DOI:10.13878/j.cnki.jnuist.2017.06.011

刘君芳[1]　邵曦[1]

多模态融合的家庭音乐相册自动生成

摘要

随着大数据以及社交网络的发展,电子相册与在线服务成为如今人们使用计算机与互联网的基础应用.尤其是近年社交网络的流行,电子相册的数量得到了爆炸增长,而如何增强相册的用户体验变得尤为重要.具有某种主题的相册一般都带有一定的情感信息,因此,本文研究了基于多模态融合的家庭音乐相册自动生成问题,旨在使用户能够在享受音乐的同时配以与音乐情感相同的相册图片.针对音乐与图片中所蕴含的情感,本文在音乐和图像中分别选取能够表达其情感的句子级别的音频特征和图像特征,然后在图像与音乐之间异构和跨模态的特征融合问题上,采用局部保持投影(LPP)方法,将图像特征与音乐特征映射到更具情感分类能力的隐式特征空间中,实现了音乐相册的自动生成.在实验中,客观评测结果表明,采用 LPP 方法在查准率方面高于纯 CCA 方法;在主观评测中 LPP 获得 72.06%的满意度,与人工推荐的评价结果(78.09%)比较接近,明显高于随机推荐和 CCA 方法的满意度.

关键词

音乐相册;情感模型;句子级别;多模态融合;隐式空间

中图分类号 TN912

文献标志码 A

收稿日期 2017-08-28

资助项目 国家自然科学基金(61401227);北京市自然科学基金(4152053)

作者简介

刘君芳,女,硕士生,研究方向为多媒体信息系统与多媒体通信.ljf846344673@163.com

邵曦(通信作者),男,博士,副教授,主要研究方向为多媒体信息系统与多媒体通信.shaoxi@njupt.edu.cn

1 南京邮电大学 通信与信息工程学院,南京,210003

0 引言

随着大数据和信息技术的飞速发展,电子相册与在线服务是如今人们使用计算机与互联网的基础应用.作为一种越来越重要的多媒体服务,自动音乐、图像检索问题逐渐成为一个引人关注的研究课题.

这些年来随着移动互联网平台的不断发展,数字图像的数量也得到了爆炸式的增长.Facebook (www.facebook.com)和 Flickr(www.flickr.com)的相册就是典型的代表.截至 2013 年 9 月,Flickr 已拥有超 60 亿张的图片.2012 年 5 月,Facebook 已拥有约 9 亿用户,每天均会上传数亿张照片.总之,不管是在线上服务还是移动平台,电子相册服务占据着越来越重要的位置,因而在改进其用户体验以及完善其功能上,具有很大的研究发展空间.

具有某种主题的电子相册一般都带有一定的情感信息,例如:一组婚礼的相册可能具有欢乐与浪漫的氛围,而一组拳击比赛的相册可能带有激烈和让人兴奋的感觉.假使用户使用不同的移动终端通过互联网浏览他们的照片时,可以同时欣赏到符合照片情境的背景音乐,便会带来与众不同的感受.但让用户自己选择背景音乐存在费时费力不够专业的缺点.因此,若能自动生成音乐相册则可以解决上述问题,提高浏览电子相册时的用户体验感.如何跨越音乐与图片之间的语义鸿沟是个很困难的问题.因为,一张图片和一段音乐是属于不同模态间的数据结构,其特征提取方式的不同使得不同模态特征间的维度往往不同,这将导致特征中所蕴含的信息无法直观地进行比较,因而具有异构性和不可比拟性,不能直接进行相似性计算,但是在情感语义上又相互关联,即不同模态的特征可以表征同一个情感语义概念,比如一张图片或者一段音乐都能同时感受出“高兴”或“悲伤”的感觉.由此可以看出多媒体时代的数据呈现出多模态数据结构复杂的特性[1],所以若要实现音乐相册的自动生成研究,就要实现跨模态检索,通过挖掘数据的潜在语义,将不同模态的数据信息投影到共同的隐藏语义空间中,并在该语义空间中利用不同模态数据间的相似度进行比较与检索,从而实现音乐相册的自动生成.

1 国内外研究现状

当人们欣赏了一定数量的音乐或浏览了一定数量的图片后难免会产生听觉和视觉疲劳,因此,如果能同时满足人们在视觉和听觉上

的需求,那么将会获得更好的用户体验.目前除了相册的发布与共享,一些软件例如 iPhoto 也能提供为相册选择背景音乐的功能.但是,让用户手动选择背景音乐也存在较大的缺点:若用户要为多个相册挑选背景音乐时,只能一一挑选着实有些麻烦,此外用户可能对音乐敏感度不高或者在较短时间内找不到合适的音乐,又或者用户正好有一首特别喜欢的音乐想为它添加一些情感比较接近的相册图片,然而却并没有这种途径.因此,若能自动生成音乐相册,即当用户在浏览相册时能同时欣赏到符合照片情境的背景音乐,就可获得更好的用户体验.

首先,音乐与图片所带来的共同体验,需要将音乐与图片进行相关性分析,即对音乐和图像特征的多模态融合分析.目前,针对各种多模态信息进行融合的方法已经广泛应用到了检索、分类、事件检测等多媒体领域中.Liu 等[2]使用分层检索结构融合音频、视频信息提高了在线视频检索的效率与准确率;Chen 等[3]利用文字信息与视觉空间的相互映射,消除用户在文字描述上的二义性,补全其在视觉空间上的信息,提升了垂直搜索的准确率;Jeon 等[4]建立 CRM(Cross-media Relevance Model)模型用于解决跨模态标注问题;Feng 等[5]在 CRM 模型的基础上使用多项伯努利分布估计图像与文本的概率分布提升了标注的准确率;Su 等[6]使用基于最近邻图学习模型,融合标签相似度以及图像相似度作为权值在图模型上传播及预测图像标签,同时还利用了图像与标签的相似度进行预测,在多个数据集上取得了不错的实验结果;Yang 等[7]基于分类器融合模型利用 3 种模态信息进行网络视频分类取得了不错的效果.

此外,音乐与图片所带来的共同体验,需要对视觉和听觉进行关联分析[8-9],本文主要研究以音乐为中心的关联.以音乐为中心的关联就是给定一定的音乐乐句,为它关联图片.目前,在播放音乐的同时能够生成简单图像的只有 Winamp 和微软的媒体播放器,但其视觉动画不一定与播放的音乐在情感上相关.Chen 等[10]提出一种音乐可视化系统,它在播放用户选取的音乐乐句的同时,播放一组基于视觉和听觉相似性的图像.其想法与本文大致相似,只是在实验时先对图像进行情感的分类,再将情感标签与音乐的情感相联系,并没有对其底层特征进行相似性研究.Xiang 等[11]挖掘美学能量作为媒介建立一个自动的图片浏览系统.美学能量的基本思想是“听见颜色,看见声音”.Hua 等[12]提出了一种家庭视频自动编辑系统.在这个系统中,用户可以指定一个音乐片段,然后系统会按一定的编辑规则自动提取一系列的视频片段.尽管上述研究在一定程度上将图像和音乐进行了关联,但这类系统的功能被限定在某些特定的情感空间,因为一个音乐片段包含了某些固定的情感.因此,照片的类型也总是收敛到一个特定的类型,并且一(音乐)对多(照片)的展示方式可能不会引起用户的兴趣.

本文提出采用多模态隐空间学习算法来解决音乐相册的自动生成问题.音乐相册的自生成其主要难点在于音乐和图片分属于 2 种媒体空间,这将导致特征中所蕴含的信息无法直观地进行比较,无法消除不同模态之间特征的异构性,为此需要找到某种方法来衡量两者之间的相关程度,是否包含了相同的情感,从而进行合适的推荐.为了解决这个问题,首先根据 MIREX 的情感分类标准,人工创建图像和音乐训练库,并对图像和音乐数据进行情感值的标注和分类;然后,通过多模态 LPP 算法,将音乐和图像特征映射到低维子空间,并分析两者之间的相关性,在每一种情感类别下,生成一种映射模型;输入测试音乐乐句样本,利用情感分类器进行情感分类后,输入到不同的映射模型,实现为音乐乐句推荐与其情感相近的图像.整个研究框架如图 1 所示.下面将分别介绍本文的音乐特征提取和图像特征提取方法,以及使用局部保持投影(Locality Preserving Projection,LPP)算法进行子空间映射的方法.

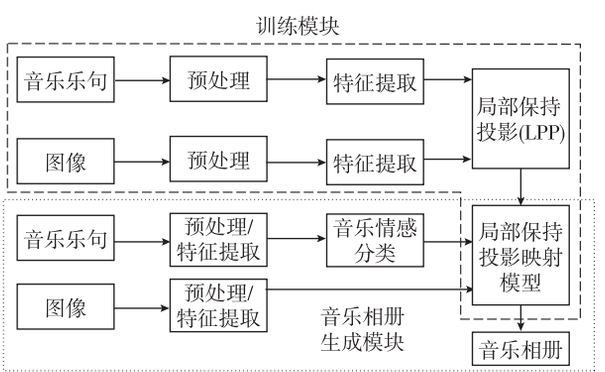

图 1　音乐相册自动生成研究框架

Fig. 1　Framework of music album generation

2　音视频特征提取

2.1　音乐特征提取

以前的音乐特征提取方法大多采用音乐片段级别的特征向量来描述音频.但是音乐所表达的情感

在整个音乐的演奏期间是有起伏的,因此音乐片段级别的特征表示方法无法准确刻画音乐所蕴含的真正情感.为了从音乐数据中挖掘出更加丰富的音乐情感信息,本文提出了句子级别的特征表示方法以刻画音乐的情感.本文主要分析图像与音频特征之间的典型相关性,根据文献[13]的研究结果,在提取典型的 Mel Frequency Cepstral Coefficients(MFCC)特征和在 Perceptual Linear Predictive (PLP)基础之上引出 Relative Spectral-Perceptual Linear Predictive (RASTA-PLP)特征.为了挖掘句子级别的音乐特征,本文根据每个句子的起始时间将音乐片段分割成多个音乐乐句,均有音频特征集 $\{v_1,v_2,v_3,\cdots,v_n\}$,计算该集合中所有特征向量的均值和方差,将此均值向量与方差向量拼接起来得到最终的句子级别音频特征.于是对于音乐乐句,取 20 阶 MFCC 系数特征,21 阶 PLP 频谱参数,9 阶 PLP 倒谱参数 RASTA-PLP,计算其均值和方差,获得 100 维的特征值,将其组成最终的特征向量来描述一个音乐乐句.

2.2 图像特征提取

在图像蕴含的众多信息中,最直观的是颜色特征.颜色特征相比于其他的视觉类特征,具有良好的稳定性,对大小和方向具有不敏感性,因而被普遍用于各类研究中.因此,在一般情况中,用颜色特征来表征一张图像比较方便并且具有重要意义,使得更多的学者更加深入地探讨了不同的颜色与其产生的不同情感之间的关系,获得了很大的收获.

从色调上,人们一般把颜色分成暖色和冷色.颜色冷暖其实和真实的温度并没有直接的联系,它只是人们心理上的一种感受.暖色,即为人们看到红色、黄色、橙色以及类似的颜色时,内心会产生愉快、调皮、温暖的感受,冷色即为在看到蓝色、紫色,白色及类似的颜色时,会产生一种清冷、高贵、神圣的感觉[14],而黑色让人想起阴暗、死亡,给人以肃穆、恐惧的感觉[15].

从饱和度上,人们对于不同饱和度的颜色也会产生不一样的感受.颜色的纯度越高,如大红、大绿等,给人的视觉冲击力越大,越会引发更加强烈的感官刺激.颜色越鲜艳,越能吸引人的注意.

Boyatzis 等[16]在观察儿童对情感的反馈中,像黄色、玫红色等暖色一般使儿童感觉积极和热情,像蓝色、绿色则容易让儿童产生平静和春意盎然之感,而如黑色、灰色等较深的颜色则会产生一些悲观之感.Itten[17]发现在艺术图像中,颜色与高阶情感语义有一定的联系,此外他还发现颜色的不同叠加会产生协调、不协调、亢奋或平和等效果.Hemphill[18]探索发现亮的颜色容易激发正面的情绪,而暗的颜色会激发负面的情绪.Saito[19]研究发现暗的颜色也可以引发正面和负面的情绪.另外,人对色性的感受也强烈受光线和邻近颜色的影响.色彩的冷暖感觉是人们在长期生活实践中由于联想而形成的.红、橙、黄色常使人联想起东方旭日和燃烧的火焰,因此有温暖的感觉,所以称为"暖色";蓝色常使人联想起高空的蓝天、阴影处的冰雪,因此有寒冷的感觉,所以称为"冷色";绿、紫等色给人的感觉是不冷不暖,故称为"中性色".色彩的冷暖是相对的.在同类色彩中,含暖意成分多的较暖,反之较冷.

考虑到颜色特征在图像情感研究中的重要性,本文选取颜色矩以及文献[20]提出的颜色对比度作为图像特征.具体特征抽取过程如下:将输入图像分成 5×5=25 张大小相等的子图,将每张子图的图像数据从 RGB(Red, Green, Blue)空间转换到 HSV(Hue,Saturation,Value)空间.颜色矩为计算每一张子图在 HSV 空间各个分量上的一阶矩(均值)、二阶矩(方差)和三阶矩(偏度).

此外,本文还提取了颜色对比度作为其特征之一,颜色对比空间 (OPP)计算公式如下:

$$\begin{aligned}&\text{red-green}: O_1=(r-g)/\sqrt{2};\\&\text{yellow-blue}: O_2=((r+g)-2b)/\sqrt{6};\\&\text{luminance}: O_3=(r+g+b)/\sqrt{3}.\end{aligned}\tag{1}$$

其中:r,g,b 为 RGB 颜色空间内任意像素点的 R、G、B 通道的值,取值范围为 0~1.颜色对比度定义如下:

$$C_{\text{ontrast}}=\sqrt{\frac{1}{N-1}\sum_{x=1}^{N}[(a_x-\bar{a})^2\times(b_x-\bar{b})^2]}.\tag{2}$$

其中:a 表示红色-绿色的颜色对比;b 表示蓝色-黄色的颜色对比;N 表示每张子图的像素;a_x,b_x 分别表示子图第 i 个像素点的 a 分量和 b 分量的值;$\bar{a}$ 和 $\bar{b}$ 分别表示 a 和 b 的平均值.颜色对比度描述图像中颜色之间的差别大小,即颜色的多彩程度.

通过该方法提取每张子图在色调、饱和度、明度分量上的均值、方差和偏度,以及颜色对比度,即 3×3×25=225 个颜色特征值和 25 个颜色对比度值,由此一张图像可以由 250 维特征向量来描述.

3 隐空间学习方法的分析

由于音乐特征与图片特征空间的异构性,使得直接挖掘这 2 个模态之间的相关性变得异常棘手.

本文提出了一种针对音乐情感分类的有监督多模态LPP隐空间学习方法,为音乐的不同模态数据学习一个共同的具有情感区分度的隐式空间,该空间不仅保持了原有空间的特性,并且拉近了同个情感类别中不同模态音乐数据间的距离,提高了特征在不同情感类别中的区分能力.

LPP是一种线性降维算法,该算法的特性是在线性投影之后仍可以保持特征在原始空间的局部特性,并将使得在原始特征空间中与新特征空间中的最近邻搜索结果相似.原始的LPP目标函数如下所示:

$$\begin{aligned}&\arg\min \boldsymbol{v}_t^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}_p\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}_t,\\&\text{s.t.}\quad \boldsymbol{v}_t^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{v}_t=1.\end{aligned}\tag{3}$$

其中$\boldsymbol{v}_t$表示一个线性变换向量,而$\boldsymbol{X}$表示样本的特征矩阵,每一行代表着一个样本.$\boldsymbol{L}_p=\boldsymbol{D}-\boldsymbol{W}$表示一个拉普拉斯(Laplacian)矩阵,其中$\boldsymbol{W}$是训练样本的距离度量矩阵,记录了训练样本间的两两距离,而$\boldsymbol{D}$是一个对角矩阵,对角线上的值是$\boldsymbol{W}$中每一列的和.

为了将LPP扩展成多模态模式,可以将式(3)改造为2个不同特征空间目标函数的联合优化:

$$\begin{aligned}&\boldsymbol{A}_1=-\boldsymbol{X}_1\boldsymbol{L}_{p1}\boldsymbol{X}_1^{\mathrm{T}},\\&\boldsymbol{B}_1=\boldsymbol{X}_1\boldsymbol{D}_1\boldsymbol{X}_1^{\mathrm{T}},\end{aligned}\tag{4}$$

$$\begin{aligned}&\boldsymbol{A}_2=-\boldsymbol{X}_2\boldsymbol{L}_{p2}\boldsymbol{X}_2^{\mathrm{T}},\\&\boldsymbol{B}_2=\boldsymbol{X}_2\boldsymbol{D}_2\boldsymbol{X}_2^{\mathrm{T}},\end{aligned}\tag{5}$$

$$\begin{aligned}&\arg\max \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{A}_1\boldsymbol{v}_1+\mu\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{A}_2\boldsymbol{v}_2,\\&\text{s.t.}\quad \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{B}_1\boldsymbol{v}_1+\gamma\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{B}_2\boldsymbol{v}_2=1.\end{aligned}\tag{6}$$

在此μ是一个正整数用于平衡2个目标,因为如果$\boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{A}_1\boldsymbol{v}_1$远大于$\boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{B}_2\boldsymbol{v}_2$则优化目标将向着优化$\boldsymbol{v}_1$的方向倾斜,反之亦然.通过优化上述联合目标函数,不仅将不同模态的异构数据映射到同个空间,并且将保持其在原有空间中的特性.

同时,为了学习一个更具类区分度的隐式空间,提高在隐式空间中的分类准确率,还需要使同个类中不同模态的样本相互接近,该目标可以通过最大化不同模态样本间的方差实现,可以通过解决如下最优化问题达到这个目标:

$$\arg\max \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{C}_1\boldsymbol{C}_2\boldsymbol{v}_2\tag{7}$$

其中,矩阵$\boldsymbol{C}_1\boldsymbol{C}_2$第$i$列应对应相同的元素,在这里设定矩阵$\boldsymbol{C}_k$的第$i$列表示第$i$个情感类所有训练样本的统计值(比如平均值),式(7)将使得属于同一个类的样本更接近的同时将不同类的样本隔离开来.

可以看出,若不对式(7)中$\boldsymbol{v}_1,\boldsymbol{v}_2$加以约束,其目标值可能会无止境的增长,所以将该式与式(6)结合,得到最终的目标函数:

$$\begin{aligned}&\begin{bmatrix}\hat{\boldsymbol{v}}_1\\\hat{\boldsymbol{v}}_2\end{bmatrix}=\arg\max\begin{bmatrix}\boldsymbol{v}_1\\\boldsymbol{v}_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{A}_1&\beta\boldsymbol{C}_1\boldsymbol{C}_2^{\mathrm{T}}\\\alpha\boldsymbol{C}_2\boldsymbol{C}_1^{\mathrm{T}}&\mu\boldsymbol{A}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1\\\boldsymbol{v}_2\end{bmatrix},\\&\text{s.t.}\quad\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}&\boldsymbol{v}_2^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{B}_1&0\\0&\gamma\boldsymbol{B}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1\\\boldsymbol{v}_2\end{bmatrix}=1.\end{aligned}\tag{8}$$

基于式(8)得到的最终投影向量$\boldsymbol{v}_1,\boldsymbol{v}_2$将平衡样本在原始空间的特征提取优化以及隐式空间中的方差得到一个最优化结果.式(8)还可以转换成如下向量表示:

$$\begin{aligned}&\begin{bmatrix}\hat{\boldsymbol{v}}_1\\\hat{\boldsymbol{v}}_2\end{bmatrix}=\arg\max\begin{bmatrix}\boldsymbol{v}_1\\\boldsymbol{v}_2\end{bmatrix}^{\mathrm{T}}\begin{bmatrix}\boldsymbol{A}_1&\beta\boldsymbol{C}_1\boldsymbol{C}_2^{\mathrm{T}}\\\alpha\boldsymbol{C}_2\boldsymbol{C}_1^{\mathrm{T}}&\mu\boldsymbol{A}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1\\\boldsymbol{v}_2\end{bmatrix},\\&\text{s.t.}\quad\begin{bmatrix}\boldsymbol{v}_1^{\mathrm{T}}&\boldsymbol{v}_2^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{B}_1&0\\0&\gamma\boldsymbol{B}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{v}_1\\\boldsymbol{v}_2\end{bmatrix}=1.\end{aligned}\tag{9}$$

通过上述目标函数,可以得到一个音乐特征、图片特征模态共同的隐式空间,该空间不禁消除了模态间的异构性,使得不同模态的特征可以进行相互计算,同时拉近了相同类的不同模态间的距离.以上特性令使用一个模态的数据补充另一个模态数据缺陷而提高音乐与图片匹配的准确率成了可能.

4 实验及分析

4.1 实验数据集

4.1.1 音乐数据集

根据MIREX(The Music Information Retrieval Evaluation eXchange)的情感分类标准,如表1所示,人工创建音乐训练库,并对音乐数据进行情感值的标注和分类.

表1 MIREX的5类情感分类标准

Table 1 Five emotion classes in MIREX

类别	中心情感词	其他情感词
1	Rowdy	Rousing, Confident, Boisterous, Passionate
2	Amiable	Sweet, Fun, Rollicking, Cheerful
3	Literate	Wistful, Bittersweet, Autumnal, Brooding, Poignant
4	Witty	Humorous, Whimsical, Wry, Campy, Quirky/Silly
5	Volatile	Fiery, Visceral, Aggressive, Tense, Intense

在音乐情感识别领域,目前还没有通用的中文音乐情感数据库,因此本文所有的训练测试数据都是自行搜集和筛选的.具体步骤如下:

1)找20位同学,根据表1的5类音乐情感描述,在百度音乐库中下载每类情感对应的歌曲,每人10首,并对音乐按句子级别进行分割,选取每首歌

中最能表达情感的5个音乐乐句,从而获得5个类别共计5 000个音乐乐句.

2)采用多人同时标注的办法,来增强音乐数据集的真实性.让20名同学对这5 000个音乐乐句进行标注,如果对同一音乐乐句有5人以上标注为同一情感的,则认定该音乐乐句属于此类情感,否则放弃该乐句数据.经此步骤筛选出了3 000个音乐乐句,每类600个.

3)让20位同学对步骤2)产生的音乐乐句进行VA(Valence,Arousal)值的标注,每人150个,为实验方便,VA值的范围取{-1,-0.8,-0.6,…,0,…,0.8,1},标注后再取平均值.

4)对标注后的3 000个音乐乐句进行筛选,划定每类情感的VA值范围如表2所示,删除超出范围的音乐乐句,选取2 000个音乐乐句作为实验数据集,每类400个.

表2 5类音乐情感的VA值范围

Table 2 VA values for five music emotion classes

类别	V值	A值
1	[-0.6,0.6]	[0.4,1]
2	[0.2,1]	[0,0.6]
3	[-1,-0.2]	[-0.6,0]
4	[0.2,1]	[-0.6,0]
5	[-1,-0.2]	[0,0.6]

5)用格式转换软件,将音乐片段统一为采样率16 kHz,wav格式,单声道.每次实验时,在每个类别中随机选取1 800个音乐乐句作为训练数据,测试时从剩下的200个音乐乐句中选择作为测试数据.另外在百度音乐库上任意下载40首歌曲,并按照句子级别进行分割得到200个音乐乐句,不进行任何情感标注处理,只摘取其中最体现情感的音乐乐句组成测试库.至此,音乐训练库共有1 800个音乐乐句,测试库有已知情感的音乐乐句200个和未知情感的音乐乐句200个.

4.1.2 图像数据集

IAPS图像库包含大多数情感的图像库,而艺术类图像集取自于一个艺术分享网站,所以这类图像的最初情感注释来源于分享的摄影师们.摄影师们通过对图像的构成、颜色和灯光等进行有意的操纵,从而激起人们某些特定的情感.为了实现基于典型相关分析的音乐相册自动生成研究,本文主要采用了这2个图像数据集:从共享网站上下载的艺术类图像集[21]和国际情绪图像系统(the International Affective Picture System,IAPS)[22]数据库.

为了与音乐的情感相对应,在图像上同样采用MIREX的5类情感分类标准.根据文献[23]提出的IAPS图像在Valance和Arousal轴上的映射,可以看出图像情感VA值所处范围为1~9,与音乐数据集的VA值相对应,本文界定出5类图像情感的VA值范围如表3所示.

表3 5类图像情感对应的VA值

Table 3 VA values for five image emotion classes

类别	V值	A值
1	[2.6,7.4]	[6.6,9]
2	[5.8,9]	[5,7.4]
3	[1,4.2]	[2.6,5]
4	[5.8,9]	[2.6,5]
5	[1,4.2]	[5,7.4]

根据表3,在图像库中摘录表1中所列5种情感类别下,满足该表标准的图像作为样本来构建图像数据集,图像数据集的情况如表4所示.训练集中共有图像样本450张,测试集中共有图像样本255张.

表4 图像训练数据集和测试数据集

Table 4 Image dataset for training and testing

类别	训练库样本/张	测试库样本/张
1	90	60
2	90	51
3	90	35
4	90	52
5	90	57
总计	450	255

4.2 实验评价指标

实验评价指标分为客观评价指标和主观评价指标.

客观评价指标采用查准率,定义为

$$\text{查准率}=\frac{\text{正确返回的图像数目}}{\text{返回的 Top}M\text{ 个图像}}\times 100\%. \quad (10)$$

其中"正确返回的图像数目"是指该返回图像与输入的音乐属于同一个情感类别.

主观评价指标采用邀请学生打分的方式,共邀请20名同学对实验结果进行评价.对于为歌曲推荐出的图像,所有这20名同学按以下规则做标记:

5分:认为实验推荐的所有图像都符合音乐所表达的情感,标记为5.

4分:认为实验推荐的所有图像中有80%符合

音乐所表达的情感,标记为 4.

3 分:认为实验推荐的所有图像中有 60%符合音乐所表达的情感,标记为 3.

2 分:认为系统推荐的所有图像中有 40%符合音乐所表达的情感,标记为 2.

1 分:认为系统推荐的所有图像中有 20%符合音乐所表达的情感,标记为 1.

0 分:认为系统推荐的所有图像中没有一张符合音乐所表达的情感,标记为 0.

定义每一类情感的满意度 r 为所有实验数据的标记分值的平均值所占的比例,计算如下:

$$r = \frac{\sum_{i}^{t} \tau_i}{t} \times 100\%, \tag{11}$$

其中 τ_i 为第 i 个音乐测试样本的平均得分,t 为音乐测试样本的总数.

4.3 实验结果及分析

本文共设计了 3 组实验并进行分析和统计,第 1 组是将 3 种算法在每一种情感类别下得到的对推荐图像平均满意度的对比;第 2 组是当输入测试音乐在情感已知和未知 2 种情况下,使用 LPP 方法获得的查准率对比;第 3 种是不同检索数量需求下,对实验查准率的影响.

在第 1 组实验中,将实验结果与以下 3 种方法进行比较:

1) Lower Bound (LB):输入几个音乐乐句,随机推荐本文测试图像数据集中的图像.由于是随机推荐,其实验结果应该作为本实验的下限.

2) Manually Selection (MS):输入几个音乐乐句,人工推荐测试图像数据集中的图像.

3) CCA:输入几个音乐乐句,采用经典 CCA 方法推荐图像.

表 5 显示了随机推荐的 LB 方法、本文的 LPP 方法和 CCA 方法、人工推荐的 MS 方法得到的满意度结果.可以看出,采用 LPP 方法在人工评价时得到了 72.06%的满意度,这一数值与随机推荐方法(满意度平均为 31.6%)和传统 CCA 方法相比有明显地提升,但与人工推荐的结果(满意度平均为 78.09%)还有一定差距,但差距并不大.因此,本文采用 LPP 方法确实能够提高音乐相册自动生成的效果,为用户推荐出一组满意度较高的图像.

第 2 组实验是不同检索数量需求下,对实验查准率的影响.

表 5 4 种图像推荐方法满意度对比

Table 5 Satisfaction percentage comparison for four image recommendation methods

方法	满意度/%
LB	31.6
CCA	69.45
LPP	72.06
MS	78.09

本文实验选取返回 10 张图片,是考虑到检索图像的数量可能会影响最终的查准率,因而对实验进行验证.在不同检索需求下,分析实验查准率的差别,实验结果如图 2 所示.可以看出,因为 LPP 能有效地描述音乐特征与图像特征之间的相关性,在实验返回 5 张图像时,查准率普遍较高.但由于数据过少,实验存在的偶然偏差也会更大,因而选择 10 张图像.

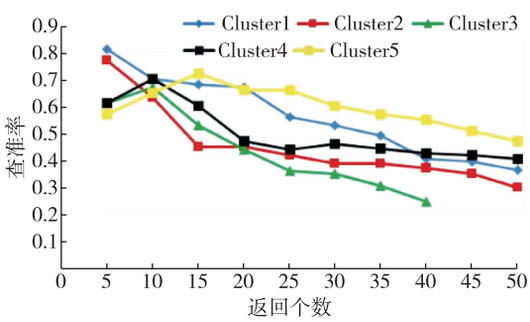

图 2 不同返回个数 M 下的查准率对比

Fig. 2 Accuracy comparison with different number of returned images

第 3 组实验是当输入测试音乐乐句在情感已知和未知情况下,使用 LPP 方法获得的查准率对比.

通过查准率来观察 LPP 方法的有效性,对测试音乐乐句在情感已知和未知 2 种情况下进行实验对比.实验返回 10 张与测试音乐乐句情感表达最接近的图像,查准率为返回的 10 张图像中与测试音乐乐句情感相同的图像所占比例.实验结果均取多次平均,如图 3 所示.可以看出,总体而言,LPP 的方法查准率高于纯 CCA 方法.在测试音乐乐句情感已知的情况下,本文提出的 LPP 方法的查准率相对较高,因为 LPP 在分析两者的相关性上,得出了较准确的映射模型,并且情感分类器对测试音乐的情感识别和分类方面效果较好.在测试音乐乐句情感未知的前提下,由于测试音乐需要先经过情感分类器进行情感的分类,再分别输入到不同的 LPP 映射模型,情感分类的偏差也会导致实验结果的偏差.

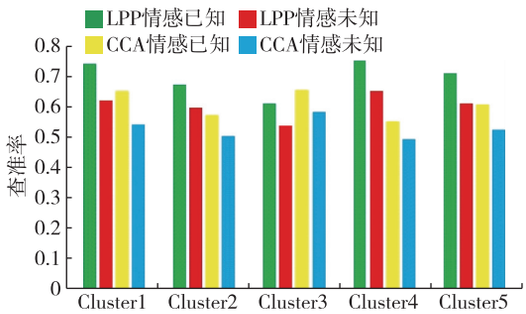

图 3 LPP 方法与 CCA 方法的查准率对比

Fig. 3 Accuracy comparison between LPP and CCA

5 结论与未来工作展望

本文主要进行了基于多模态融合的家庭音乐相册自动生成研究,采用 LPP 方法分析图像与音乐特征之间潜在的相关性,实现了为音乐推荐出与之情感表达相近的图像,并与人工推荐和随机推荐等方法进行比较,结果表明 LPP 方法是有效的,同时也表明本文提取的图像特征与音乐特征之间存在着一定的相关性.

未来可以继续展开的工作:

1)本文实验是在国际情绪系统 IAPS 数据集上完成的,主要是国外的图像,今后还应选取更多符合中国人审美的图像数据集进行测试,以减少人工评价或者人工标注时可能产生的误差.

2)本文的音乐情感数据库,仅仅依靠少数同学一起创建,获得的音乐库只代表了一部分人的意愿,并不具有权威性,所以希望在未来的研究中,可以创建一个更完整、更有权威的音乐情感数据库.

3)本文在音乐特征提取方面,只是选取了音频特征参数来描述音乐乐句,在以后的研究中,可以尝试结合歌词文本特征和音频特征进行多模态融合进行更多的实验与筛选,寻找更准确的特征来表达音乐的情感.

4)本文在图像特征提取时,提取的是图像的颜色特征,在以后的实验中可以尝试挖掘图像的纹理、形状特征等更多能表征图像情感的特征.

参考文献

References

[1] Zhang H, Zhuang Y T, Wu F. Cross-modal correlation learning for clustering on image-audio dataset[C]// ACM International Conference on Multimedia, 2007:273-276

[2] Liu W, Mei T, Zhang Y D, et al. Listen, look, and gotcha: Instant video search with mobile phones by layered audio-video indexing [C] // ACM International Conference on Multimedia, 2013:887-896

[3] Chen Y X, Yu N H, Luo B, et al. iLike: Integrating visual and textual features for vertical search[C]// ACM International Conference on Multimedia, 2010:221-230

[4] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models [C]// International ACM SIGIR Conference on Research and Development in Information Retrieval, 2003:119-126

[5] Feng S L, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation [C] // IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, DOI: 10.1109/CVPR.2004.1315274

[6] Su F, Xue L K. Graph learning on K nearest neighbours for automatic image annotation [C] // ACM International Conference on Multimedia Retrieval, 2015:403-410

[7] Yang L, Liu J, Yang X, et al. Multi-modality web video categorization [C] // ACM SIGMM International Workshop on Multimedia Information Retrieval, 2007: 265-274

[8] Hanjalic A. Extracting moods from pictures and sounds: Towards truly personalized TV [J]. IEEE Signal Processing Magazine, 2006, 23(2):90-100

[9] Wang J C, Yang Y H, Jhuo I H, et al. The acoustic visual emotion Gaussians model for automatic generation of music video [C] // ACM International Conference on Multimedia, 2012:1379-1380

[10] Chen C H, Weng M F, Jeng S K. Emotional-based music visualization using photos[C]// International Conference on Advances in Multimedia Modeling, 2008:358-368

[11] Xiang Y Y, Kankanhalli M S. A synesthetic approach for image slideshow generation [C] // IEEE International Conference on Multimedia & Expo, 2012:985-990

[12] Hua X S, Lu L, Zhang H J. Optimization-based automated home video editing system[J]. IEEE Transactions on Circuit and Systems for Video Technology, 2004, 14(5): 572-583

[13] 查美丽.基于情感的音乐分类系统的研究与实现[D].南京:南京邮电大学通信与信息工程学院,2014
ZHA Meili. The research and realization of music classification system based on emotion[D]. Nanjing: College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, 2014

[14] Hayashi T, Hagiwara M. Image query by impression words: The IQI system [J]. IEEE Transactions on Consumer Electronics, 1998, 44 (2):347-352

[15] 古大治.色彩与图形视觉原理[M].北京:科学出版社,2000
GU Dazhi. Color and graphic visual principles [M]. Beijing: Science Press, 2000

[16] Boyatzis C J, Varghese R. Children's emotional associations with colors [J]. Journal of Genetic Psychology, 1994, 155(1):77-85

[17] Itten J. Art of color (Kunst der Farbe)[M]. Ravensburg, Germany: Otto Maier Verlag, 1961

[18] Hemphill M. A note on adults' color-emotion associations

[J]. Journal of Genetic Psychology, 2010, 157(3): 275-281
[19] Saito M.Comparative studies on color preference in Japan and other Asian regions, with special emphasis on the preference for white[J].Color Research and Application, 21(1):35-49
[20] Ruiz-Del-Solar J, Jochmann M. On determining human description of textures[C]// Proceedings of SCIA 2001 Scandinavian Conference on Image Analysis, 2001: 288-294
[21] Cuthbert B N,Lang P J,Bradley M M.International affective picture system (IAPS):Affective ratings of pictures and instruction manual [R]. Technical Report of University of Florida,2008
[22] Yanulevshaya V,Van Gemert J C,Roth K.Emotion alvalence categorization using holistic image features[C]// IEEE International Conference on Image Processing, 2008:101-104
[23] Rao M A,Vazquez D,Lopez A M.Opponent colors for human detection[J].Iberian Conference on Pattern Recognition and Image Analysis,2011,6669:363-370

Automatic generation of family music album based on multi-modal fusion

LIU Junfang[1] SHAO Xi[1]
1 College of Telecommunications and Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003

Abstract With the development of the big data and social network,electronic albums and online services have become basic uses of computers and the Internet.Especially in recent years,the number of electronic albums has exploded with the popularity of social network.So how to improve the user experience of music album becomes particularly important.A photo album with certain topic usually has some emotion information.This paper studies the problem of automatic generation of family music album based on multi-modal fusion,so that users can enjoy music when browsing album photos with matched emotion.According to the emotions in music and images,the representative sentence-level features both for music and images are selected,and the LPP (Locality Preserving Projection) is employed to study the relevance between the music and the images in the same emotion.The image feature and the music feature are mapped into the latent space with more emotional classification ability to realize the automatic generation of music album.In the experiments,the objective evaluation result shows that the LPP method is higher than pure CCA (Canonical Correlation Analysis) method in precision;and in the subjective evaluation,the proposed LPP method achieves 72.06% at satisfaction level,which is close to the results of manually recommended approach (78.09%) and is higher than the results of randomly recommended approach and pure CCA approach.

Key words music album;emotion model;sentence-level;multi-modal fusion;latent space