跨模态表征向量，这极大的扩展了其应用面，任何跨模态学习问题都可以使用这个方法，而 Kobus 的方法则局限于聚类任务.

Rasiwasia 等人在文献[11]中提出了基于相关性分析(correlation analysis)和摘要分析(abstraction analysis)的跨模态学习方法. 据我们所知，这篇文章第一次提出了“*image-query-text*”和“*text-query-image*”等真正意义上的跨模态检索. Rasiwasia 等人在文献[11]中假设文本特征空间和视觉特征空间之间存在一个双方共享的隐式子空间. 他使用典型相关分析算法(CCA)学习两组映射向量，并将文本特征和视觉特征映射到一个共享的子空间，在此特征空间内，可以使用任意距离度量方法来计算样本间的相似度. 在实验中，相关性分析显示了非常好的效果. 摘要分析(abstraction analysis)是另外一种结合文本特征和视觉特征的方法，Nikhil 使用一些较为常见或者随机选取的类别，为每个类别训练两个分类器：一个基于文本特征，另一个基于视觉特征. 虽然使用的特征不同，但分类器的输出在本质上是相同的，所以可以视为同一特征空间. 共享的核信息嵌入(shared Kernel Information Embedding)[22]、共享的基于高斯过程的隐式变量模型(the shared Gaussian Process Latent Variable Model)[23-25]等方法使用不同的技术寻找多模态信息间的共享特征，与上述方法称得上殊途同归，所以我们将这些方法视为同一类方法.

多模态词典学习方法(multi-modal dictionary learning)[2-4]遵循同样的思想，假设不同模态的特征之间隐式地存在共享特征空间. 在文献[4]中，通过控制各个词典的稀疏性，使所有的重构系数共享相同的稀疏性，可以学习到多个模态的词典，并且使这些词典之间具有一定的对应关系，这类方法可以统称为共轭词典学习(coupled dictionary learning). 在文献[2]中，共轭词典学习则是通过设计一个连接方程来实现. 文献[3]学习共轭词典的思路同前两篇文献非常相似，不同点在于其借助共享的标签空间来控制词典的对应关系，也就是共享的特征空间. 多模态词典学习方法大多针对跨模态的图像检索，可以将不同模态的特征互相转换，往往不区分共享特征和各模态独有特征，所以不易用于跨模态的图像聚类之中.

上述方法，多侧重于多模态数据的共享特征学习，往往会忽略各个模态独有的信息. 如何全面的利用多模态数据中的共享特征与独有特征进行社交数据处理，据我们所知，目前仍处于空白状态，也是本文与上述其他方法最大的不同. 下面我们介绍一下与本文方法关系较为密切的一些工作. 文献[26]提出了一个非冗余的隐式空间分解方法，在将隐式空间分解为共享和独有两部分的同时，去优化隐式空间的维度，并且在人体姿态估计上取得了不错的效果. 但是，文献[26]的优化过程较为复杂，代价较高. 文献[6]提出了一个有效的分解方法，并避免了文献[26]中过度复杂的计算过程. 据我们所知，文献[6]是第一次在词典上添加稀疏性限制，使其学习到的特征可以是任意几个角度间共享的特征，而非一定是所有角度共享的特征. 本文的工作也可以视为文献[6]的一个变种，与文献[6]不同的是，我们设计的重构系数稀疏项并非 $L_{1,\infty}$ 范数的正则项，我们

使用的优化方法也与文献[6]中使用的方法不同.文献[8]同样借鉴了文献[6]的工作,并提出了更为有效的优化方法,我们在目标函数的优化上主要借鉴自文献[8].文献[8]使用了半监督的训练数据以添加一定量的语义信息,而本文属于无监督学习.

3 跨模态特征学习

在这部分,我们首先介绍如何学习共享的隐式特征空间,然后给出整个跨模态特征学习的公式.我们使用 $\boldsymbol{X}^v$ 和 $\boldsymbol{X}^s$ 分表表示视觉特征和社交标签特征,其中 $\boldsymbol{X}^v\in R^{P_v\times N}$,$\boldsymbol{X}^s\in R^{P_s\times N}$.$\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$ 为各模态对应的词典,$\boldsymbol{D}^v\in R^{P_v\times K}$,$\boldsymbol{D}^s\in R^{P_s\times K}$.$N$ 为样本数量,K 为词典长度.重构系数矩阵我们用 $\boldsymbol{\alpha}$ 表示,$\boldsymbol{\alpha}\in R^{K\times N}$.跨模态的共享隐式特征空间可以用下面的公式进行学习:

$$\min_{\boldsymbol{D}^v,\boldsymbol{D}^s,\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{X}^v-\boldsymbol{D}^v\boldsymbol{\alpha}\|_F^2+\frac{1}{2}\|\boldsymbol{X}^s-\boldsymbol{D}^s\boldsymbol{\alpha}\|_F^2 \quad (1)$$

$$\text{s.t. } \boldsymbol{D}^v_{ik}\geq 0,\boldsymbol{D}^s_{ik}\geq 0,\boldsymbol{\alpha}_{kj}\geq 0,\ \forall i,j,k$$

设置各元素非负可以模拟一个物体由多个部分组成这种情况,更为接近人的直观感受,这一点在(Non-negative Matrix Factorization,NMF)[27]已有论述.

对于 $\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$ 我们对其添加一个结构化的稀疏性正则项,使某些 $\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$ 的列向量为零(即让某些词为零).当 $\boldsymbol{D}^v$ 中某一列为零,而 $\boldsymbol{D}^s$ 中对应的列不为零时,说明视觉特征空间与所学习的跨模态特征空间的对应维度无关,而标签特征空间与此维度相关,换言之,跨模态特征空间的对应维度为标签特征空间所独有的特征.我们可以通过在词典上添加一个 $L_{1,q}$ 范数来实现这种结构化的稀疏性,其中 q 的范围为 1 到∞.假设 $\boldsymbol{D}$ 表示两个模态中任一模态,则

$$\|\boldsymbol{D}\|_{1,q}=\sum_{k=1}^{K}\|\boldsymbol{D}_k\|_q,$$

其中 $\boldsymbol{D}_k$ 代表 $\boldsymbol{D}$ 的第 k 列,即第 k 个词.为了保证优化问题的凸性,通常设置 $q=2$ 或 $q=\infty$.在本文中,我们选择 $q=\infty$,因为在文献[28]中显示 $q=\infty$ 比 $q=2$ 效果更好.$L_{1,\infty}$ 定义如下:

$$\|\boldsymbol{D}\|_{1,\infty}=\sum_{k=1}^{K}\max_{1\leq i\leq M}|\boldsymbol{D}_{ik}| \quad (2)$$

其中,M 为 $\boldsymbol{D}$ 的行数.

除了在词典上添加结构化的稀疏性,我们在重构系数矩阵上也添加 L_1 范数稀疏项:

$$\|\boldsymbol{\alpha}\|_{1,1}=\sum_{i=1}^{N}\sum_{k=1}^{K}|\boldsymbol{\alpha}_{ki}| \quad (3)$$

综合式(1)(2)(3),我们跨模态特征学习的目标函数可以总结如下:

$$\min_{\boldsymbol{D}^v,\boldsymbol{D}^s,\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{X}^v-\boldsymbol{D}^v\boldsymbol{\alpha}\|_F^2+\frac{1}{2}\|\boldsymbol{X}^s-\boldsymbol{D}^s\boldsymbol{\alpha}\|_F^2+\gamma(\|\boldsymbol{D}^v\|_{1,\infty}+\|\boldsymbol{D}^s\|_{1,\infty})+\varphi\|\boldsymbol{\alpha}\|_{1,1} \quad (4)$$

$$\text{s.t. } \boldsymbol{D}^v_{ik}\geq 0,\boldsymbol{D}^s_{ik}\geq 0,1\geq\boldsymbol{\alpha}_{kj}\geq 0,\ \forall i,j,k$$

式(4)的优化,对于 $\boldsymbol{D}^v,\boldsymbol{D}^s,\boldsymbol{\alpha}$ 来讲是非凸的,所以我们只能找到局部最优解.在下面部分,我们介绍式(4)的优化方法.

4 优 化

当 $\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$ 固定时,目标函数(4)对于 $\boldsymbol{\alpha}$ 是凸函数,反之亦然.我们通过固定词典或者重构系数矩阵,来迭代的优化目标函数(4),过程如算法 1.我们的优化方法主要借鉴自文献[8],下面我们介绍下这两个子优化过程.

算法 1. 跨模态特征学习的优化过程.

输入:$\boldsymbol{X}^v,\boldsymbol{X}^s,\gamma,\varphi$

输出:$\boldsymbol{D}^v,\boldsymbol{D}^s,\boldsymbol{\alpha}$

开始

随机初始化 $\boldsymbol{D}^v_{ik}\geq 0,\boldsymbol{D}^s_{ik}\geq 0,1\geq\boldsymbol{\alpha}_{kj}\geq 0,\ \forall i,j,k$

重复

固定 $\boldsymbol{\alpha}$,优化目标函数(4),求解 $\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$

固定 $\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$,优化目标函数 4,求解 $\boldsymbol{\alpha}$

直到收敛或者达到最大迭代次数

结束

4.1 $\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$ 的优化

不难发现,如果固定了 $\boldsymbol{\alpha}$,$\boldsymbol{D}^v$ 和 $\boldsymbol{D}^s$ 的优化是独立的.既然每个词典的优化过程是相同的,那么我们只需介绍一种词典的优化过程,下面我们用 $\boldsymbol{X}$ 和 $\boldsymbol{D}$ 代指输入数据和所要优化的词典,则我们所要解决的子优化问题可以归纳为下面的公式:

$$\min_{\boldsymbol{D}}\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{\alpha}\|_F^2+\gamma\|\boldsymbol{D}\|_{1,\infty},\ \text{s.t. } \boldsymbol{D}_{ik}\geq 0,\forall i,k \quad (5)$$

我们用 $\varnothing(\boldsymbol{D})$ 表示目标函数(5),$\varnothing(\boldsymbol{D})$ 是一个复合目标函数,其中第一项具有强凸性和可微性,第二项具有凸性.我们可以根据文献[29]提出的复合梯度映射技术(composite gradient mapping)来设计优化算法.通过迭代地最小化一个辅助目标函数和 $\varnothing(\boldsymbol{D})$ 第一项的 Lipschitz 常数,$\varnothing(\boldsymbol{D})$ 的函数值可以快速的下降.$f(\boldsymbol{D})=0.5\cdot\|\boldsymbol{X}-\boldsymbol{D}\boldsymbol{\alpha}\|_{Fro}^2$,$\boldsymbol{D}^t$ 表示第 t 次迭代 $\boldsymbol{D}$ 的值.在此,辅助目标函数定义如下:

$$m_L(\boldsymbol{D}^t;\boldsymbol{D})=f(\boldsymbol{D}^t)+tr[\nabla f(\boldsymbol{D}^t)^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{D}^t)]+\frac{L}{2}\|\boldsymbol{D}-\boldsymbol{D}^t\|_F^2+\gamma\|\boldsymbol{D}\|_{1,\infty} \quad (6)$$

《计算机学报》

其中 L 是函数 $f(\cdot)$的 Lipschitz 常数估计值，L_f，$\nabla f(\boldsymbol{D}^t)$是函数 $f(\cdot)$在 $\boldsymbol{D}^t$处的梯度：

$$\nabla f(\boldsymbol{D}^t)=\boldsymbol{D}^t\boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}-\boldsymbol{X}\boldsymbol{\alpha}^{\mathrm{T}} \tag{7}$$

通过最小化 $m_L(\boldsymbol{D}^t;\boldsymbol{D})$，我们可以获取一个 $\boldsymbol{D}^{t+1}$的潜在值，此处用 $T_L(\boldsymbol{D}^t)$表示：

$$T_L(\boldsymbol{D}^t)=\underset{\boldsymbol{D}_{ik}\geq 0,\forall i,k}{\arg\min} m_L(\boldsymbol{D}^t;\boldsymbol{D}) \tag{8}$$

注意 $m_L(\boldsymbol{D}^t;T_L(\boldsymbol{D}^t))\leq\varnothing(\boldsymbol{D}^t)$，因为 $m_L(\boldsymbol{D}^t;\boldsymbol{D}^t)=\varnothing(\boldsymbol{D}^t)$并且 $T_L(\boldsymbol{D}^t)$是 $m_L(\boldsymbol{D}^t;\boldsymbol{D})$的最小元. 在文献[40]中已经证明了，当 $L\geq L_f$时，我们有 $\varnothing(T_L(\boldsymbol{D}^t))\leq m_L(\boldsymbol{D}^t;T_L(\boldsymbol{D}^t))$. 所以，目标函数(5)的优化算法从一个估计值 L_0开始，$0\leq L_0\leq L_f$，每次迭代调整 L 直到我们得到 $\varnothing(T_L(\boldsymbol{D}^t))\leq m_L(\boldsymbol{D}^t;T_L(\boldsymbol{D}^t))$. 具体算法如算法 2.

算法 2.　复合梯度映射.

输入：$\eta_u>1$，$\eta_d>1$；L 的尺度参数

开始

随机初始化 $\boldsymbol{D}_{ik}^0\geq 0,\forall i,k$，以及 $L_0:0<L_0\leq L_f$；$t=0$；

　重复

　　$L=L_t$；

　　重复

　　优化(8)得到 $T_L(\boldsymbol{D}^t)$；

　　如果$\varnothing(T_L(\boldsymbol{D}^t))>m_L(\boldsymbol{D}^t;T_L(\boldsymbol{D}^t))$，那么$L=L\eta_u$；

　　直到$\varnothing(T_L(\boldsymbol{D}^t))\leq m_L(\boldsymbol{D}^t;T_L(\boldsymbol{D}^t))$

　$\boldsymbol{D}^{t+1}=T_L(\boldsymbol{D}^t)$；

　$L_{t+1}=\max(L_0,L/\eta_d)$；

　$t=t+1$

　直到收敛

结束

在算法 2 中，内层循环寻找合适的 $\boldsymbol{D}^{t+1}$. 当 $\eta_u=\eta_d=2$ 时，经过 t 层外层迭代后，内层循环的迭代次数上界为 $2(t+1)+\log_2\frac{L_f}{L_0}$[29]. 经文献[29]证明，算法 2 收敛的时间复杂度为 $O(1/T)$，其中 T 为外层迭代的总次数.

那么，剩下的问题就是如何优化目标函数(8). 首先，将 $m_L(\boldsymbol{D}^t;\boldsymbol{D})$重写为($\boldsymbol{Y}$ 等于 $\boldsymbol{D}-\boldsymbol{D}^t$)：

$$\begin{aligned}
&m_L(\boldsymbol{D}^t;\boldsymbol{D})=\\
&f(\boldsymbol{D}^t)+tr[\nabla f(\boldsymbol{D}^t)^{\mathrm{T}}\boldsymbol{Y}]+\frac{L}{2}\|\boldsymbol{Y}\|_F^2+\gamma\|\boldsymbol{D}\|_{1,\infty}\\
&=\frac{L}{2}\left\{\|\boldsymbol{Y}\|_F^2+\frac{2}{L}tr[\nabla f(\boldsymbol{D}^t)^{\mathrm{T}}\boldsymbol{Y}]+\frac{1}{L^2}\|\nabla f(\boldsymbol{D}^t)\|_F^2\right\}+\\
&\gamma\|\boldsymbol{D}\|_{1,\infty}+f(\boldsymbol{D}^t)-\frac{1}{2L}\|\nabla f(\boldsymbol{D}^t)\|_F^2\\
&=\frac{L}{2}\left\|\boldsymbol{Y}+\frac{1}{L}\nabla f(\boldsymbol{D}^t)\right\|_F^2+\gamma\|\boldsymbol{D}\|_{1,\infty}+const.
\end{aligned}$$

用 $\boldsymbol{D}-\boldsymbol{D}^t$替代 $\boldsymbol{Y}$，则目标函数(8)的优化问题变成了：

$$\begin{aligned}
&\min_{\boldsymbol{D}}\frac{L}{2}\left\|\boldsymbol{D}-\boldsymbol{D}^t+\frac{1}{L}\nabla f(\boldsymbol{D}^t)\right\|_F^2+\\
&\gamma\sum_{k=1}^{K}\max_{1\leq i\leq M}|\boldsymbol{D}_{ik}|\\
&\text{s.t. }\boldsymbol{D}_{ik}\geq 0,\ \forall i,k
\end{aligned} \tag{9}$$

目标函数(9)的第一项是和 $\boldsymbol{D}$ 的每个元素相关的一个求和项，第二项按列计算，所以目标函数(9)可以按照 $\boldsymbol{D}$ 的列向量，进一步分解为许多独立的优化问题. 我们用 $\boldsymbol{d}$ 代表 $\boldsymbol{D}$ 中任一列向量，$\boldsymbol{b}$ 代表 $\boldsymbol{D}^t-\frac{1}{L}\nabla f(\boldsymbol{D}^t)$的对应的列向量，那么我们所要优化的子问题为

$$\min_{\boldsymbol{d}}\frac{1}{2}\|\boldsymbol{d}-\boldsymbol{b}\|_2^2+\frac{\gamma}{L}\|\boldsymbol{d}\|_\infty,\ \text{s.t. }\boldsymbol{d}_i\geq 0,\forall i \tag{10}$$

对于目标函数(10)，我们可以使用 SPAMS[30] 工具箱提供的相应方法来求解.

4.2　$\boldsymbol{\alpha}$ 的优化

当 $\boldsymbol{D}^v$和 $\boldsymbol{D}^s$固定后，对于 $\boldsymbol{\alpha}$ 的子优化问题可以写成：

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}\|\boldsymbol{X}^v-\boldsymbol{D}^v\boldsymbol{\alpha}\|_F^2+\frac{1}{2}\|\boldsymbol{X}^s-\boldsymbol{D}^s\boldsymbol{\alpha}\|_F^2+\varphi\|\boldsymbol{\alpha}\|_{1,1} \tag{11}$$
$$\text{s.t. }1\geq\boldsymbol{\alpha}_{kj}\geq 0,\ \forall j,k$$

这是一个有界的非负二次规划问题. 对于此类问题，文献[31]提出了一个优化方案. 根据文献[31]的优化方案，我们设计了 $\boldsymbol{\alpha}$ 的优化算法.

首先，我们将目标函数(11)主要部分重写为

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}\sum_{r=1}^{2}\|\boldsymbol{X}^{(r)}-\boldsymbol{D}^{(r)}\boldsymbol{\alpha}\|_F^2+\varphi\|\boldsymbol{\alpha}\|_{1,1},$$

其中 $r=1$ 时，$\boldsymbol{X}^{(r)}$和 $\boldsymbol{D}^{(r)}$代表视觉特征模态，$r=2$ 时代表标签特征模态. 对上式第一部分，我们做一下转化：

$$\begin{aligned}
&\frac{1}{2}\sum_{r=1}^{2}\|\boldsymbol{X}^{(r)}-\boldsymbol{D}^{(r)}\boldsymbol{\alpha}\|_F^2\\
&=\frac{1}{2}\sum_{r=1}^{2}tr[(\boldsymbol{X}^{(r)}-\boldsymbol{D}^{(r)}\boldsymbol{\alpha})^{\mathrm{T}}(\boldsymbol{X}^{(r)}-\boldsymbol{D}^{(r)}\boldsymbol{\alpha})]\\
&=\frac{1}{2}\sum_{r=1}^{2}(tr[\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{D}^{(r)})^{\mathrm{T}}\boldsymbol{D}^{(r)}\boldsymbol{\alpha}]-\\
&\quad 2tr[\boldsymbol{\alpha}^{\mathrm{T}}(\boldsymbol{D}^{(r)})^{\mathrm{T}}\boldsymbol{X}^{(r)}])+const.
\end{aligned}$$

为了式子进一步清晰，设 $\boldsymbol{P}=\sum_{r=1}^{2}(\boldsymbol{D}^{(r)})^{\mathrm{T}}\boldsymbol{D}^{(r)}$，$\boldsymbol{Q}=\sum_{r=1}^{2}(\boldsymbol{D}^{(r)})^{\mathrm{T}}\boldsymbol{X}^{(r)}$，则目标函数(11)可以转化为如下形式：

$$\min_{\boldsymbol{\alpha}}\frac{1}{2}tr[\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\alpha}]-tr[\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{Q}]+\varphi\|\boldsymbol{\alpha}\|_{1,1} \tag{12}$$
$$\text{s.t. }1\geq\boldsymbol{\alpha}_{kj}\geq 0,\ \forall j,k$$

《计算机学报》

目标函数(12)的优化几乎可以完全参照文献[31],用 $\boldsymbol{\alpha}$ 的任一列向量代替 $\boldsymbol{\alpha}$,除 $\boldsymbol{\alpha}$ 的选项外,即文献[31]中所优化的问题.比如目标函数(12)第一项可转化为

$$\frac{1}{2}tr[\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\alpha}]=\frac{1}{2}\sum_{j=1}^{N}(\boldsymbol{\alpha}_j)^{\mathrm{T}}\boldsymbol{P}\boldsymbol{\alpha}_j.$$

仿照文献[31]我们可以设计一个辅助函数,并迭代地更新 $\boldsymbol{\alpha}$,具体的推导过程此处不再给出,下面我们给出辅助函数(13)与更新规则(14):

$$\mathcal{G}(\boldsymbol{\alpha}^t;\boldsymbol{\alpha})=\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{K}\frac{(\boldsymbol{P}\boldsymbol{\alpha}_j^t)_k}{\boldsymbol{\alpha}_{kj}^t}(\boldsymbol{\alpha}_{kj})^2+\sum_{j=1}^{N}\sum_{k=1}^{K}(\varphi-\boldsymbol{Q}_{kj})\boldsymbol{\alpha}_{kj} \tag{13}$$

$$\boldsymbol{\alpha}_{kj}^{t+1}=\min\left\{1,\frac{-(\varphi-\boldsymbol{Q}_{kj})+|\varphi-\boldsymbol{Q}_{kj}|}{2(\boldsymbol{P}\boldsymbol{\alpha}_j^t)_k}\boldsymbol{\alpha}_{kj}^t\right\} \tag{14}$$

4.3 时间复杂度分析

时间复杂度包含两部分:词典的优化和重构系数的优化.优化每个 $\boldsymbol{D}$ 时,我们需要运行算法 2.核心循环是目标函数(8)的优化,其要求对词典的每一列来优化目标函数(10),代价为 $O(p)$,p 为对应特征向量的维数,则目标函数(8)的优化代价为 $O(pK)$.在算法 2 的外层循环我们需要计算 $\nabla f(\boldsymbol{D})$,代价为 $O(pK^2)$.假设我们运行了 T 次外层循环,则内层循环的迭代次数上界为 $2(T+1)+\log_2\frac{L_f}{L_0}$[29].对于单个词典的优化,主要的时间代价为 $O\left(pK\left(2(T+1)+\log_2\frac{L_f}{L_0}\right)+TpK^2\right)$.对于 $\boldsymbol{\alpha}$ 的优化,每次迭代主要计算 $\boldsymbol{Q}(O(NK))$ 和 $\boldsymbol{P}\boldsymbol{\alpha}(O(NK^2))$,总代价为 $O(T'(NK+NK^2))$,T' 为迭代次数.在实际应用中,T 和 T' 的值可以设置的小一些,比如 10 或者 15.

5 语义相似度度量框架

本文提出的语义相似度度量框架并不复杂,简单讲就是在社交标签的词义基础上从两个角度来度量:概念距离和释义相似度.首先,我们通过一个简单的关联度最大方法为多义词标签确定两个通用词义,然后从上面所提的两个角度来度量标签之间的语义相似度.在本文中,我们通过标签间的语义相似度来计算样本间的语义相似度,并将相似度矩阵作为社交标签的特征矩阵来使用.

5.1 通用词义检测

我们假设多义词最通用的词义与其他词义的关联度最大,但必须承认这是一个非常强的假设,与实际情况可能有些出入,所以为了增强算法的泛化能力,我们将通用词义的数量经验性地设置为二.词义之间的关联性体现在其定义和 gloss 上.我们借助一种名为 Gloss Vector[32] 的语义相似性度量算法来计算词义间的关联性.多义词可能包含多个词性,我们按照"名词-动词-形容词-副词"的优先级顺序进行计算.首先,按照词性优先级确定通用词义候选集,确保候选集中词义数量至少为二.然后,使用 Gloss Vector 算法度量每个候选词义与其他词义的相似度.最后,对候选词义按其与其他词义相似度之和进行排序,前两位为通用词义.

5.2 Gloss Vector 算法

Gloss Vector 算法是 Context Vector 的 WordNet 版本.Context Vector 最早由 Schutze[33] 提出,由 SiddharthPatwardhan[32] 结合 WordNet 进行了改进.Context vector 算法基于一个假设[34]:人通过单词的上下文信息判断其词义.Schutze 按照这个观点,将某个词的"上下文词"映射至一个词空间,表示成一个个的词向量,然后将这些词向量相加作为这个词的表征向量.为了获得泛化能力较强的词空间,首先需要一个较大的语料库.词空间简单说就是一张词表,用于向量量化.Siddharth 使用 WordNet 中所有 synsets 的词义解释作为语料库,结合其中每个词的出现频率及一个停用词列表来确定词表.词表确定完成后,下一步是为词表中的每个词生成一个词向量.按照下面的步骤,我们可以为词 w 生成一个词向量:

(1)初始化一个全零的向量 $\boldsymbol{w}$.

(2)在 WordNet 的全部 gloss 中找到所有 w 出现的地方.

(3)对每一个出现 w 的 gloss,根据预设的窗口大小,统计 w 周围的词,并且在 w 相应的维度上增加计数.

这样生成的词向量 $\boldsymbol{w}$,就对词 w 的共生信息进行了编码.对每个词都生成这样的词向量,每个词向量可以视为对应了真实世界中的某个确定的主题或方面.我们的目的是度量 WordNet 中任意两个词义之间的语义相似度,所以需要生成一个词向量以表征具体的词义.每个词义 gloss 中的词,我们视为此词义的"上下文词",将这些"上下文词"的词向量

进行相加，得到的向量我们称之为此词义的 gloss vector，也就是用来表征词义的向量. 对于任意两个词义，计算它们的 gloss vector 之间的余弦值，即为它们之间的语义相似度.

5.3 语义相似度度量

WordNet 中的“*Is-a*”关系反映了概念之间的相关性，它将大量概念连接成一个语义网络. 概念在这个语义网络上的距离，即间隔的节点数量，反映了概念之间的语义相关性.

我们使用两种算法来度量两个标签之间的语义相关性：通过“*Is-a*”语义网络度量其概念距离(conceptual distance)；通过 Gloss Vector 算法度量它们之间的释义相似度(gloss similarity). 假设我们有两个 synset，用 S_1 和 S_2 表示. 它们之间的概念距离用 $length(S_1,S_2)$ 表示，概念距离即 S_1 和 S_2 之间的节点数量. 通过概念距离，我们按照下面的公式计算 S_1 和 S_2 在“*Is-a*”语义网络上的相似度，

$$C(S_1,S_2)=-\log\left(\frac{length(S_1,S_2)}{2depth}\right) \tag{15}$$

其中，$depth$ 代表“*Is-a*”语义网络的最大层数，本文设为 16.

注意，式(15)只能计算名词词性的 synset 之间的语义相似度. 当 S_1 和 S_2 有一个为非名词时，我们只能使用 Gloss Vector 计算它们之间的语义相似度. 在此，我们用 $G(S_1,S_2)$ 来表示 S_1 和 S_2 通过 Gloss Vector 算法计算出的语义相似度. 当 S_1 和 S_2 均为名词词性的 synset 时，通过加权求和来结合两种算法，总的语义相似度为

$$Simi(S_1,S_2)=\frac{C(S_1,S_2)+G(S_1,S_2)}{2} \tag{16}$$

当 S_1 和 S_2 中存在非名词词性的 synset 时，总的语义相似度为 $G(S_1,S_2)$.

假设我们有两个标签 A 和 B，其通用词义分别为 $\{a_1,a_2\}$ 和 $\{b_1,b_2\}$，则 A 和 B 之间的语义相似度为

$$Simi(A,B)=0.25\sum_{i=1}^{2}\sum_{j=1}^{2}Simi(a_i,b_j).$$

假设我们有两个样本 P 和 Q，其标签集合分别为 $\{p_1,p_2,\cdots,p_m\}$ 和 $\{q_1,q_2,\cdots,q_n\}$，则样本 P 和 Q 之间的语义相似度为

$$Simi(P,Q)=\frac{\left(\sum_{i=1}^{m}Simi(p_i,q)+\sum_{i=1}^{n}Simi(q_i,p)\right)}{2},$$

$$Simi(p_i,q)=\max_{j=1,\cdots,n}Simi(p_i,q_j),$$

$$Simi(q_i,p)=\max_{j=1,\cdots,m}Simi(q_i,p_j).$$

6 聚　类

样本间的相似度矩阵根据下面的公式进行计算：

$$A_{ij}=\exp\left(\frac{-d^2(I_i,I_j)}{\sigma_i\sigma_j}\right),\ i\neq j;\ A_{ii}=0,$$

其中 $d(I_i,I_j)$ 为样本 i 和样本 j 根据对应的特征所计算出的欧式距离，尺度参数 σ_i 和 σ_j 由一个自适应的方法[35]自动进行调节. 文献[35]通过统计样本的近邻信息来调节 σ_i 和 σ_j，并称之为局部尺度化. 每个样本对应一个参数 σ，σ 等于样本的第 K 个最近邻与样本的距离，K 值为经验值，在文献[35]中设为 7.

在得到样本间的相似度矩阵后，鉴于谱聚类在许多工作中取得的出色表现，我们使用谱聚类算法进行聚类. 谱聚类[36]为 Ng 等人在 Ncut 和 Malik[37]基础上进行小幅度修改而来，其算法流程如下：

1. 给出一个 R^l 空间内的样本点集 $S=\{s_1,\cdots,s_n\}$，假设我们要将 S 分为 K 个子集，$\boldsymbol{A}$ 为样本间的相似度矩阵.

2. 定义 $\boldsymbol{D}$ 为一对角阵，其 (i,i) 元素值为矩阵 $\boldsymbol{A}$ 第 i 行的和，然后构造矩阵 $\boldsymbol{L}$，$\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$.

3. 对 $\boldsymbol{L}$ 计算出 k 个特征值最大的特征向量，并组成矩阵 $\boldsymbol{X}$，$\boldsymbol{X}$ 中每一列对应一个特征向量.

4. 对 $\boldsymbol{X}$ 的行进行归一化处理，得到矩阵 $\boldsymbol{Y}$.

5. 将 $\boldsymbol{Y}$ 中的每一行视作 R^k 空间内的一个点，使用 K-means 算法将它们分至 k 个簇.

6. 最终，将原始样本 s_i 分配至簇 j，当且仅当 $\boldsymbol{Y}$ 的第 i 行被分配至簇 j.

7 实　验

7.1 数据集

在实验中，我们一共使用了两个数据集，其中一个从 NUS-WIDE[38]中收集而来，在此我们称之为 NUS-WIDE 数据集. NUS-WIDE 数据集由新加坡国立大学的多媒体搜索实验室从图片分享网站 Flickr 上收集整理而来，我们从中整理了一个 20 000 幅图像的数据集，共包含 9935 个不同的社交标签，每幅图像对应的标签数量从 3～30 不等，平均约为 7 个. 另外一个数据集从 IAPR TC-12 Benchmark[39]中收集而来. IAPR TC-12 Benchmark 数据集是一个公开的图像标注数据集，每幅图片有 4～5 个标签，共 292 个标签，我们从中选取了 12 000 幅图像，在此简称 IAPR.

7.2 实验指标

为了全面的对聚类结果进行评估，我们使用了两个指标[40]：Normalized Mutual Information（NMI）和 Clustering Error（CE）. *NMI* 从一定程度上反映了各个簇的内聚性，*NMI* 值越大，聚类结果越好；*CE* 则可视为聚类当中的分类错误率，越小越好. 结合两个指标进行分析可以更为全面、准确的分析聚类的结果.

对于两个随机变量 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$，它们间的 *NMI* 值定义如式(17)：

$$NMI(X,Y)=\frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{17}$$

其中，$I(X,Y)$是X和Y之间的互信息，$H(X)$和$H(Y)$是它们各自的熵. 显而易见，*NMI* 的最大值为 1. 对于聚类结果的评估，上式的具体形式为式(18)：

$$NMI=\frac{\sum_{l=1}^{c}\sum_{h=1}^{c}n_{l,h}\log\left(\frac{n\times n_{l,h}}{n_l\hat{n}_h}\right)}{\sqrt{\left(\sum_{h=1}^{c}n_l\log\frac{n_i}{n}\right)\left(\sum_{h=1}^{c}\hat{n}_h\log\frac{\hat{n}_h}{n}\right)}} \tag{18}$$

其中，n_l代表簇$C_l(1\leq l\leq c)$包含的样本数量，$\hat{n}_h$ 是真值中第h 个簇$(1\leq h\leq c)$包含的样本数量，$n_{l,h}$代表这两个簇的交集包含的样本数量.

为了从聚类结果中计算 *CE*，我们需要一个映射函数，为聚类结果中的簇与真值中的簇建立对应关系. 当映射函数确定，我们便可以按式(19)计算一个误差值：

$$err=1-\frac{\sum_{i=1}^{n}\delta(y_i,map(c_i))}{n} \tag{19}$$

其中，y_i和c_i分别是样本i 的真值簇索引和聚类结果中的簇索引. 当$x=y$，$\delta(x,y)$等于1；否则，等于0. 不同的映射函数可以得到不同的误差值，通过 Hungarian 算法进行优化，我们选择其最小值作为 *CE* 值.

7.3 对比方法

我们使用下面四种方法作为对比方法：

(1) NMF-b：对两种模态分别使用 NMF 进行重编码[27]. 我们在所有图表中报告两种模态中较好的那个结果.

(2) ConcateNMF：将两种模态的特征直接相连，然后使用 NMF.

(3) MultiNMF：将 γ 和 φ 设置为零，只学习两种模态共享的信息.

(4) CCA：使用典型相关分析算法[1]学习两种模态共享的信息.

7.4 实验结果对比

依据经验，对于 NUS-WIDE 数据集，我们将词典长度设置为 75，对于 IAPR 设置为 50. 通过第五部分的方法，我们计算得到样本之间的语义相似度矩阵，此矩阵包含了样本间的语义关系. 在此矩阵基础上，我们求其距离直方图作为标签信息模态的特征. 对于视觉模态信息，我们使用 Alex-CNN[41] 倒数第二层的激活值作为特征.

图 1 和图 2 展示了各种方法的对比结果. 对于 CCA，我们按照典型相关系数的排序，选取前 d 对典型得分作为特征向量，并进行聚类. 随着 d 的增长，CCA 特征的聚类效果逐渐变好，我们在图 1 和图 2 中报告聚类结果稳定后的值，CCA 的实验结果展示在图 13 中. 我们使用 S&P 表示本文方法. 根据 γ 和 φ 值的变化，S&P 方法效果有所不同，我们在图 1 和图 2 中的报告较接近于平均值的结果.

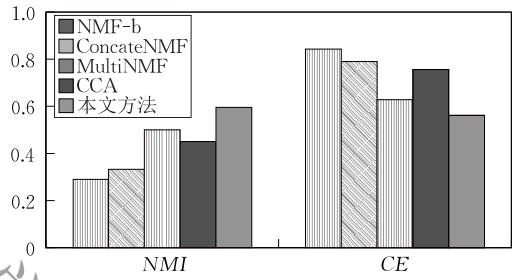


图 1 NUS-WIDE 数据集上的实验结果

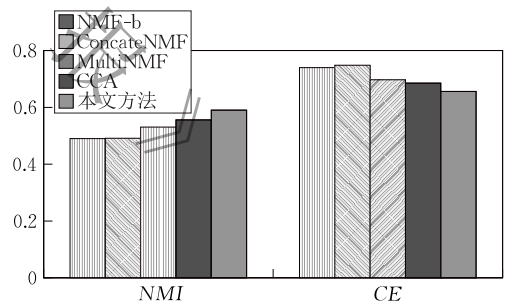


图 2 IAPR 数据集上的实验结果

从图 1 和图 2 中可以看出：(1) 结合不同模态的方法在聚类任务上的表现优于单模态聚类，这一点符合之前在跨模态学习或者说多角度学习方面的研究结论，贯穿于各种不同模态的信息往往具有很强的描述力；(2) 本文方法通过学习共享＋独有信息，效果优于只学习共享信息的方法，这一点印证了我们的研究思路与方法.

7.5 参数研究

本文的方法有两个参数：γ 和 φ，分别控制了词典和重构系数的稀疏性. 参数值与词典长度 K 的设置对本文方法的效果有着直接的影响，我们通过实

验来观察不同参数值对聚类结果的影响，并希望读者在使用本文方法时有所参照，结合自身的实际应用情况进行试验来确定各个值.

对于 γ，从图 3 和图 4 中可以看出，随着其值的变大聚类效果也在逐渐变好，直至一定程度后基本稳定，这说明词典的稀疏性限制的确帮助我们学习到了一个更好的特征空间. 同时还可以看出，本文方法对 γ 值不是很敏感，只要大于某一值，其表现基本稳定. 从图 5 和图 6 可以看出，φ 值对本文方法的表现贡献较小，取稍微大点的值可以增强聚类效果，但继续增大，则有可能使特征过度稀疏，从而影响表现. K 值的设置同 γ 值有些类似，当 K 值大于一定值后，表现趋于稳定.

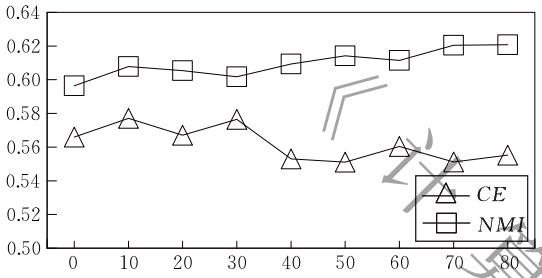


图 3　NUS 数据集(φ 为 0.005，字典长度为 75)

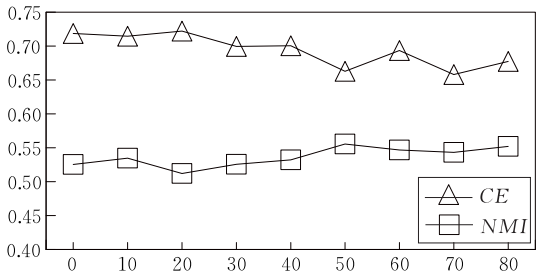


图 4　IAPR 数据集(φ 为 0.005，字典长度为 50)

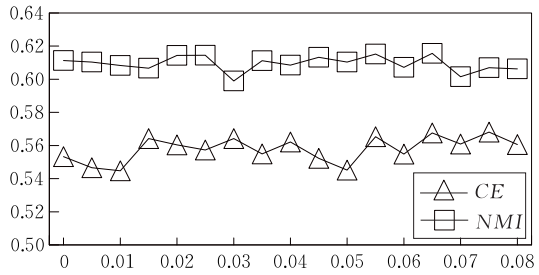


图 5　NUS 数据集(γ 为 25，字典长度为 75)

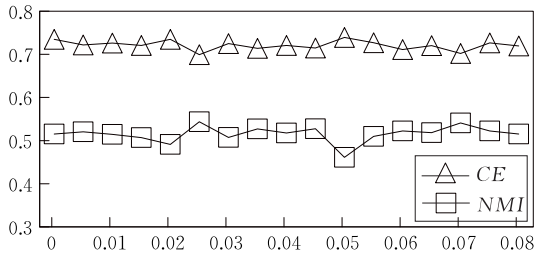


图 6　IAPR 数据集(γ 为 25，字典长度为 50)

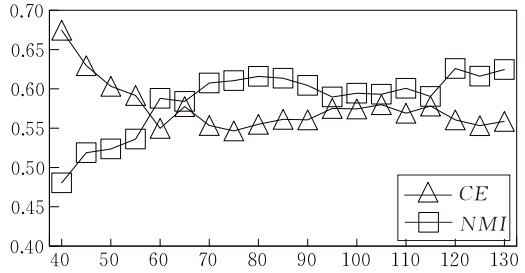


图 7　NUS 数据集(γ 为 25，φ 为 0.005)

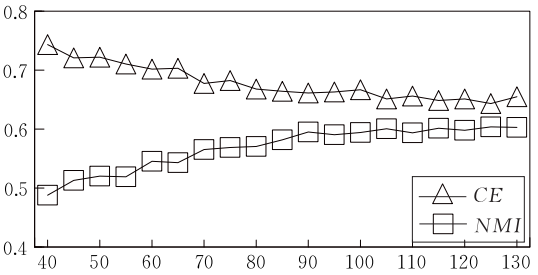


图 8　IAPR 数据集(γ 为 25，φ 为 0.005)

7.6　收敛分析

本文对目标函数(4)的优化只能找到其局部最小值，所以对算法 1 的收敛性有必要做一个经验性的研究分析. 图 9 与图 11 展示了函数值下降的速度，Y 轴为每次下降的幅度，图 10 和图 12 为对应的聚类表现. 从这四个图中可以看出，随着算法的运行，函数值先是猛烈下降，然后下降的幅度变得平缓，基本上大于 10 次之后，函数值下降的速度就已经十分缓慢，大于 25 次之后，函数值基本收敛. 与此对应的是，聚类表现的收敛显然更快，在迭代 5 次左右，聚类表现便已达到峰值，而此时函数并未收敛，

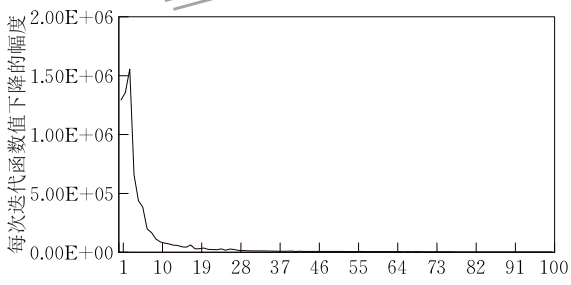


图 9　NUS 数据集

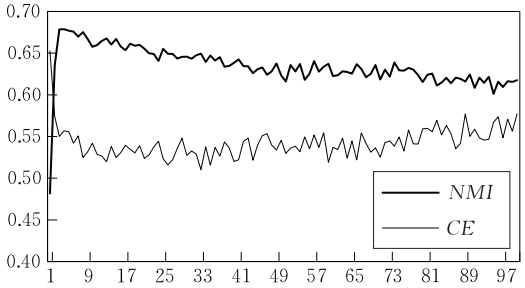


图 10　NUS 数据集

《计算机学报》

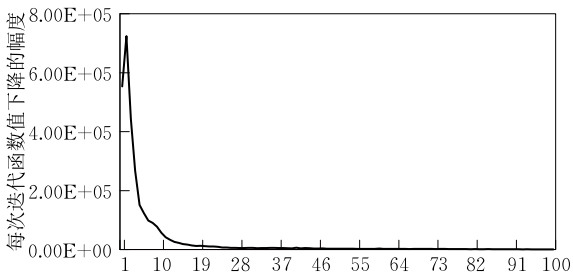


图 11 IAPR 数据集

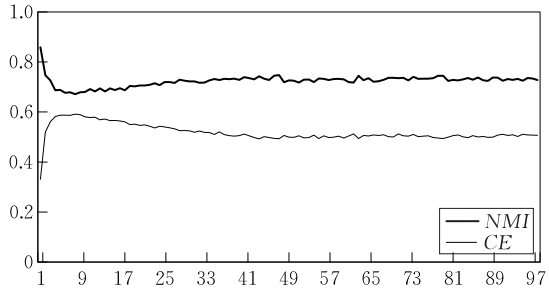
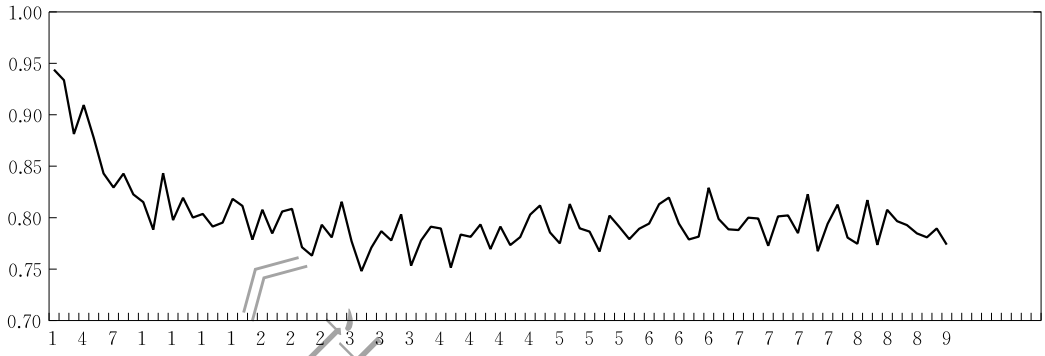
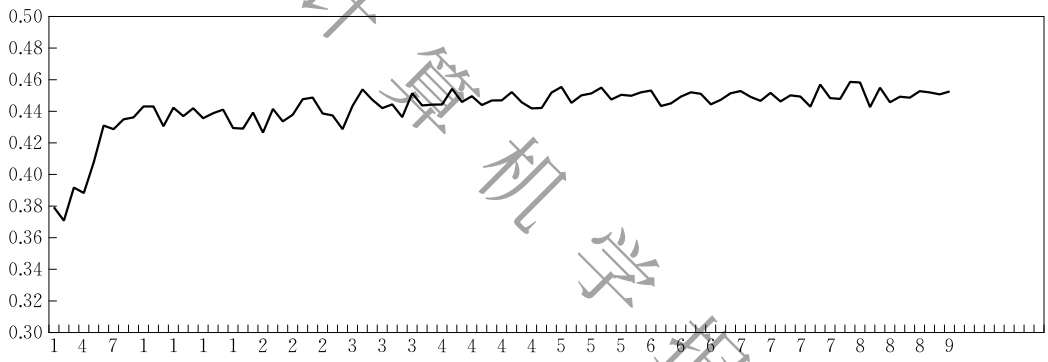
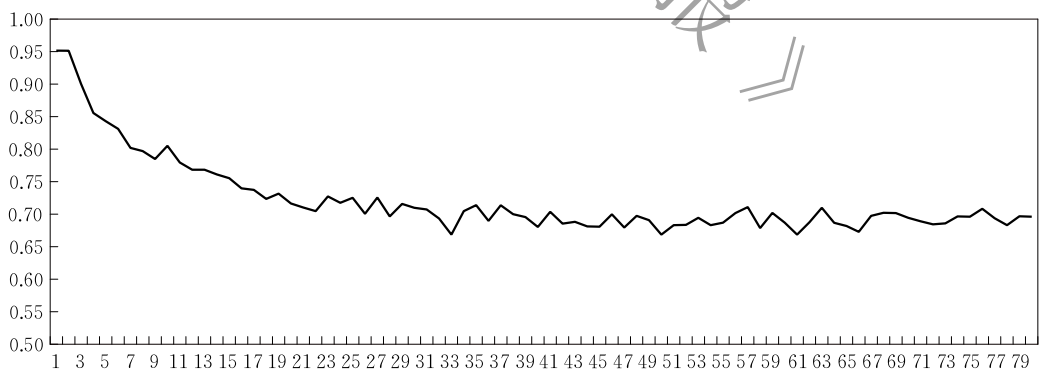
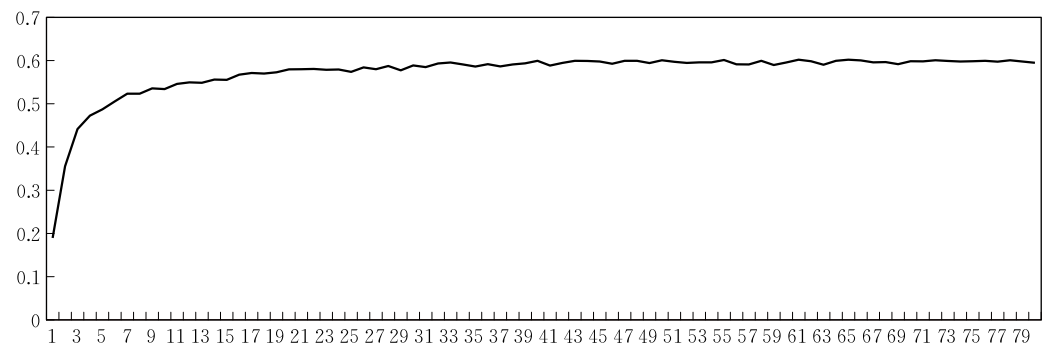


图 12 IAPR 数据集









《计算机学报》

图 13 典型相关分析算法在两个数据集上的实验结果(从上到下依次为:NUS-CE,NUS-NMI,IAPR-CE 和 IAPR-NMI)

在迭代 10 次之后，聚类效果开始出现轻微下降. 从这 4 幅图来看，算法 1 仅仅运行有限次迭代，便可达到较好的聚类效果，这对实际应用中，降低计算量有很好的指导意义.

8 总　结

在跨模态学习领域，大多数研究人员将精力集中于多模态信息的共享特征空间学习上，往往忽略了各模态信息本身所特有的特征，这显然并不符合实际应用情况. 本文提出了一个基于共轭词典学习的跨模态特征学习机制，以学习一个“共享&独有”模式的跨模态特征. 与此同时，为了深入挖掘标签所蕴含的语义关系，我们提出了一个基于 WordNet 的计算框架.

从方法上讲，本文最独特的地方在于在词典上加入结构化稀疏性限制，这一点同大多数基于稀疏编码的词典学习机制不同. 大多数基于稀疏编码的词典学习机制通过控制重构系数的稀疏性来达到较好的特征学习. 在词典上添加稀疏性限制，不仅可以使我们学习到“共享&独有”的跨模态特征，而且还可以使词典更为紧凑. 实验结果表明，较小的词典长度就可以获得较好的任务表现，这一点对于实际应用非常有价值.

在两个较大的数据集上我们进行了充分的实验，并从参数设置与收敛性分析方面，给出了经验性的指导. 实验结果表明，本文提出的方法对参数设置敏感性较低，这一点对实际应用非常有意义. 与此同时，较低次数的迭代可以在目标收敛前获得较好的聚类效果，也就是说，我们可以通过较低的计算量获得较好的跨模态特征.

参　考　文　献

[1] Hotelling H. Relations between two sets of variates. Breakthroughs in Statistics, 1992, 14: 162-190

[2] Xu X, Shimada A, Taniguchi R, et al. Coupled dictionary learning and feature mapping for cross-modal retrieval// Proceedings of the International Conference on Multimedia and Expo. London, UK, 2015: 1-6

[3] Deng C, Tang X, Yan J, et al. Discriminative dictionary learning with common label alignment for cross-modal retrieval. IEEE Transactions on Multimedia, 2016, 18(2): 208-218

[4] Zhuang Y T, Wang Y F, Wu F, et al. Supervised coupled dictionary learning with group structures for multi-modal retrieval//Proceedings of the AAAI Conference on Artificial Intelligence. Bellevue, USA, 2013: 1070-1076

[5] Blei D M, Jordan M I. Modeling annotated data//Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval. Toronto, Canada, 2003: 127-134

[6] Jia Y, Salzmann M, Darrell T. Factorized latent spaces with structured sparsity//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2010: 982-990

[7] Jia K, Wang X, Tang X, et al. Image transformation based on learning dictionaries across image spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(2): 367-380

[8] Guan Z, Zhang L, Peng J, et al. Multi-view concept learning for data representation. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(11): 3016-3028

[9] Harris Z S. Distributional structure. Word, 1954, 10(2-3): 146-162

[10] Costa Pereira J, Coviello E, Doyle G, et al. On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 521-535

[11] Rasiwasia N, Costa Pereira J, Coviello E, et al. A new approach to cross-modal multimedia retrieval//Proceedings of the 18th ACM International Conference on Multimedia. Firenze, Italy, 2010: 251-260

[12] Miller G A. WordNet: A lexical database for English. Communications of the ACM, 1995, 38(11): 39-41

[13] Cai D, He X, Li Z, et al. Hierarchical clustering of WWW image search results using visual, textual and link information// Proceedings of the 12th Annual ACM International Conference on Multimedia. New York, USA, 2004: 952-959

[14] Wang X J, Ma W Y, Xue G R, et al. Multi-model similarity propagation and its application for web image retrieval// Proceedings of the 12th Annual ACM International Conference on Multimedia. New York, USA, 2004: 944-951

[15] Rege M, Dong M, Hua J. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering//Proceedings of the 17th International Conference on World Wide Web. Beijing, China, 2008: 317-326

[16] Gao B, Liu T Y, Qin T, et al. Web image clustering by consistent utilization of visual features and surrounding texts//Proceedings of the 13th Annual ACM International Conference on Multimedia. Singapore, 2005: 112-121

[17] Dhillon I S. Co-clustering documents and words using bipartite spectral graph partitioning//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 269-274

计算机学报

[18] Barnard K, Forsyth D. Learning the semantics of words and pictures//Proceedings of the International Conference on Computer Vision. Vancouver, Canada, 2001, 2: 408-415

[19] Hofmann T. Learning and representing topic—A hierarchical mixture for word occurrence in document databases//Proceedings of the Workshop on Learning from Text and the Web. Pittsburgh, USA, 1998

[20] Barnard K, Duygulu P, Forsyth D. Clustering art//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, USA, 2001: 434-441

[21] Barnard K, Duygulu P, Forsyth D, et al. Matching words and pictures. The Journal of Machine Learning Research, 2003, 3(6): 1107-1135

[22] Sigal L, Memisevic R. Shared kernel information embedding for discriminative inference. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4): 778-790

[23] Ek C H, Torr P H, Lawrence N D, et al. Gaussian process latent variable models for human pose estimation//Proceedings of the International Conference on Machine Learning. Corvallis, USA, 2007: 132-143

[24] Navaratnam R M, Fitzgibbon A. The joint manifold model for semi-supervised multi-valued regression//Proceedings of the International Conference on Computer Vision. Rio de Janeiro, Brazil, 2007: 1-8

[25] Shon A P, Grochow K, Hertzmann A, et al. Learning shared latent structure for image synthesis and roboticimitation//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2005: 1233-1240

[26] Salzmann M, Ek C H, Urtasun R, et al. Factorized orthogonal latent spaces. Journal of Machine Learning Research, 2010, 9: 701-708

[27] Lee D D, Seung H S. Learning the parts of objects by nonnegative matrix factorization. Nature, 1999, 401(6755): 788-791

[28] Quattoni A, Carreras X, Collins M, Darrell T. An efficient projection for L1 infinity regularization//Proceedings of the International Conference on Machine Learning. Brisbane, Australia, 2009: 857-864

[29] Nesterov Y. Gradient methods for minimizing composite functions. Mathematical Programming, 2013, 140(1): 125-161

[30] Mairal J, Bach F, Ponce J. Sparse modeling for image and vision processing. Foundations and Trends in Computer Graphics and Vision, 2014, 8(2-3): 85-283

[31] Sha F, Lin Y, Saul L K, et al. Multiplicative updates for nonnegative quadratic programming. Neural Computation, 2007, 19(8): 2004-2031

[32] Patwardhan S. Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness [M. S. dissertation]. University of Minnesota, Duluth, 2003

[33] Schutze H. Automatic word sense discrimination. Computational Linguistics, 1998, 24(1): 97-123

[34] Miller G A, Charles W G. Contextual correlates of semantic similarity. Language and Cognitive Processes, 1991, 6(1): 1-28

[35] Zelnik-Manor L, Perona P. Self-tuning spectral clustering//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2004: 1601-1608

[36] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2002: 849-856

[37] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[38] Chua Tat-Seng, Tang Jin-Hui, Hong Ri-Chang, et al. NUS-WIDE: Areal-world Web image database from National University of Singapore//Proceedings of the ACM International Conference on Image and Video Retrieval. Island of Santorini, Greece, 2009: 8-10

[39] Grubinger M, Clough P D, Müller H, Deselaers T. The IAPR benchmark: Anew evaluation resource for visual information systems//Proceedings of the International Conference on Language Resources and Evaluation. Genoa, Italy, 2006: 21-32

[40] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research, 2003, 3: 583-617

[41] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097-1105

ZHAO Qi-Lu, born in 1987, Ph. D. candidate. His research interests include computer vision and machine learning.

LI Zong-Min, born in 1965, Ph. D., professor. His research interests include image processing, pattern recognition and computer graphics.

Background

In this paper, we focus on the research of cross-modal clustering, which is a branch of multi-view learning as known in machine learning community. The research is part of the project "Research on the Key Points of Visual Technologies in Clothing Product Search", funding by the National Natural Science Foundation of China. Clothing product search is a complex project, involving many aspects. Clustering can provide the structure of the product data. With a more effective organization of structure, we can achieve more precise retrieval results and faster speed.

Like social image, clothing image often has some tags, such as "red", "round collar" and so on. We are still harvesting the clothing images from the Internets, so we didn't use any clothing image in experiments here. In this paper, we proved that leveraging an information database can provide sufficient semantic relations in cross-modal learning, which lead to more correlations and powerful descriptive ability. For clothing image organization, we need a specific information database. Apparently, WordNet is not suitable for cross-modal clothing image clustering. Designing and generating a novel information database is our next move.

The canonical correlation analysis algorithm we used in this paper has not considered the specific information of each feature modality. We can view each feature modality as a combination of shared and specific information. Abandoning the specific information is wasteful. After completion of information database, we intend to design an algorithm, which treats specific information seriously.

Above all, we have introduced the role of this paper in the project "Research on the Key Points of Visual Technologies in Clothing Product Search" and the problems of cross-modal learning. Thanks for all the supports we have. This work is partly supported by the National Natural Science Foundation of China (Grant No. 61379106), the Shandong Provincial Natural Science Foundation (Grant Nos. ZR2009GL014, ZR2013FM036, ZR2015FM011), the Open Project Program of the State Key Laboratory of CAD&CG (Grant No. A1315), Zhejiang University.

《计算机学报》