

一种基于内容相关性的跨媒体检索方法

张 鸿^{1),2)} 吴 飞²⁾ 庄越挺²⁾ 陈建勋¹⁾

¹⁾ (武汉科技大学计算机科学与技术学院 武汉 430081)

²⁾ (浙江大学人工智能研究所 杭州 310027)

摘 要 针对传统基于内容的多媒体检索对单一模态的限制,提出一种新的跨媒体检索方法.分析了不同模态的内容特征之间在统计意义上的典型相关性,并通过子空间映射解决了特征向量的异构性问题,同时结合相关反馈中的先验知识,修正不同模态多媒体数据集在子空间中的拓扑结构,实现跨媒体相关性的准确度量.实验以图像和音频数据为例验证了基于相关性学习的跨媒体检索方法的有效性.

关键词 跨媒体检索;异构性;典型相关性;子空间映射;相关反馈

中图法分类号 TP391

Cross-Media Retrieval Method Based on Content Correlations

ZHANG Hong^{1),2)} WU Fei²⁾ ZHUANG Yue-Ting²⁾ CHEN Jian-Xun¹⁾

¹⁾ (College of Computer Science & Technology, Wuhan University of Science & Technology, Wuhan 430081)

²⁾ (Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027)

Abstract Most traditional content-based multimedia retrieval methods are designed for multimedia data of single modality. Such methods include image retrieval, audio retrieval, video retrieval, etc. This paper proposes a novel cross-media retrieval approach, which can process multimedia data of different modalities and measure cross-media similarity, such as image-audio similarity. First statistical method is used to learn canonical correlations between low-level feature spaces of different modalities. Then, sub-space mapping is designed to build an isomorphic subspace and solve the heterogeneity problem between different low-level feature vectors. This subspace contains media objects of different modalities, and each media object is represented with isomorphic vector. Since canonical correlations among multimedia objects are furthest preserved during the mapping process, cross-media similarity can be estimated with defined distance metric. Furthermore, relevance feedback provided by users is utilized to learn prior knowledge and refine multimedia topology in the subspace. In this way cross-media similarity is more consistent with human perception with the incorporation of user interaction. Both image and audio data are selected for experiments and comparisons. Given the same visual and auditory features the new approach outperforms ICA, PCA and PLS methods both in precision and recall performance. Overall cross-media retrieval results between images and audios are very encouraging.

Keywords cross-media retrieval; heterogeneity; canonical correlation; subspace mapping; relevance feedback

收稿日期:2006-06-16;最终修改稿收到日期:2007-12-27. 本课题得到国家自然科学基金(60525108,60533090)、国家科技支撑计划课题(2006BAH02A13-4)、国家“八六三”高新技术研究发展计划(2006AA010107)、高等学校科技创新工程重大项目培育资金项目、长江学者和创新团队发展计划(IRT0652)资助. 张 鸿,女,1979年生,博士,讲师,主要研究方向为多媒体分析与检索、机器学习. E-mail: zhanghong_zju@yahoo.com.cn. 吴 飞(通信作者),男,1973年生,博士,副教授,主要研究方向为多媒体分析与检索、统计学习理论. 庄越挺,男,1965年生,博士,教授,主要研究领域为多媒体数据库、人工智能、基于内容的多媒体检索、视频动画等. 陈建勋,男,1957年生,博士,教授,主要研究领域为基于 Web 的多媒体检索和计算机图形学等

1 引言

基于内容的多媒体检索是计算机视觉和信息检索领域的研究热点,其根据视觉、听觉或者几何等底层特征进行相似度匹配而实现检索^[1]. 20 世纪 90 年代初期人们提出了基于内容的图像检索技术,从图像中提取颜色、纹理、形状等底层视觉特征作为图像的索引^[2]. 这种技术后来也被运用到视频检索^[3]和音频检索^[4],只是针对不同媒体内容所采用的底层特征也不同. 这些研究都取得了一定成果,提交任何一种模态的媒体对象例子都可检索到同种模态的相似结果.

然而,现有的多媒体检索系统大都只能检索包含单一模态的多媒体数据库,或虽能处理多模态的媒体数据,却不支持跨媒体的检索,即根据一种模态的多媒体对象检索到其他模态的多媒体对象,例如:以描述“松鼠”外貌的图像为查询例子,检索记录了“松鼠”叫声的音频数据. 早在 1976 年,McGurk 就已经揭示了人脑对外界信息的认知需要跨越和综合不同类型的信息,来自视觉、听觉等不同感官的信息相互刺激、共同作用而形成整体性的理解^[5]. 因此,目前迫切需要研究一种支持不同模态的跨媒体检索方法,突破传统基于内容的多媒体检索对检索模态的限制.

本文以图像视觉特征和音频听觉特征之间的典型相关性分析(Canonical Correlation Analysis, CCA)^①为基础,通过子空间映射解决了不同模态数据特征异构性问题,并提出基于增量学习的相关反馈方法,实现跨媒体相关性的准确度量. 虽然 CCA 方法已经成功应用到图像语义自动标注^[6]和基于内容的图像检索^[2]等领域,但这些研究大多需要文本信息的辅助,而文本数据本身就代表了一定的语义. 因此,对于图像和音频这种非结构化、难以应用文本描述的多媒体数据,需要研究新的方法以挖掘两者间所蕴涵的相关性. 本文的方法实现了图像和音频之间的跨媒体检索,能够通过底层异构的视觉特征和听觉特征较好地理解高层的跨媒体语义关系,在检索过程中实现不同模态之间的灵活跨越.

2 挑战及相关研究

跨媒体检索是基于内容的多媒体检索中一个新的研究领域,目前国际上还没有较成熟的跨媒体检

索算法和技术. 跨媒体检索需要处理不同模态的媒体数据,例如:一个 500 维的视觉特征向量和一个 650 维的听觉特征向量,两者可能都表达了相似的语义概念,如爆炸和画面与爆炸的声音,但是计算机却很难根据两个特征向量度量两者在语义层面上的相关程度. 以图像和音频为例,跨媒体检索面临的主要挑战包括:

(1) 图像视觉特征与音频听觉特征之间不但维数不同,而且具有不同属性,这种异构性造成跨媒体的相关性度量十分困难;

(2) 即使解决了特征异构性问题,还需要进一步缩小底层特征与高层语义之间的鸿沟,以提高跨媒体检索精度.

近年来,一些研究者先后提出了类似跨媒体检索思想的研究,挖掘不同模态之间的相关性,与本文的跨媒体内容相关性学习有相似之处,这些方法主要分为以下两类:

(1) 基于相关性学习的自动语义标注和检索. 文献[6]通过学习图像和关键字之间的典型相关性,以解决图像自动标注问题;类似地,有研究者根据典型相关性实现基于关键字的图像检索^[2];除了典型相关性之外,还有其他的相关性挖掘研究,如文献[7]将 Web 图像的视觉特征和伴随文字特征看成两种不同模态,根据两者之间的标注关系非线性地在图像相似度矩阵和文本相似度矩阵间传递相关性,以提高 Web 图像检索效率.

(2) 多模态特征关联和索引. 视频内容同时包含不同模态的底层特征,许多研究通过关联分析和互索引等方法帮助理解视频语义. 文献[8]提出一种主题结构模型以组织不同模态的新闻内容,在新闻中出现频率较高的语义概念和多媒体对象之间建立关联,使得用户可以查询不同模态的新闻信息;文献[9]通过挖掘多模态特征进行视频数据库的索引和检索.

上述研究分析了不同模态、异构特征之间的相关性,用来缩小语义鸿沟、提高检索效率. 然而,其中大部分方法是为了提高单模态检索效率(如图像检索、视频检索),而不是解决不同模态之间的相关性匹配(如图像和音频之间). 另一方面,一些研究在实现不同模态之间的交叉检索时依赖于相关的文本数据(如基于关键字的图像检索),而文本自身就代表了一定的语义信息. 因此,这些方法难以有效应用到

① Magnus Borga. Canonical correlation a tutorial. January 12, 2001. <http://people.imt.liu.se/~magnus/cca/tutorial/>

缺乏文本描述的跨媒体检索中,如直接根据视觉和听觉特征度量图像和音频数据在语义上的相似度。

3 视觉和听觉特征的相关性保持映射

不同模态的媒体数据在底层特征匹配上面临异构性和不可度量挑战,如果采用传统的特征分析方法,如主成分分析(Principal Component Analysis, PCA)^[10]、独立成分分析(Independent Component Analysis, ICA)^[11]和偏最小二乘法(Partial Least Squares)^[12],需要分别对不同模态的特征矩阵进行降维,从而得到两个维数相同的子空间.但这样会丢弃不同模态在底层特征上潜在的相关性信息,使得降维得到的子空间难以准确反映高层语义联系,造成跨媒体检索效率较低。

与上述方法不同,本节通过典型相关性分析同时对视觉特征矩阵和听觉特征矩阵进行相关性求解和子空间映射,解决异构性和不可比性问题,更为重要的是映射得到的子空间在最大程度上保证了视觉和听觉特征之间的典型相关性不变。

3.1 相关性学习

相同语义、不同模态的媒体数据在底层特征上具有潜在相关性,例如,“松鼠”图像的视觉特征和“松鼠”音频的听觉特征在统计意义上存在一定相互关联.本节采用典型相关分析(Canonical Correlation Analysis, CCA)方法挖掘这种不同模态之间的典型相关性。

两个变量场 \mathbf{X} 与 \mathbf{Y} 之间的相关性定义如下:设有 n 个样本、 p 个变量组成的变量场,记为 $\mathbf{X}_{(n \times p)}$,另有 n 个样本、 q 个变量组成的变量场 $\mathbf{Y}_{(n \times q)}$,以最大限度地提取 \mathbf{X} 与 \mathbf{Y} 之间相关性的主要特征为准,从 \mathbf{X} 中提取组合变量 \mathbf{L} ,从 \mathbf{Y} 中提取组合变量 \mathbf{M} ,如下所示:

$$\begin{aligned} \mathbf{X}_{(n \times p)} &\xrightarrow{\mathbf{W}_X(p \times m)} \mathbf{L}_{(n \times m)}; \\ \mathbf{Y}_{(n \times q)} &\xrightarrow{\mathbf{W}_Y(q \times m)} \mathbf{M}_{(n \times m)} \quad (m < p, m < q) \end{aligned} \quad (1)$$

其中, $\mathbf{W}_X, \mathbf{W}_Y$ 为空间特征向量,又称为典型变量.按式(1)把具有较多个变量的变量场 \mathbf{X} 与 \mathbf{Y} 之间的相关化为较少组合变量 \mathbf{L} 与 \mathbf{M} 间的相关,通过 $\mathbf{W}_X, \mathbf{W}_Y$ 的数值分布来确定 \mathbf{X} 与 \mathbf{Y} 的空间相关分布形式,而 $\mathbf{W}_X, \mathbf{W}_Y$ 的数值大小则表示了所对应变量的重要程度.于是问题归结为如何求解典型变量 $\mathbf{W}_X, \mathbf{W}_Y$. 定义相关系数为 $\rho = r(\mathbf{L}, \mathbf{M})$,在式(3)的约束下,使相关系数最优化。

$$\rho = r(\mathbf{L}, \mathbf{M}) = \frac{\mathbf{W}_X^T \mathbf{C}_{XY} \mathbf{W}_Y}{\sqrt{\mathbf{W}_X^T \mathbf{C}_{XX} \mathbf{W}_X \mathbf{W}_Y^T \mathbf{C}_{YY} \mathbf{W}_Y}} \quad (2)$$

$$v(\mathbf{L}) = \mathbf{L}^T \mathbf{L} = \mathbf{W}_X^T \mathbf{X}^T \mathbf{X} \mathbf{W}_X = 1;$$

$$v(\mathbf{M}) = \mathbf{M}^T \mathbf{M} = \mathbf{W}_Y^T \mathbf{Y}^T \mathbf{Y} \mathbf{W}_Y = 1 \quad (3)$$

其中式(2)的 \mathbf{C}_{XY} 表示 $\mathbf{X}_{(n \times p)}$ 和 $\mathbf{Y}_{(n \times q)}$ 构成的协方差矩阵.结合式(2)和(3),使用拉格朗日乘法可以得到 $\mathbf{C}_{XY} \mathbf{C}_{YY}^{-1} \mathbf{C}_{YX} \mathbf{W}_X = \lambda^2 \mathbf{C}_{XX} \mathbf{W}_X$,即将最优化问题转换为形如 $\mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x}$ 的特征根问题,并进一步根据式(1)得到最小变量组合 $\mathbf{L}_{(n \times m)}, \mathbf{M}_{(n \times m)}$,以最大限度地揭示 $\mathbf{X}_{(n \times p)}, \mathbf{Y}_{(n \times q)}$ 之间的相关性。

3.2 同构子空间的映射

给定多个语义类别的图像和音频作为训练数据,设已知语义类别的个数为 z ,未知每幅图像和每段音频例子与语义类别之间的所属关系,可以采用如下所示的半监督式相关性保持映射方法构建同时容纳图像和音频对象同构子空间 S^* 。

半监督式相关性保持映射。

1. 对每个语义类别 $C_i (i \in [1, z])$, 随机选择一些图像 A_i 和音频 B_i 进行语义标注;
2. 分别求出 A_i, B_i 聚类质心^[13] $CtrA_i, CtrB_i$;
3. 分别以 $CtrA_i, CtrB_i$ 为初始质心对图像数据集和音频数据集进行 K-Means 聚类^[14];
4. 聚类结果中与初始聚类质心 $CtrA_i$ 划分到相同类别的图像被赋予与 $CtrA_i$ 相同的语义;
5. 聚类结果中与初始聚类质心 $CtrB_i$ 划分到相同类别的音频被赋予与 $CtrB_i$ 相同的语义;
6. 对每个语义类别 C_i 中所有图像和音频数据提取视觉特征矩阵 \mathbf{X} 和听觉特征矩阵 \mathbf{Y} , 计算 \mathbf{X}, \mathbf{Y} 之间的典型变量,以此为基础向量映射得到低维子空间。

上述方法在只对少量图像和音频数据进行语义标注的情况下,通过 K-Means 聚类划分语义类别,分别求取每个类别的视觉和听觉典型变量,将典型变量映射得到的子空间命名为 CCA 子空间 S^* 。

4 CCA 子空间中的跨媒体检索

4.1 不同模态间的相关性度量

设 $\mathbf{x}_i = (x_{i1}, \dots, x_{ik}, \dots, x_{ip}) (x_{ik} \in R)$ 表示初始的视觉特征向量, $\mathbf{y}_j = (y_{j1}, \dots, y_{jk}, \dots, y_{jq}) (y_{jk} \in R)$ 表示初始的听觉特征向量.经过半监督式的相关性保持映射后生成大量复数,定义 \mathbf{x}_i 经过子空间映射后的向量为 $\mathbf{x}'_i = (x'_{i1}, \dots, x'_{ik}, \dots, x'_{im}) (x'_{ik} = a + b \times i, (a, b \in R))$,同理可得 \mathbf{y}_j 对应 CCA 子空间中的映射结果 \mathbf{y}'_j 。

由于存在大量复数而无法直接在 CCA 子空间

S^* 中计算距离,因此,将子空间中每一维上的坐标值转换为极坐标形式:

$$\begin{aligned} x'_{ik} &= (\beta_{ik}, |x'_{ik}|) \\ \beta_{ik} &= \arctg(b/a), |x'_{ik}| = \sqrt{a^2 + b^2} \end{aligned} \quad (4)$$

对 y'_j 也用式(4)的方法进行变换,则图像 x'_i 和音频 y'_j 之间的距离定义为每一维上极坐标距离的平方和的 2 次方根,即

$$\begin{aligned} CCAdis(x'_i, y'_j) &= \text{sqrt} \sum_{k=1}^m (|x'_{ik}|^2 + |y'_{jk}|^2 - \\ &2 \times |x'_{ik}| \times |y'_{jk}| \times \cos|\beta_{ik} - \beta_{jk}|) \end{aligned} \quad (5)$$

从而,对于用户提交的图像查询例子 R ,可以采用 $CCAdis$ 计算子空间中图像 R 与音频对象之间的距离以衡量跨媒体相关性大小。然而,由于语义鸿沟的存在,CCA 子空间 S^* 的映射过程虽然保留了视觉和听觉特征间的典型相关性,但是 $CCAdis$ 的计算结果不能准确反映整个数据集范围内的跨媒体语义关系。因此,需要对 $CCAdis$ 的结果进行修正,定义修正后的跨媒体相关性为

$$CrossCor(x'_i, y'_j) = CCAdis(x'_i, y'_j) + \gamma(x'_i, y'_j) \quad (6)$$

其中 $\gamma(x'_i, y'_j)$ 为修正因子,表示子空间中不同模态样本之间 $CCAdis$ 与真实的跨媒体语义关系之间的差值。 $\gamma(x'_i, y'_j)$ 初始化为 0,并在基于增量学习的相关反馈过程中通过提取用户交互中的先验知识进行更新。

4.2 基于增量学习的相关反馈

相关反馈方法的使用可以结合用户的感知先验知识,以修正查询向量和整个数据集的拓扑关系,从而提高查询效率。本节提出的基于增量学习的跨媒体相关反馈作用于 CCA 子空间 S^* ,而不是初始的视觉和听觉特征空间。因此,子空间 S^* 中数据集的分布关系直接影响反馈算法的设计和效率。

子空间 S^* 是基于相关性保持映射而得到的,这种相关性保持特性使得图像和音频数据在子空间中形成一定的聚类效果(实验部分将给出子空间中单模态数据的聚类分析),因此我们有如下假设:

假设. 在子空间 S^* 中,相似语义、相同模态的媒体对象分布在比较集中的区域。

基于上述假设,本节以增量学习方式传播相关反馈中的跨媒体语义信息,修正图像和音频数据集在 CCA 子空间中的拓扑结构,同时更新修正因子 γ 的取值,使得式(6)的计算结果更准确地反映图像和音频对象在语义上的跨媒体相关程度。设 R 为提交的图像查询例子,用户对返回的音频例子进行评判,

得到音频正例集合 P 和音频负例集合 N ,相关反馈算法描述如下。

相关反馈算法:

1. $\forall p_i \in P$,调用 $CCAdis$ 找到 p_i 在音频数据库中的 k -近邻 $T = \{t_1, \dots, t_j, \dots, t_k\}$,并按距离进行升序排列;
2. 令 $\gamma(R, p_i) = -\tau(\tau > 0)$,以等差的方式依次修改集合 T 中每个元素对应的修正因子 γ 值: $\gamma(R, t_j) = -\tau + j \times d_1 (d_1 = \tau/k)$;
3. $\forall n_i \in N$,调用 $CCAdis$ 找到 n_i 在音频数据库中的 k -近邻 $H = \{h_1, \dots, h_j, \dots, h_k\}$,并按距离进行升序排列;
4. 令 $\gamma(R, n_i) = \tau(\tau > 0)$,以等差的方式依次修改集合 H 中每个元素对应的修正因子 γ 值: $\gamma(R, h_j) = \tau - j \times d_2 (d_2 = \tau/k)$;
5. 根据式(6)重新计算与查询例子 R 相似的音频对象,作为新的查询结果返回。

为了更好地说明上述相关反馈机制修正不同模态的数据集在子空间中的拓扑结构,以松鼠、汽车、鸟类三个语义类别的图像和音频为例进行描述:CCA 子空间中同时保留了松鼠、汽车、鸟类三个语义类别中的跨媒体典型相关性,这些相关关系在子空间中可能会产生“交叉”或“重叠”,使得在修正因子 $\gamma=0$ 的情况下式(6)找到的与鸟类音频之间距离最小的图像有可能是一幅汽车的照片;相关反馈过程中重新对修正因子 γ 进行赋值,使得式(6)度量的鸟类音频与鸟类图像之间的距离变小,而鸟类音频与汽车图像之间的距离变大,从而更加符合高层的跨媒体语义关系。

4.3 新媒体对象在 CCA 子空间中的定位

如果查询例子不在数据库中,则此查询例子定义为“新”媒体对象。同构子空间映射算法不能对单一的媒体对象分析相关性并降维映射。为了实现“新”媒体对象在 CCA 子空间中的定位,需要结合用户反馈中的先验知识。设“新”媒体对象为 Z ,如果可以准确计算出 Z 的 CCA 坐标,则以 Z 为查询例子的跨媒体检索可以用上述方法实现。 Z 的 CCA 坐标的计算如下:

(1) 提取 Z 的底层特征,使用欧氏距离,检索与 Z 同模态的媒体对象数据库,找到 Z 的 K -近邻作为返回结果;

(2) 用户标注两个反馈正例 $\{y_1, y_2\}$,设 $y_j (j=1, 2)$ 的 CCA 坐标表示为 $y_j = (y_{j1}, y_{j2}, \dots, y_{jm})$,则 Z 的 CCA 坐标为 $Z = \{z_1, \dots, z_k, \dots, z_m\}$,其中 $z_k = \text{Mean}(y_{1k} + y_{2k})$ 。

此外,还可以根据反馈正例对应的典型变量实现 Z 的子空间坐标映射。

5 实验结果与分析

为了验证上述算法的有效性,我们在 Win XP 下用 VC6.0 实现了一个原型系统,支持图像和音频间的跨媒体检索.实验数据集包括 10 个语义(鸟类、狗、汽车、爆炸、老虎、飞机等等)的多媒体对象,每个语义类别中分别有 100 幅图像和 70 段音频数据,其中 60 幅图像和 60 段音频例子作为训练数据,其余共 400 幅图像和 100 段音频数据作为“新”媒体对象.实验中的图像来自于 Corel 图像库,并从课题组参加 TRECVID2006 竞赛^[3]中搜集了与图像数据相应的音频例子,以进行图像-音频之间的跨媒体检索实验.

从图像数据集提取的视觉特征有 HSV 颜色直方图、颜色聚合向量(color coherence vector)和 Tamura 纹理.从音频数据集提取的听觉特征包括质心(centroid)、衰减截止频率(folloff)、频谱流量(spectral flux)和均方根(root mean square),这些特征在一定程度上综合反映了音频数据所具有的短时间内平稳、长时间动态变化的特性.音频是时序数据,所收集的音频例子持续时间均不超过 8s,不同持续时间的音频例子中提取的听觉特征向量的维数也不同,本文采用模糊聚类算法^[4],对音频数据集中提取的所有特征向量统一降维,得到维数相同的音频例子索引.

以下实验结果中的“平均”是指分别在每个语义类别中随机选择了 10 个不同的查询例子,得到检索结果的平均值.

5.1 不同方法得到的跨媒体检索结果

为验证本文方法对图像和音频两种不同模态之间跨媒体检索的有效性,实验根据第 3 节的方法分析视觉特征和听觉特征之间的典型相关性,并提取典型变量,映射得到保持相关性的 CCA 子空间 S^* ,用式(5)计算图像和音频在子空间中的距离,得出在没有相关反馈情况下的跨媒体检索结果.

实验与传统的 PCA、ICA 和 PLS 方法做了对比,分别用这三种方法通过相同的降维映射步骤实现跨媒体检索,过程如下:(1) 计算视觉特征矩阵的子空间基向量,映射得到子空间 S_1 ; (2) 同样将听觉特征向量都映射到与 S_1 相同维数的子空间 S_2 中; (3) 根据图像和音频在 S_1, S_2 中的坐标计算两者间的欧氏距离,以度量跨媒体相关性从而实现检索.

图 1 列出了本文的方法与传统 PCA、ICA 以及

PLS 方法得到的跨媒体检索结果,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计算^[7].

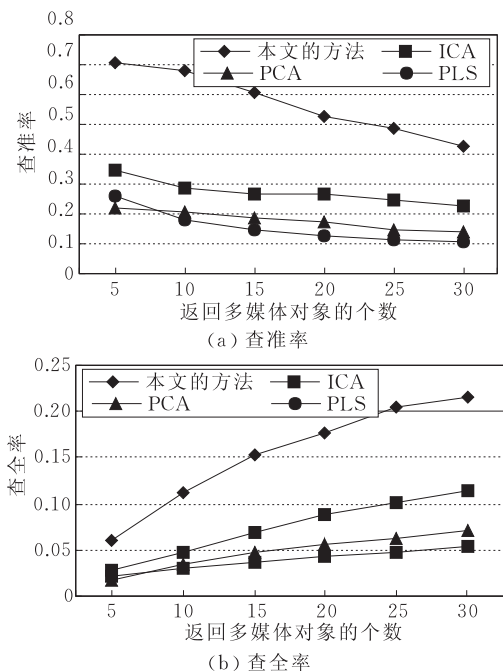



图 1 没有相关反馈时的跨媒体检索结果对比

图 1 中的结果是以图像为查询例子检索音频和以音频为查询例子检索图像得到的平均值.可见,在选择相同的视觉和听觉特征作为输入的情况下,本文方法优于传统的 PCA、ICA 和 PLS 方法.这是因为典型变量的计算过程是根据视觉和听觉特征的协方差矩阵分析潜在的跨媒体相关性信息,从而映射得到的 CCA 子空间 S^* 可以更好地反映高层的语义关系;而传统的 PCA、ICA 和 PLS 方法虽然已证明在处理单一模态的特征矩阵时十分有效,但是难以挖掘两种不同的特征矩阵之间的潜在关联.

图 2 是一个具体的跨媒体检索例子,其中输入

输入:  一个 5.3s 的汽车音频
返回的前 15 个相似图像:

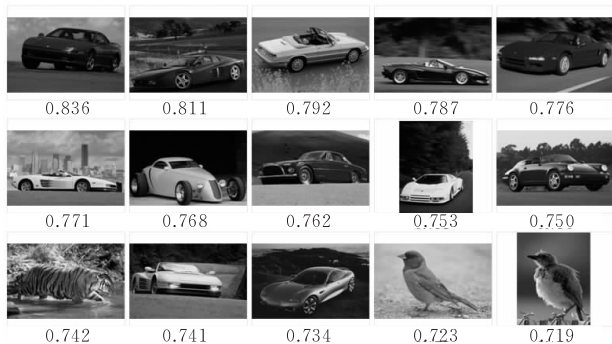


图 2 以汽车的音频为查询例子返回的相似图像

为一段 5.3s 的汽车音频,系统根据本文的方法计算相关性大小(见图 2 中每幅图像下方的数字),并返回前 15 个相似图像.可见,返回结果中有 12 幅图像与音频查询例子描述了相同语义.

5.2 相关反馈对跨媒体检索性能的改善

实验在每轮反馈时分别提供 2 个反馈正例和 2 个反馈负例,并设定基于增量学习的相关反馈算法中参数 τ 为

$$\tau = \text{Max}(\text{CrossCor}) - \text{Min}(\text{CrossCor}) \quad (7)$$

由于在新一轮反馈之后 $\text{CrossCor}(x_i, y_j)$ 的值随着 $\gamma(x_i, y_j)$ 的改变而更新(见第 4 节中式(6)),因此参数 τ 可以根据不同的反馈情况而动态更新.

图 3 显示了当返回结果个数固定为 15 时,随着相关反馈中用户交互的不断融入,返回结果中正确结果个数的变化过程,包括以音频为查询例子检索图像(I-by-A)和以图像为查询例子检索音频(A-by-I)两部分.

可以看到,经过两次相关反馈 I-by-A 和 A-by-I 得到的正确结果个数分别比反馈之前提高了 44.9% 和 24.2%,当反馈次数大于等于 3 时,跨媒体检索结果趋于稳定.由此可见,本文的方法能够快速学习,并修正图像与音频数据集的拓扑结构,从而有效地提高跨媒体检索效率.

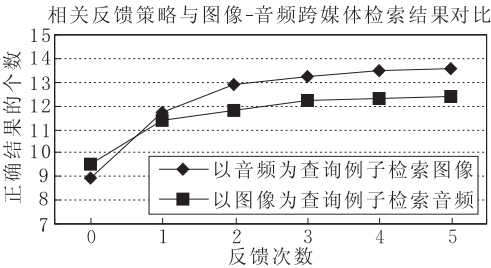


图 3 相关反馈对跨媒体检索性能的改善

5.3 相关性保持映射对单模态数据的聚类检测

上述跨媒体检索结果反映的是全局范围上图像和音频数据集之间的拓扑关系,而在 CCA 子空间 S^* 中的单模态检索效率则取决于相关性保持映射得到的局部(即图像数据集内部和音频数据集内部)聚类效果.全局数据关系和局部数据关系在一定程度上相互影响,因此,单模态检索虽然不是本文的研究重点,但是为保持实验的完整性,除了验证全局意义上跨媒体检索的有效性之外,实验还从局部意义上说明了跨媒体检索可达到较好检索性能的原因.

图 4 显示了在相关性保持映射得到的 CCA 子空间 S^* 中根据式(5)得到的图像检索和音频检索

结果.

当返回结果个数为 35 时,图像检索和音频检索中分别返回 25.5 和 29.1 个正确结果,这说明了:相关性保持映射不但将初始的图像和音频特征向量投影到一个同构的子空间中,而且变换后的特征向量在各自的单模态内部形成较好的聚类效果.这也验证了 4.2 节中在介绍相关反馈算法之前给出的假设,这也是全局范围内的跨媒体检索和相关反馈策略得以有效实施的前提.

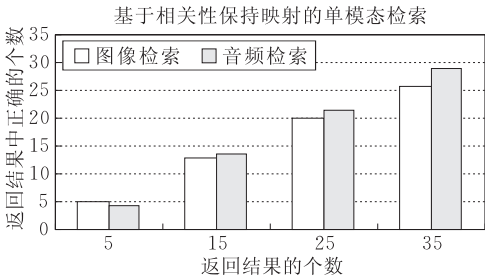


图 4 同构子空间中的图像检索和音频检索

6 结束语

传统基于内容的多媒体检索技术受限于单模态,不同模态之间的异构性、相关性度量以及语义提取一直是它面临的主要问题.本文提出并实现了一种基于内容相关性的跨媒体检索方法,使得用户可以通过提交一种模态的查询例子,检索不同模态的媒体对象,并且通过相关反馈有效地修正跨媒体查询结果.

该方法是以分析不同模态媒体数据的底层内容特征为基础,因此除了实验中的音频和图像数据,同样适用于其他两种模态的多媒体对象.不足之处在于相关性学习只针对两种不同模态的特征矩阵,当引入第三种模态的媒体数据时,需要建立新的模型以分析三种模态间的相关性,我们可以通过邻接图和多示例学习等方法来改进.

此外,进一步的研究还包括:跨媒体相关性的传递和求精,例如用信念度传递的方式从多个数据空间的角度,而不是从单一角度度量相关性.

参 考 文 献

[1] Zhang Hong-Jiang, Zhong Di. Schema for visual feature-based image indexing//Proceedings of the SPIE, Storage and Retrieval for Image and Video Database. San Diego, USA, 1995: 36-46

- [2] David R H, John S T. KCCA for different level precision in content-based image retrieval//Proceedings of the 3rd International Workshop on Content-Based Multimedia Indexing. Rennes, France, 2003; 51-56
- [3] Snoek C G M, Worring M, Geusebroek J M. Semantic video search engine//Proceedings of the TRECVID Workshop. Gaithersburg, USA, 2004; 102-105
- [4] Zhao Xue-Yan, Zhuang Yue-Ting, Wu Fei. Audio clip retrieval with fast relevance feedback based on constrained fuzzy clustering and stored Index table//Proceedings of the Pacific-Rim Conference on Multimedia. Taiwan, China, 2002; 237-244
- [5] McGurk J M. Hearing lips and seeing voices. *Nature*, 1976, 264(5588); 746-748
- [6] Hardoon D R. A correlation approach for automatic image annotation//Proceedings of the 2nd International Conference on Advanced Data Mining and Applications. Xi'an, China, 2006; 681-692
- [7] Wang Xin-Jing, Ma Wei-Ying, Xue Gui-Rong, Li Xing. Multi-model similarity propagation and its application for web image retrieval//Proceedings of the ACM Multimedia Conference. New York, USA, 2004; 944-951
- [8] Ma Qiang, Akiyo Nadamoto, Katsumi Tanaka. Complementary information retrieval for cross-media news content. *Proceedings of Information Systems*, 2006, 31(7); 659-678
- [9] Adams W H, Iyengar G, Lin C Y. Semantic indexing of multimedia content using visual, audio and text cues. *Eurasip Journal on Applied Signal Processing*, 2003(2); 170-185
- [10] Jolliffe I T. *Principal Component Analysis*. New York; Springer-Verlag, 1986; 74-81
- [11] Hansen L K, Larsen J, Kolenda T. On independent component analysis for multimedia signals//Guan L, Kung S Y, Larsen J. *Multimedia Image and Video Processing*. London; CRC Press, 2000; 175-200
- [12] Lu Wen-Cong, Chen Nian-Yi, Li Guo-Zheng, Yang Jie. Multitask learning using partial least square method//Proceedings of the 7th International Conference on Information Fusion. Stockholm, Sweden, 2004; 79-84
- [13] Zhang Hong, Zhuang Yue-Ting, Wu Fei. Cross-modal correlation learning for clustering on image-audio dataset//Proceedings of the ACM Multimedia. Augsburg, German, 2007; 273-276
- [14] Xing E P, Ng A Y, Jordan M I, Russell S. Distance metric learning with application to clustering with side-information. *Advances in Neural Information Processing Systems*. Canada, 2003, 15; 505-512



ZHANG Hong, born in 1979, Ph. D., lecturer. Her research interests include content-based multimedia analysis and retrieval, machine learning.

WU Fei, born in 1973, Ph. D., assistant professor. His research interests include content-based multimedia retrieval

and statistical learning theory.

ZHUANG Yue-Ting, born in 1965, Ph. D., professor. His research interests include multimedia database, artificial intelligence, content-based multimedia retrieval and video-based cartoon.

CHEN Jian-Xun, born in 1957, Ph. D., professor. His research interests include Web-based multimedia retrieval and computer graphics.

Background

Cross-media retrieval discussed in this paper is a new research topic in content-based multimedia analysis and retrieval area. Most researchers focus on how to calculate the similarity between two multimedia objects of the same modality. Cross-media similarity between multimedia objects of different modalities is difficult to measure because of content heterogeneity. This paper solves the problem of cross-media similarity measure with semi-supervised learning methods, and support user interaction in relevance feedback. This paper basically implements a primary cross-media retrieval system. The main limitation is that cross-media indexing strategies need to be incorporated when the size of multimedia database is huge. This work is supported by the National Natural Science Foundation of China (Nos.60533090, 60525108), Key Technology R&D Program (2006BAH02A13-4), the

National High Technology Research and Development Program (863 Program) of China(2006AA010107), Program for Changjiang Scholars and Innovative Research Team in University(IRT0652,PCSIRT). Heterogeneous multimedia data stored in digital libraries and data centers is semi-structured or unstructured, and these multimedia data is connected from both semantic and content level. Above projects focus on intelligent processing and integrative retrieval techniques to better utilize multimedia resources. The research team focuses on multimedia semantic learning by content analysis, cross-media retrieval algorithms, multimedia database indexing, etc., and has published some papers. This paper focuses on the part of cross-media retrieval algorithm, which solves the problem of cross-media similarity measure.