ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.19, No.11, November 2008, pp.2853−2868
DOI: 10.3724/SP.J.1001.2008.02853

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

基于张量表示的直推式多模态视频语义概念检测*

吴 飞+, 刘亚楠, 庄越挺

(浙江大学 计算机科学与技术学院 数字媒体计算与设计实验室,浙江 杭州 310027)

Transductive Multi-Modality Video Semantic Concept Detection with Tensor Representation

WU Fei+, LIU Ya-Nan, ZHUANG Yue-Ting

(Digital Media Computing & Design Lab., College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China)

+ Corresponding author: E-mail: wufei@zju.edu.cn

Wu F, Liu YN, Zhuang YT. Transductive multi-modality video semantic concept detection with tensor representation. *Journal of Software*, 2008,19(11):2853−2868. http://www.jos.org.cn/1000-9825/19/2853.htm

Abstract: A higher-order tensor framework for video analysis and understanding is proposed in this paper. In this framework, image frame, audio and text are represented, which are the three modalities in video shots as data points by the 3rd-order tensor. Then a subspace embedding and dimension reduction method is proposed, which explicitly considers the manifold structure of the tensor space from temporal-sequenced associated co-occurring multimodal media data in video. It is called TensorShot approach. Transductive learning uses a large amount of unlabeled data together with the labeled data to build better classifiers. A transductive support tensor machines algorithm is proposed to train effective classifier. This algorithm preserves the intrinsic structure of the submanifold where tensorshots are sampled, and is also able to map out-of-sample data points directly. Moreover, the utilization of unlabeled data improves classification ability. Experimental results show that this method improves the performance of video semantic concept detection.

Key words: multi-modality; TensorShot; temporal associated cooccurrence (TAC); higher order SVD (HOSVD); dimensionality reduction; transductive support tensor machine (TSTM)

摘 要: 提出了一种基于高阶张量表示的视频语义分析与理解框架.在此框架中,视频镜头首先被表示成由视频中所包含的文本、视觉和听觉等多模态数据构成的三阶张量;其次,基于此三阶张量表达及视频的时序关联共生特性设计了一种子空间嵌入降维方法,称为张量镜头;由于直推式学习从已知样本出发能对特定的未知样本进行学习和识别,最后在这个框架中提出了一种基于张量镜头的直推式支持张量机算法,它不仅保持了张量镜头所在的流形空间的本征结构,而且能够将训练集合外数据直接映射到流形子空间,同时充分利用未标记样本改善分类器的学习性能.实验结果表明,该方法能够有效地进行视频镜头的语义概念检测.

关键词: 多模态;张量镜头;时序关联共生;高阶 SVD;降维;直推式支持张量机

* Supported by the National Natural Science Foundation of China under Grant Nos.60603096, 60533090 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2006AA010107 (国家高技术研究发展计划(863); the National Key Technology R&D Program of China under Grant No.2007BAH11B01 (国家科技支撑计划); the Program for Changjiang Scholars and Innovative Research Team in University of China under Grant Nos.IRT0652, PCSIRT (长江学者和创新团队发展计划)

Received 2008-03-01; Accepted 2008-08-26

中图法分类号: TP391 文献标识码: A

随着信息技术的发展和数字化手段的提高,出现了越来越多的大规模视频数据库.视频数据蕴含了任务、场景、对象和事件等丰富的语义信息;同时,视频也是时序复杂媒体,主要包含图像、音频和文本三种媒质数据.如何有效地利用视频所体现的多模态及时序特性来挖掘其蕴涵丰富语义,从而支持视频有效检索,发挥视频数据的资源共享优势,是一个具有挑战性的研究问题[1−3].

视频中多种模态的融合与交互对于缩减底层特征与高层语义之间的“语义鸿沟(semantic gap)”起到了重要的作用.文献[4]在考虑视频中的多种信息流,如视觉流、听觉流以及文本流的基础上,提出了一种基于多模态组合的视频事件索引方法.文献[5]提出了时间间隔多媒体事件(time intervals multimedia events,简称 TIME)的概念,TIME 通过考虑视频流中的上下文和同步关系,将多种异构信息源结合起来共同表达某一语义事件.文献[6]从视频语义特征的角度出发,将与视频语义相关的声音、字幕、音乐、剧情脚本、新闻文稿等信息特征进行整合,通过人像、字幕、语音、视频镜头识别和剧情脚本分析等组合手段,实现视频语义特征的多模式提取和检索.在本文中,我们认为视频在本质上是时序数据,呈现出时序关联共生(temporal associated cooccurrence,简称 TAC)的特性[7].视频中所包含的多种模态之间是相互影响甚至是互补的.一方面,比如在某个时间段内,视频帧、音频信号及转录文本等多媒质数据可能并不在同一时刻出现,存在不同步现象,但却共同表达一个语义,在语义持续时间内彼此耦合关联;另一方面,多种模态之间语义的互补性也很重要.例如,表达同一语义的不同镜头可能在视觉上看起来有很大区别.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,从而弥补了其他模态较弱的相关性.

传统视频表达所采用的向量模型除了会产生高维向量而导致“维度灾难”问题以外,同时,在降维过程中,如文献[8]所描述的,由于特征向量过高的维度及训练样本的数据不足,将不同类型特征进行拼合会引起“过压缩(over-compression)”问题,以致丢失大量信息.另外,不同类型特征通过简单向量拼接也在一定程度上减弱或忽略了视频中这些多种模态特征之间的时序关联共生性.因此,本文采用多线性几何即高阶张量来表达和分析包含多模态媒体特征的视频镜头.张量是对向量和矩阵的自然扩展,张量几何也已定义了一系列在向量空间上的多线性运算法则[9].在张量表达中,视频数据中所包含的同一类型媒体数据特征被表达为张量的一阶,在一定程度上避免了从不同类型媒体数据中所提取特征因为拼合而产生的维数灾难及过压缩问题.

研究者已经普遍认识到,高维数据其实很可能在本质上是由有限的自由度(degrees of freedom)来决定的,并且分布在较低维度的流形空间上.为了寻找这个低维度空间,一些典型的流形学习方法包括 Principle Component Analysis(PCA)[10],Locally Linear Embedding(LLE)[11],ISOMAP[12],Laplacian Eigenmaps[13]及 Locality Preserving Projections(LPP)[14]等被相继提出和使用.近年来,流形学习已被应用于人脸识别[15]、图形学[16]、图像检索[17]等领域,并在表达和处理高维数据方面表现出较好的能力.我们也注意到,已有一些工作以视频分布在流形上为前提假设,希望在此基础上寻找合适的几何拓扑或相关关系来分析和处理视频高维数据,如文献[18,19].因此,我们的工作也以视频镜头数据分布在流形上作为假设前提.

由于视频分别包含图像、音频和文本三种模态特征数据,可将视频镜头表示为三阶张量.但是,因为这 3 种媒质数据具有较高的维度,并且根据语义理解的需要及已有的标注信息,仍需寻找一个能够保持张量空间本征结构并将底层特征空间映射到高层语义空间的降维方法,使得语义上相似的视频镜头在经过映射后在高层语义空间上仍然保持相关关系,本文称其为张量镜头(TensorShot)方法.与 Lapacian Eigenmaps 和 LPP 等降维方法类似,张量镜头也将张量空间的流形结构模拟为一个近邻图,通过近似估计流形上的 Laplace-Beltrami 算子的特征方程来获得局部等距意义上的最优张量子空间[20].但是,Lapacian Eigenmaps 和 LPP 只适用于向量,而本文提出的张量镜头方法可以应用于张量空间.同时,正如 LPP 是一种线性降维方法而能够直接应用于训练集外数据,本文张量镜头降维方法得到的映射转换矩阵同样可直接将训练集合外的数据映射到低维子空间.另外,在具体计算过程中,本文采用了高阶奇异值分解(higher order SVD,简称 HOSVD)[21]来分别发现图像、音频和文本三个模态语义子空间.张量不仅在表达上突出了多种模态的层次性,而且在寻找本征低维流形空间时可以更方便地

融合和传递 3 种模态特征之间的相关性,从而更好地描述视频镜头的语义关系.

张量的表达方法不仅可以减少原始数据的输入参数,也可以解决“过学习(overfitting)”问题.文献[22]提出了一个以张量作为输入的监督张量学习(supervised tensor learning,简称 STL)框架.STL 是凸优化和多线性几何运算的结合,它采用交替投影优化来实现.基于 STL 可将传统支持向量机(support vector machine,简称 SVM)[23]扩展到支持张量机(support tensor machine,简称 STM).

在传统的监督学习中,学习机通过对大量已标记的(labeled)训练样本进行学习,训练得到映射模型,再通过映射模型来预测未知样本的类别标记.随着数据收集和存储技术的飞速发展,大量未标记(unlabeled)样本的收集已相当容易,而由于人力、物力的原因则使得获取大量已标记的样本则变得相对困难.因此,在已标记样本较少时,如何利用大量未标记样本来改善学习性能的半监督学习成为当前最受关注的问题之一.半监督学习方式事实上假设了同类别的未标记数据与已标记样本在特征空间上的距离是比较近的,在已标记类别样本提供的监督信息的引导下,去学习全部样本或未标记类别样本的标记信息.现有的半监督学习算法包括生成式模型、协同训练、直推式支持向量机以及基于图正则化框架的半监督学习算法等[24].其中,Joachims 提出的直推式支持向量机(transductive support vector machine,简称 TSVM)[25]是传统支持向量机算法在半监督学习问题上的一种扩展.对于二值分类问题来说,传统 SVM 是利用已标记数据在样本空间中寻找一个最优超平面,使两类样本间的分类间隔最大.而 TSVM 则同时使用已标记样本和未标记样本来寻找最优分类边界,使得到的分类间隔能够最大限度地分割原始已标记样本和未标记样本(经 TSVM 学习后其标记将变为已知),新找到的最优分类边界应满足对原始未标记样本的分类具有最小泛化误差.

结合本文提出的张量镜头表达以及支持张量机学习算法,并考虑半监督学习的优点,本文提出一种扩展的直推式支持张量机(transductive support tensor machine,简称 TSTM),训练扩展直推式支持向量机为语义分类器,以实现对视频镜头语义概念的检测.TSTM 算法不仅能够充分利用张量表达方式的优势,而且能够在训练过程中有效地利用未标记样本分布信息,更好地刻画整个样本空间上的数据特性,优化了分类器的分类性能,得到更好的分类检测结果.

本文第 1 节简要介绍涉及到的相关知识.第 2 节介绍本文提出的张量镜头和直推式支持张量机算法.第 3 节给出实验结果并加以讨论.第 4 节总结全文.

1 相关知识

本节简要介绍本文所涉及到的张量几何[26]相关定义、监督张量学习框架基础[22]以及 Joachims 提出的直推式支持向量机算法[25].

首先说明本文所使用的各种符号的含义.斜体小写字母($a,b,\ldots$)表示标量,粗斜体小写字母($\boldsymbol{a},\boldsymbol{b},\ldots$)表示向量,矩阵由粗斜体大写字母($\boldsymbol{A},\boldsymbol{B},\ldots$)表示,高阶张量则由手写体字母($\mathcal{A},\mathcal{B},\ldots$)或($\mathcal{a},\mathcal{b},\ldots$)来表示.

1.1 张量几何

张量(tensor)又称为多维数组.张量 $\mathcal{A}\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_N}$ 的阶(order)是 N,$\mathcal{A}$ 的第 n 模(mode 或 way)的维度大小是 I_n,$\mathcal{A}$ 中的元素用 $\mathcal{A}(i_1,i_2,\ldots,i_N)$ 或 $a(i_1,i_2,\ldots,i_N)$ 表示.标量是零阶张量,n 维向量是大小为 n 的一阶张量,$m\times n$ 维的矩阵是大小为 $m\times n$ 的二阶张量.

定义 1(张量积). 张量 $\mathcal{A}\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_P}$ 和 $\mathcal{B}\in\mathbb{R}^{J_1\times J_2\times\ldots\times J_Q}$ 的张量积 $\mathcal{A}\circ\mathcal{B}$ 定义为

$$(\mathcal{A}\circ\mathcal{B})_{i_1i_2\ldots i_Pj_1j_2\ldots j_Q}\overset{\text{def}}{=}a_{i_1i_2\ldots i_P}b_{j_1j_2\ldots j_Q} \tag{1}$$

定义 2(标量积). 张量 $\mathcal{A},\mathcal{B}\in\mathbb{R}^{I_1\times I_2\times\ldots\times I_N}$ 的标量积 $\langle\mathcal{A},\mathcal{B}\rangle$ 定义为

$$\langle\mathcal{A},\mathcal{B}\rangle\overset{\text{def}}{=}\sum_{i_1}\sum_{i_2}\cdots\sum_{i_N}a_{i_1i_2\ldots i_N}b_{i_1i_2\ldots i_N}. \tag{2}$$

定义 3(正交性). 标量积为 0 的张量之间相互正交.

定义 4(Frobenius 范数). 根据张量内积的定义,张量 $\mathcal{A}$ 的 Frobenius 范数可以表示为

$$\|\mathcal{A}\| \overset{\text{def}}{=} \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle} \tag{3}$$

定义 5(张量矩阵展开). 张量矩阵展开(matrix unfolding)是指将一个张量中的元素重新排列,得到一个矩阵的过程.张量 $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ 的 n 模(mode-n)展开矩阵表示为 $\mathcal{A}_{(n)} \in \mathbb{R}^{I_n \times (I_1 \ldots I_{n-1} I_{n+1} \ldots I_N)}$.

定义 6(张量乘法). 由于高阶张量的维数是任意的,因此在计算张量与矩阵乘法时需指明是张量中哪一维与给定矩阵的列或行相乘.张量与矩阵的乘法由 n 模乘积(mode-n product)所定义.张量 $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ 和矩阵 $\boldsymbol{U} \in \mathbb{R}^{J_n \times I_n}$ 的 n 模乘积表示为 $\mathcal{A} \times_n \boldsymbol{U}$, 是一个 $I_1 \times I_2 \times \ldots \times I_{n-1} \times J_n \times I_{n+1} \times \ldots \times I_N$ 阶张量,定义如下:

$$(\mathcal{A} \times_n \boldsymbol{U})_{i_1 i_2 \ldots j_n \ldots i_N} = \sum_{i_n} a_{i_1 i_2 \ldots i_n \ldots i_N} \cdot u_{j_n i_n} \tag{4}$$

张量矩阵展开后得到的矩阵形式张量可以直接作为矩阵进行运算,因此,n 模乘积 $\mathcal{B} = \mathcal{A} \times_n \boldsymbol{U}$ 可以由矩阵乘法 $\mathcal{B}_{(n)} = \boldsymbol{U} \mathcal{A}_{(n)}$ 得到.

令 $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$, 如果 $\boldsymbol{U} \in \mathbb{R}^{J_m \times I_m}$ 且 $\boldsymbol{V} \in \mathbb{R}^{J_n \times I_n}$, 则有如下性质:

$$\mathcal{A} \times_m \boldsymbol{U} \times_n \boldsymbol{V} = \mathcal{A} \times_n \boldsymbol{V} \times_m \boldsymbol{U} \tag{5}$$

在本文中,为简化符号表示,有以下表达:

$$\mathcal{A} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times \ldots \times_M \boldsymbol{U}_M \triangleq \mathcal{A} \prod_{k=1}^{M} \times_k \boldsymbol{U}_k \tag{6}$$

定理 1(高阶奇异值分解(HOSVD)). 每个张量 $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N}$ 均可唯一分解为 $\mathcal{A} = \mathcal{T} \times_1 \boldsymbol{U}_1 \times_2 \boldsymbol{U}_2 \times \ldots \times_N \boldsymbol{U}_N$, 使得:1) $\boldsymbol{U}_{(n)}$均为 $I_n \times I_n$ 维正交矩阵;2) $\mathcal{T}_{(n)}$的行都是正交的;3) 对于任意 n,都有 $\left\|\mathcal{T}_{i_n=1}\right\| \geq \left\|\mathcal{T}_{i_n=2}\right\| \geq \ldots \geq \left\|\mathcal{T}_{i_n=N}\right\| \geq 0$.

张量 $\mathcal{T}$ 也称为核心张量(core tensor),类似于常规矩阵 SVD 分解得到的特征根对角矩阵.对于张量 $\mathcal{A}$,其高阶奇异值分解算法如下:

1) 对于任意 n,$\boldsymbol{U}_{(n)}$是张量 $\mathcal{A}$ 的 n 模展开矩阵 $\mathcal{A}_{(n)}$进行 SVD 分解得到的左矩阵;

2) 计算 $\mathcal{T} = \mathcal{A} \times_1 \boldsymbol{U}_1^T \times_2 \boldsymbol{U}_2^T \times \ldots \times_N \boldsymbol{U}_N^T$.

1.2 监督张量学习框架

文献[22]中提出了一个监督张量学习框架.作者认为张量表达方式可以有效解决"过学习"问题,并实现了应用张量最小最大概率机(tensor minimax probability machine)进行图像分类.在 STL 框架中,传统的基于向量的学习机都可以扩展为以张量作为输入进行训练学习.

给定 N 个张量所表示的训练数据 $\mathcal{X}_i \in \mathbb{R}^{I_1 \times I_2 \times \ldots I_N}$, 每个训练数据都有类别标记信息 $y_i \in \{+1, -1\}$. 将正($y_i = +1$)、负($y_i = -1$)两类数据分开的过程可以按照如下优化问题进行定义:

$$\begin{bmatrix} \min\limits_{\boldsymbol{w}_k|_{k=1}^M, b, \boldsymbol{\xi}} f(\boldsymbol{w}_k|_{k=1}^M, b, \boldsymbol{\xi}) \\ \text{s.t. } y_i c_i \left(\mathcal{X}_i \prod_{k=1}^{M} \times_k \boldsymbol{w}_k + b \right) \geq \xi_i, 1 \leq i \leq N \end{bmatrix} \tag{7}$$

其中, $f: \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_M} \to \mathbb{R}$ 是一个凸函数,作为分类准则;对于任意 $1 \leq i \leq N$, $c_i: \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_M} \to \mathbb{R}$ 是一个凸约束方程; $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_N]^T \in \mathbb{R}^N$ 是松弛变量;判别函数为 $y(\mathcal{X}) = sign\left(\mathcal{X} \prod_{k=1}^{M} \times_k \boldsymbol{w}_k + b \right)$, 其中, $\boldsymbol{w}_k \in \mathbb{R}^{I_k} (1 \leq k \leq M), b \in \mathbb{R}$, 张量分类超平面是 $\mathcal{X} \prod_{k=1}^{M} \times_k \boldsymbol{w}_k + b = 0$.

在 STL 的拉格朗日(Lagrangian)求解中可以发现,任意 $\boldsymbol{w}_k (1 \leq k \leq M)$ 之间相互影响,不能直接计算得到结果,因此,文献[22]提出了交替投影优化(alternating projection)算法,其关键步骤在于根据确定的 $\boldsymbol{w}_k$, 以迭代方式来求得 $\boldsymbol{w}_j (1 \leq k, j \leq M, k \neq j)$, 同时证明了该方法的收敛性.

1.3 直推式支持向量机

下面简要介绍 Joachims 提出的直推式支持向量机的算法原理和实现.具体的描述和证明参见文献[25].

给定一组独立同分布的 l 个已标记训练样本集 $(x_1,y_1),(x_2,y_2),...,(x_l,y_l),y_i\in\{+1,-1\}$ 和另一组具有同一分布的 u 个未标记测试样本点 $x_1^*,x_2^*,...,x_u^*$.

在一般线性不可分条件下,Joachims 提出的直推式支持向量机训练过程可以描述为如下优化求解问题:

$$\left[\begin{array}{l}\text{Minimize over } (y_1^*,y_2^*,...,y_u^*,\boldsymbol{w},b,\xi_1,...,\xi_l,\xi_1^*,...,\xi_u^*):\\ \qquad\qquad \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\xi_i+C^*\sum_{j=1}^{u}\xi_j\\ \text{s.t.}\quad \forall_{i=1}^{l}: y_i(\boldsymbol{w}\cdot x_i+b)\ge 1-\xi_i, 1\le i\le l\\ \qquad \forall_{j=1}^{u}: y_j(\boldsymbol{w}\cdot x_j^*+b)\ge 1-\xi_j^*, 1\le j\le u\\ \qquad \forall_{i=1}^{l}: \xi_i>0, 1\le i\le l\\ \qquad \forall_{j=1}^{u}: \xi_j^*>0, 1\le j\le u\\ \qquad \frac{1}{u}\sum_{j=1}^{u}\max[0,sign(\boldsymbol{w}\cdot x_j^*+b)]=r\end{array}\right. \tag{8}$$

求解上述优化问题的目的就是为了找到一个超平面,使其对已标记训练样本和未标记测试样本都能满足分类间隔最大,且能预测未标记样本的类别标记.其中,C 和 C^*是由用户指定及调节的参数,C^*可被看作是未标记样本在训练过程中的影响因子,它们可以通过交叉验证来获得最优值.另外,r 是在训练过程中待训练未标记样本集合中将要标记为正的样本数,其初始值可设定为已标记样本集中正标记样本所占的比例,同时也可以在训练过程中通过交叉验证来调节 r 的大小.

TSVM 训练过程就是求解上述优化问题的过程,首先对已标记样本集进行初始学习得到一个初始分类器并对未标记样本进行初始分类,对判别函数输出值最大的 r 个未标记样本暂时赋予正标记,其余赋予负标记;接下来对所有样本重新训练,使用新得到的分类器对所有样本进行标记,且按一定规则交换一对标记值不同的测试样本所具有的类别标识,使得式(8)中目标函数值获得最大下降.上述步骤反复执行,直到找不出符合交换条件的样本对为止.通过均匀增加未标记样本影响因子的值,迭代训练分类器,以得到对未标记样本具有尽可能小的分类误差的分类函数.

陈毅松等人在文献[27]中指出,在已标记样本数较少的情况下,TSVM 算法中指定 r 值很容易导致较大估计误差,而一旦错误估计了 r 值,则将导致训练算法产生一个不能正确描述样本分布特征的学习机.因此,文献[27]中提出了一种渐进直推式支持向量机(progressive transductive support vector machine,简称 PTSVM),在 PTSVM 中无须事先设定未标记样本中的正标记样本数,而是在训练过程中根据渐进赋值和动态调整的规则对未标记样本逐一赋予可能的标记,并对新得到的已标记样本集重新训练.

2 张量镜头:直推式多模态视频语义概念检测

本文将视频切分为一个个镜头,以镜头作为语义识别的基本处理单元.我们希望通过合适的视频镜头表达方式,融合和传递视频中多种模态媒质之间存在的相关性,更加准确地获得镜头之间的相似度关系,通过降维、分类和聚类等处理来识别镜头所蕴涵的高层语义.

2.1 基于张量的视频镜头表达

2.1.1 底层特征提取

底层特征是指直接从视频源数据中提取的特征,有别于语义概念所代表的高层特征.本文从每一个镜头中分别提取图像、音频和文本等底层特征.

图像特征:镜头是基本处理单元,从每个镜头中选取一个关键帧作为代表图像,然后提取关键帧的颜色直方图、纹理和 Canny 边界作为图像特征.

音频特征:将镜头相对应的音频作为一个音频例子(audio clip),并将该音频例子分割成迭加短时音频帧,提取每个短时音频帧特征,包括 MFCC、质心、衰减截止频率、频谱流量及过零率,形成短时帧特征向量,然后计算短时音频帧特征向量的统计值(均值或方差)作为镜头的音频特征.

文本特征:本文从视频中经过识别的转录(transcript)文本提取特征.由于文本特征的维数远大于其他模态特征,并且文本中包含了单词共生等丰富信息,可以先采用隐含语义分析(latent semantic analysis,简称 LSA)对文本特征进行降维处理.

视频本身是时序数据,在提取上述底层特征时,考虑到了视频的时序特性(temporal characteristic).如在视频采样帧中提取关键帧,以及短时音频帧特征的统计值(均值或方差)提取.由于张量是由以上不同类型特征组成的,所提取的特征体现了视频的时序特性,因此,在所组成的张量中,一定程度上也体现了视频时序特性.

2.1.2 张量镜头的表达

在提取视频中图像、音频和文本等特征后,将每个视频镜头用一个三阶张量 $S\in\mathbb{R}^{I_1\times I_2\times I_3}$ 来表示,其中,I_1,I_2 和 I_3 分别是图像特征向量、音频特征向量及文本特征向量的维数.每个元素 $s_{i_1i_2i_3}$ 的值定义如下:

- $s_{i_1,1,1}(1\le i_1\le I_1)$ 为图像特征向量对应的值;
- $s_{2,i_2,2}(1\le i_2\le I_2)$ 为音频特征向量对应的值;
- $s_{3,3,i_3}(1\le i_3\le I_3)$ 为文本特征向量对应的值;
- 其他值均初始设为 0.通过第 2.2 节的计算以后,由于 3 种模态特征之间的融合、传递及互补,使得这些初始的零值有所改变并具有意义.

2.2 张量镜头的语义子空间嵌入和降维

如前所述,视频镜头集合本质上分布在一个内嵌低维流形空间内.然而,现有的降维和流形学习方法均限于处理向量而无法映射得到内嵌流形空间.下面介绍本文提出的张量镜头降维方法.本算法不仅能够保持镜头的局部近邻信息,而且根据已有的标注信息,能够发现其所在语义流形空间的本征非线性结构.此外,本算法是线性映射,因此能够直接对训练集合外的数据进行降维.

给定空间 $\mathbb{R}^{I_1\times I_2\times I_3}$ 上的镜头数据集合 $X=\{\mathcal{X}_1,\mathcal{X}_2,...,\mathcal{X}_N\}$,我们希望为 X 上的每个张量镜头 $\mathcal{X}_i|_{i=1}^N$ 寻找下面 3 个语义投影矩阵:$J_1\times I_1$ 维的 $\boldsymbol{T}_1^i$,$J_2\times I_2$ 维的 $\boldsymbol{T}_2^i$ 及 $J_3\times I_3$ 维的 $\boldsymbol{T}_3^i$,使其映射这 N 个数据点到空间 $\mathbb{R}^{J_1\times J_2\times J_3}$ ($J_1<I_1$, $J_2<I_2$,$J_3<I_3$)上的 $Y=\{\mathcal{Y}_1,\mathcal{Y}_2,...,\mathcal{Y}_N\}$,并满足 $\mathcal{Y}_i|_{i=1}^N=\mathcal{X}_i\times_1\boldsymbol{T}_1^i\times_2\boldsymbol{T}_2^i\times_3\boldsymbol{T}_3^i$.于是,低维数据集 Y 就反映了 X 所在流形空间的本征几何拓扑结构并体现语义关联,将底层特征空间映射到高层语义空间;同时,这个映射也具有线性特性,也就是说,对于训练集合之外的数据点 $\mathcal{X}_t$,可以直接由预先训练出的语义投影矩阵来计算以得到它在低维语义子空间中的映射.

令 $\mathcal{X}\in\mathbb{R}^{I_1\times I_2\times I_3}$ 代表一个三阶张量镜头,给定分布在 N 个张量镜头流形空间 $\mathcal{M}\in\mathbb{R}^{I_1\times I_2\times I_3}$ 上的数据集合 $X=\{\mathcal{X}_1,\mathcal{X}_2,...,\mathcal{X}_N\}$,可以构建一个最近邻图 $\boldsymbol{G}$ 来模拟 $\mathcal{M}$ 的局部几何结构.在已知视频镜头的语义类别标注信息时,定义 $\boldsymbol{G}$ 的权重矩阵 $\boldsymbol{W}$ 如下:

$$\boldsymbol{W}_{ij}=\begin{cases}1, & \text{当}\mathcal{X}_i\text{和}\mathcal{X}_j\text{被标注为同一类别}\\ 0, & \text{其他}\end{cases} \tag{9}$$

对于每个张量镜头 $\mathcal{X}_i|_{i=1}^N$,根据高阶奇异值分解(HOSVD),可分别对 $\mathcal{X}_i$ 的 $k|_{k=1}^3$ 模展开矩阵(mode-k unfolding matrix) $\mathcal{X}_{(1)}^i,\mathcal{X}_{(2)}^i,\mathcal{X}_{(3)}^i$ 进行奇异值 SVD 分解,分别得到左矩阵 $\boldsymbol{U}_1^i,\boldsymbol{U}_2^i,\boldsymbol{U}_3^i$,称为 3 种模态的互补代表矩阵.举例来说,$\boldsymbol{U}_1^i$ 是对 $\mathcal{X}_i$ 的一模展开矩阵(mode-1 unfolding matrix) $\mathcal{X}_{(1)}^i$ 进行 SVD 分解后得到的左矩阵,可作为图像特征的代表矩阵.在这一展开和分解过程中,不同模态之间实现了传递、融合及互补,得到更完整的表达.而传统的通过拼合方式来“融合”形成不同类型媒体的向量表达则无法实现这种交互影响.如果视频中图像、音频和文本特征分别表示为 I,A 和 T,这一形成的拼合向量为 I,A,T 的任意一种组合形式.这种向量形式不仅忽视了不同类型媒体数据量纲有所不同,而且在后继处理中无法实现视频中这些不同媒体数据的相互影响.在张量表

达中,不同类型的媒体数据分别被表达成一阶(order)形式,体现了其量纲上的不同,且在张量的矩阵展开过程中,由于对组成张量的所有阶按交错次序采样,并非简单地先采样完某一类型特征再采样其他类型特征,而是在采样过程中对不同类型特征混合在一起交错采样,这样的采样展开过程体现了不同类型特征的传递和融合.我们在实验中也对比了张量表达和向量表达在视频语义理解方面的准确率,张量表达取得了较好的结果.

现以图像特征的互补代表矩阵 $\boldsymbol{U}_1^i|_{i=1}^N \in \mathbb{R}^{I_1\times I_1}$ 为例,来寻找 $I_1\times J_1$ 维的中间转换矩阵 $\boldsymbol{V}_1$,将 $\boldsymbol{U}_1^i$ 映射得到 $\boldsymbol{T}_1^i=\boldsymbol{V}_1^T\boldsymbol{U}_1^i\in\mathbb{R}^{J_1\times I_1}$, 即图像特征模态的语义投影矩阵.可以从两个角度来考虑这个问题.一方面,要保持流形的本征结构,需求取目标函数(10)的最优解[28]:

$$\min_{\boldsymbol{V}_1}\sum_{ij}\left\|\boldsymbol{V}_1^T\boldsymbol{U}_1^i-\boldsymbol{V}_1^T\boldsymbol{U}_1^j\right\|^2\boldsymbol{W}_{ij} \tag{10}$$

也就是说,最小化 $\sum_{ij}\left\|\boldsymbol{V}_1^T\boldsymbol{U}_1^i-\boldsymbol{V}_1^T\boldsymbol{U}_1^j\right\|^2\boldsymbol{W}_{ij}$ 如果能够确保 $\boldsymbol{U}_1^i$ 和 $\boldsymbol{U}_1^j$ 是相近的,那么 $\boldsymbol{V}_1^T\boldsymbol{U}_1^i$ 和 $\boldsymbol{V}_1^T\boldsymbol{U}_1^j$ 也是相近的.

令 $\boldsymbol{D}$ 为 $\boldsymbol{W}$ 的对角矩阵,即 $\boldsymbol{D}_{ii}=\sum_j\boldsymbol{W}_{ij}$. 对于一个矩阵 $\boldsymbol{A}$,它的迹(trace)满足 $\|\boldsymbol{A}\|^2=tr(\boldsymbol{A}\boldsymbol{A}^T)$, 则有

$$\begin{aligned}\frac{1}{2}\sum_{ij}\left\|\boldsymbol{V}_1^T\boldsymbol{U}_1^i-\boldsymbol{V}_1^T\boldsymbol{U}_1^i\right\|^2\boldsymbol{W}_{ij}&=\frac{1}{2}\sum_{ij}tr((\boldsymbol{T}_1^i-\boldsymbol{T}_1^j)(\boldsymbol{T}_1^i-\boldsymbol{T}_1^j)^T)\boldsymbol{W}_{ij}\\&=\frac{1}{2}\sum_{ij}tr(\boldsymbol{T}_1^i\boldsymbol{T}_1^i+\boldsymbol{T}_1^j\boldsymbol{T}_1^j-\boldsymbol{T}_1^i\boldsymbol{T}_1^{jT}-\boldsymbol{T}_1^j\boldsymbol{T}_1^{iT})\boldsymbol{W}_{ij}\\&=tr\left(\sum_i\boldsymbol{D}_{ii}\boldsymbol{T}_1^i\boldsymbol{T}_1^{iT}-\sum_{ij}\boldsymbol{W}_{ij}\boldsymbol{T}_1^i\boldsymbol{T}_1^{jT}\right)\\&=tr\left(\sum_i\boldsymbol{D}_{ii}\boldsymbol{V}_1^T\boldsymbol{U}_1^i\boldsymbol{U}_1^{iT}\boldsymbol{V}_1-\sum_{ij}\boldsymbol{W}_{ij}\boldsymbol{V}_1^T\boldsymbol{U}_1^i\boldsymbol{U}_1^{jT}\boldsymbol{V}_1\right)\\&=tr\left(\boldsymbol{V}_1^T\left(\sum_i\boldsymbol{D}_{ii}\boldsymbol{U}_1^i\boldsymbol{U}_1^{iT}-\sum_{ij}\boldsymbol{W}_{ij}\boldsymbol{U}_1^i\boldsymbol{U}_1^{jT}\right)\boldsymbol{V}_1\right)\\&\doteq tr(\boldsymbol{V}_1^T(\boldsymbol{D}_U-\boldsymbol{W}_U)\boldsymbol{V}_1)\end{aligned} \tag{11}$$

其中, $\boldsymbol{D}_U=\sum_i\boldsymbol{D}_{ii}\boldsymbol{U}_1^i\boldsymbol{U}_1^{iT}$, $\boldsymbol{W}_U=\sum_{ij}\boldsymbol{W}_{ij}\boldsymbol{U}_1^i\boldsymbol{U}_1^{jT}$. 从上面的推导可以看出,若想求解 $\min_{\boldsymbol{V}_1}\sum_{ij}\left\|\boldsymbol{V}_1^T\boldsymbol{U}_1^i-\boldsymbol{V}_1^T\boldsymbol{U}_1^j\right\|^2\boldsymbol{W}_{ij}$, 则需要最小化 $tr(\boldsymbol{V}_1^T(\boldsymbol{D}_U-\boldsymbol{W}_U)\boldsymbol{V}_1)$.

另一方面,除了要保持流形的本征几何及拓扑结构以外,还需要最大化流形空间上的全局方差.一般地,一个随机变量 x 的方差为

$$var(x)=\int_{\mathcal{M}}(x-\mu)^2\mathrm{d}P(x),\quad \mu=\int_{\mathcal{M}}x\mathrm{d}P(x) \tag{12}$$

其中,$\mathcal{M}$ 是数据的流形,μ是期望值,dP 是概率密度函数.根据谱图理论(spectral graph theory)[29],dP 可以由样本点对角矩阵 $\boldsymbol{D}(\boldsymbol{D}_{ii}=\sum_j\boldsymbol{W}_{ij})$ 的离散化估计得到,因此有如下推导:

$$\begin{aligned}var(\boldsymbol{T}_1)&=\sum_i\left\|\boldsymbol{T}_1^i\right\|^2\boldsymbol{D}_{ii}\\&=\sum_i tr(\boldsymbol{T}_1^i\boldsymbol{T}_1^{iT})\boldsymbol{D}_{ii}\\&=\sum_i tr(\boldsymbol{V}_1^T\boldsymbol{U}_1^i\boldsymbol{U}_1^{iT}\boldsymbol{V}_1)\boldsymbol{D}_{ii}\\&=tr\left(\boldsymbol{V}_1^T\left(\sum_i\boldsymbol{D}_{ii}\boldsymbol{U}_1^i\boldsymbol{U}_1^{iT}\right)\boldsymbol{V}_1\right)\\&\doteq tr(\boldsymbol{V}_1^T\boldsymbol{D}_U\boldsymbol{V}_1)\end{aligned} \tag{13}$$

根据式(11)和式(13)两个方面的约束条件,得到如下优化问题:

$$\min_{\boldsymbol{V}_1}\frac{tr(\boldsymbol{V}_1^T(\boldsymbol{D}_U-\boldsymbol{W}_U)\boldsymbol{V}_1)}{tr(\boldsymbol{V}_1^T\boldsymbol{D}_U\boldsymbol{V}_1)} \tag{14}$$

显然,$\boldsymbol{V}_1$ 的最优解是 $(\boldsymbol{D}_U-\boldsymbol{W}_U,\boldsymbol{D}_U)$ 的广义特征向量.因此,可以通过计算下面的广义特征向量来获得最优化 $\boldsymbol{V}_1$ 值:

$$(\boldsymbol{D}_U-\boldsymbol{W}_U)\boldsymbol{V}_1=\lambda\boldsymbol{D}_U\boldsymbol{V}_1 \tag{15}$$

当计算得到 $\boldsymbol{V}_1$ 后,由 $\boldsymbol{U}_1^i|_{i=1}^N\in\mathbb{R}^{I_1\times I_1}$,可求取 $\boldsymbol{T}_1^i=\boldsymbol{V}_1^T\boldsymbol{U}_1^i\in\mathbb{R}^{J_1\times I_1}$.同理,对于音频以及文本这两种模态的中间转换矩阵 $\boldsymbol{V}_2$ 和 $\boldsymbol{V}_3$,也可用同样的方法来计算,那么由音频特征模态的互补代表矩阵 $\boldsymbol{U}_2^i|_{i=1}^N\in\mathbb{R}^{I_2\times I_2}$ 和 $\boldsymbol{V}_2$ 便可求取此模态的语义投影矩阵 $\boldsymbol{T}_2^i=\boldsymbol{V}_2^T\boldsymbol{U}_2^i\in\mathbb{R}^{J_2\times I_2}$,同理,由文本特征模态的互补代表矩阵 $\boldsymbol{U}_3^i|_{i=1}^N\in\mathbb{R}^{I_3\times I_3}$ 和 $\boldsymbol{V}_3$ 可求取其语义投影矩阵 $\boldsymbol{T}_3^i=\boldsymbol{V}_3^T\boldsymbol{U}_3^i\in\mathbb{R}^{J_3\times I_3}$.这样,低维语义空间的视频张量镜头集合 Y 中的数据可计算为 $\mathcal{Y}_i|_{i=1}^N=\mathcal{X}_i\times_1\boldsymbol{T}_1^i\times_2\boldsymbol{T}_2^i\times_3\boldsymbol{T}_3^i\in\mathbb{R}^{J_1\times J_2\times J_3}$.算法 1 给出了张量镜头的子空间嵌入和降维流程.

算法 1. 张量镜头的子空间嵌入和降维算法.

输入:原始训练张量镜头集合 $X=\{\mathcal{X}_1,\mathcal{X}_2,...,\mathcal{X}_N\}\in\mathbb{R}^{I_1\times I_2\times I_3}$.

输出:映射后的低维张量镜头集合 $Y=\{\mathcal{Y}_1,\mathcal{Y}_2,...,\mathcal{Y}_N\}\in\mathbb{R}^{J_1\times J_2\times J_3}$;中间转换矩阵 $\boldsymbol{V}_1,\boldsymbol{V}_2,\boldsymbol{V}_3$;以及语义投影矩阵 $\boldsymbol{T}_1^i=\boldsymbol{V}_1^TU_1^i\in\mathbb{R}^{J_1\times I_1},\boldsymbol{T}_2^i=\boldsymbol{V}_2^T\boldsymbol{U}_2^i\in\mathbb{R}^{J_2\times I_2},\boldsymbol{T}_3^i=\boldsymbol{V}_3^T\boldsymbol{U}_3^i\in\mathbb{R}^{J_3\times I_3}$,且满足 $\mathcal{Y}_i|_{i=1}^N=\mathcal{X}_i\times_1\boldsymbol{T}_1^i\times_2\boldsymbol{T}_2^i\times_3\boldsymbol{T}_3^i$.

步骤 1. 为训练集合构建一个最近邻图 $\boldsymbol{G}$;

步骤 2. 按照式(9)计算权重矩阵 $\boldsymbol{W}$;

步骤 3. For k=1 to 3

步骤 4. For i=1 to N

步骤 5. 计算 X_i 的 k 模展开矩阵 $\mathcal{X}_{(k)}^i$ 的 SVD 分解的左矩阵 $\boldsymbol{U}_{(k)}^i$;

步骤 6. End;

步骤 7. $\boldsymbol{D}_U=\sum_i\boldsymbol{D}_{ii}\boldsymbol{U}_{(k)}^i\boldsymbol{U}_{(k)}^{iT}$;

步骤 8. $\boldsymbol{W}_U=\sum_{ij}\boldsymbol{W}_{ij}\boldsymbol{U}_{(k)}^i\boldsymbol{U}_{(k)}^{j\ T}$;

步骤 9. 求解下列广义特征向量问题以得到最优化的 $\boldsymbol{V}_k$: $(\boldsymbol{D}_U-\boldsymbol{W}_U)\boldsymbol{V}_k=\lambda\boldsymbol{D}_U\boldsymbol{V}_k$;

步骤 10. For i=1 to N

步骤 11. $\boldsymbol{T}_{(k)}^i=\boldsymbol{V}_{(k)}^T\boldsymbol{U}_{(k)}^i\in\mathbb{R}^{J_k\times I_k}$;

步骤 12. End;

步骤 13. End;

步骤 14. For i=1 to N

步骤 15. $\mathcal{Y}_i=\mathcal{X}_i\times_1\boldsymbol{T}_1^i\times_2\boldsymbol{T}_2^i\times_3\boldsymbol{T}_3^i$;

步骤 16. End.

2.3 直推式支持张量机

视频语义概念检测的本质是分类问题,因此需要寻找适合的分类器和学习机制.直推式学习是直接从已知样本出发对特定的未知样本进行识别和分类,对本文所研究的视频镜头集语义识别具有实际意义,因此,本文在支持张量机和直推式支持向量机的启发下,提出一种对视频张量镜头进行学习和分类的直推式支持张量机算法.

TSTM 与基于向量表示的 TSVM 不同之处在于,训练及测试样本点均采用降维后的张量来表示,也就是说,TSTM 是 TSVM 在张量空间的自然扩展.当然,与 TSVM 相似的是,TSTM 在训练过程中需要不断修改 STM 分类超平面两侧某些未标记样本的可能标记,使得 STM 在所有数据(包括已标记和未标记)上的分类间隔最大化,从而得到一个既通过相对稀疏分布的数据区域又尽可能正确划分已标记样本的超平面.TSTM 中一个关键的步骤是用支持张量机 STM 来训练分类器,所以下面首先介绍张量镜头的支持张量机训练过程.

2.3.1 支持张量机

根据上一节提出的张量镜头子空间嵌入及降维算法,先对原始的张量镜头 $\mathcal{X}_i$ 进行预处理,得到低维张量 $\mathcal{Y}_i$.这样的处理不仅能够提高精确度,而且能提高训练和分类效率.

对于 l 个已知标记的训练样本,根据算法 1 得到输出降维后的张量镜头 $\mathcal{Y}_i\,|_{i=1}^{l}\in\mathbb{R}^{J_1\times J_2\times J_3}$.

由于算法 1 的方法是线性降维,因此给定 u 个未知标记的测试样本,可以将其直接映射到低维子空间.令 $\mathcal{X}_t$ 是一个训练集合外的测试样本,其降维映射过程见算法 2.

算法 2. 未标记样本的投影算法.

输入:测试未标记样本 $\mathcal{X}_t\in\mathbb{R}^{I_1\times I_2\times I_3}$;中间转换矩阵 $\boldsymbol{V}_1,\boldsymbol{V}_2,\boldsymbol{V}_3$.

输出:降维后的 $\mathcal{Y}_t\in\mathbb{R}^{J_1\times J_2\times J_3}$.

步骤 1. For k=1 to 3

步骤 2. 计算 $\mathcal{X}_t$ 的 k 模展开矩阵 $\mathcal{X}_{(k)}^{i}$ 的 SVD 分解的左矩阵 $\boldsymbol{U}_{(k)}^{i}$;

步骤 3. 计算 $\boldsymbol{T}_{(k)}^{t}=\boldsymbol{V}_{(k)}^{T}\boldsymbol{U}_{(k)}^{t}$;

步骤 4. End;

步骤 5. 计算 $\mathcal{Y}_t=\mathcal{X}_t\times_1\boldsymbol{T}_1^t\times_2\boldsymbol{T}_2^t\times_3\boldsymbol{T}_3^t$;

步骤 6. End.

在对训练和测试张量镜头都进行预处理后,接下来的分类器训练和检测过程中均用降维后的张量镜头 $\mathcal{Y}$ 取代原始的 $\mathcal{X}$,以得到更准确的结果及更好的性能.

在监督张量学习(STL)框架的基础上,文献[22]所推出的支持张量机仅针对已标记样本进行学习,对于视频张量镜头所需求解的优化问题为

$$\left[\begin{array}{l}\min\limits_{\boldsymbol{w}_j,b,\boldsymbol{\xi}} J(\boldsymbol{w}_j,b,\boldsymbol{\xi})=\dfrac{\eta}{2}\left\|\boldsymbol{w}_j\right\|_{Fro}^{2}+c\sum\limits_{i=1}^{N}\xi_i\\ \quad\text{s.t.}\quad y_i\left[\boldsymbol{w}_j^{T}\left(\mathcal{Y}_i\prod\limits_{k=1}^{M}\times_k\boldsymbol{w}_k+b\right)\right]\geq\xi_i,1\leq i\leq N\\ \qquad\xi_i\geq\boldsymbol{0}\end{array}\right.\tag{16}$$

本文针对直推式半监督学习对文献[22]中的 STM 进行了改进,使其能够对所有样本(包括已标记和未标记)同时进行学习,进而求解优化函数,更新分类器.本文提出的 STM 所需求解的优化函数见算法 3 中的步骤 4.两者的区别在于,由于引入了标注和非标注数据,支持张量机和本文提出的扩展直推式支持张量机(即 TSTM)在目标函数以及优化函数上均有所不同.

本文提出的支持张量机 STM 训练分类器的算法 $solve\_stm((\mathcal{Y}_i,y_i)\,|_{i=1}^{l},(\mathcal{Y}_j^{*},y_j)\,|_{j=1}^{u},C,C_+^{*},C_-^{*})$ 见算法 3.

算法 3. 支持张量机训练分类器的算法 solve_stm.

输入:映射后的低维已标记张量镜头(训练)集合 $\mathcal{Y}_i\,|_{i=1}^{l}=\mathcal{X}_i\times_1\boldsymbol{T}_1^i\times_2\boldsymbol{T}_2^i\times_3\boldsymbol{T}_3^i\in\mathbb{R}^{J_1\times J_2\times J_3}$ 及相应的类别标识 $y_i\in\{+1,-1\}$;映射后的低维未标记张量镜头(测试样本)集合 $\mathcal{Y}_j^{*}\,|_{j=1}^{u}=\mathcal{X}_j^{*}\times_1\boldsymbol{T}_1^j\times_2\boldsymbol{T}_2^j\times_3\boldsymbol{T}_3^j\in\mathbb{R}^{J_1\times J_2\times J_3}$ 及在训练过程中被赋予的标记 $y_j^{*}\in\{+1,-1\}$.

参数:C,C_+^{*},C_-^{*} 为优化问题的目标函数中的影响因子,由用户指定和调节.

输出:分类器模型的张量超平面参数:$\boldsymbol{w}_k\,|_{k=1}^{3}\in\mathbb{R}^{J_k}$ 和 $b\in\mathbb{R}$;相应的松弛变量 $\boldsymbol{\xi},\boldsymbol{\xi}^{*}$.

步骤 1. 设置 $\boldsymbol{w}_k\,|_{k=1}^{3}$ 为 $\mathbb{R}^{I_k}$ 中的随机单位向量.

步骤 2. 重复步骤 3~步骤 5 直至收敛.

步骤 3. For m=1 to 3

步骤 4. 通过求解优化问题:

$$\left[\begin{array}{l}\min\limits_{\boldsymbol{w}_m,b,\boldsymbol{\xi},\boldsymbol{\xi}^*} J_{C\text{-}STM}(\boldsymbol{w}_m,b,\boldsymbol{\xi},\boldsymbol{\xi}^*)=\dfrac{\eta}{2}\|\boldsymbol{w}_m\|_{Fro}^2+C\sum\limits_{i=1}^{l}\xi_i+C_+^*\sum\limits_{\substack{1\le j\le u\\ y_j^*=1}}\xi_j^*+C_-^*\sum\limits_{\substack{1\le j\le u\\ y_j^*=-1}}\xi_j^* \\ \text{s.t.}\quad y_i\left[\boldsymbol{w}_m^T\left(y_i\prod\limits_{k=1}^{3}\times_k\boldsymbol{w}_k+b\right)\right]\ge 1-\xi_i, 1\le i\le l \\ \quad\quad y_j^*\left[\boldsymbol{w}_m^T\left(y_i\prod\limits_{k=1}^{3}\times_k\boldsymbol{w}_k+b\right)\right]\ge 1-\xi_j^*, 1\le j\le u \\ \quad\quad \xi_i\ge 0, 1\le i\le l \\ \quad\quad \xi_j^*\ge 0, 1\le j\le u\end{array}\right]$$

得到 $\boldsymbol{w}_j\in\mathbb{R}^{I_j}$,其中调节参数 $\eta=\prod_{1\le k\le 3}^{k\ne m}\|\boldsymbol{w}_k\|_{Fro}^2$.

步骤 5. End.

步骤 6. 检查是否收敛:如果 $\sum_{k=1}^{3}\left[\left|\boldsymbol{w}_{k,t}^T\boldsymbol{w}_{k,t-1}\right|\left(\|\boldsymbol{w}_{k,t}\|_{Fro}^{-2}\right)-1\right]\le\varepsilon$,那么计算得到的 $\boldsymbol{w}_k|_{k=1}^3$ 已经收敛.这里,$\boldsymbol{w}_{k,t}$ 是当前的投影向量,$\boldsymbol{w}_{k,t-1}$ 是前一个投影向量.

步骤 7. End.

2.3.2 直推式支持张量机

Joachinms 的直推式 TSVM 算法中每次只交换 1 对标记相异的测试样本的标记符号以使目标函数值获得最大下降.本文提出每次交换 S 对测试样本的标签,以此来减少到达收敛所需要的训练次数,提高算法的速度.

S 是由用户来调节的参数,当 S=1 时就是原始的 TSVM 算法,而当 S=u/2 则表示可以交换的最大样本对数.算法 4 给出了本文提出的直推式支持张量机对张量镜头进行训练和分类检测的算法过程.与文献[22]中的定理 2 相似,可以证明我们的 TSTM 算法能够在有限次的循环后收敛终止,并输出结果.

算法 4. 张量镜头的直推式支持张量机算法.

输入:映射后的低维已标记张量镜头(训练)集合 $y_i|_{i=1}^l=\mathcal{X}_i\times_1\boldsymbol{T}_1^i\times_2\boldsymbol{T}_2^i\times_3\boldsymbol{T}_3^i\in\mathbb{R}^{J_1\times J_2\times J_3}$ 及相应的类别标识 $y_i\in\{+1,-1\}$;映射后的低维未标记张量镜头(测试样本)集合 $y_j^*|_{j=1}^u=\mathcal{X}_j^*\times_1\boldsymbol{T}_1^j\times_2\boldsymbol{T}_2^j\times_3\boldsymbol{T}_3^j\in\mathbb{R}^{J_1\times J_2\times J_3}$.

参数:C,C_+^*,C_-^* 为优化问题的目标函数中的影响因子,由用户指定和调节;r 表示训练过程中对测试样本集合赋予正标记值的样本个数;S 代表循环中每一次交换的相异标记的样本对数.

输出:测试样本的类别标记 $y_j^*|_{j=1}^u\in\{+1,-1\}$.

步骤 1. 对已标记训练样本进行一次初始学习,得到一个初始分类器,即

$$(\boldsymbol{w}_k|_{k=1}^3,b,\boldsymbol{\xi},\_):=solve\_stm((y_i,\boldsymbol{y}_i)|_{i=1}^l,\square,C,0,0);$$

步骤 2. 用初始分类器 $(\boldsymbol{w}_k|_{k=1}^3,b)$ 对未标记测试样本进行分类,对判别函数输出值 $y_j^*\prod_{k=1}^{3}\times_k\boldsymbol{w}_k+b$ 最大的 r 个未标记样本暂时赋予正标记值,其余未标记样本赋予负标记值;

步骤 3. 指定临时影响因子 $C_+^*=10^{-5}\cdot\dfrac{r}{u-r}$ 及 $C_-^*=10^{-5}$;

步骤 4. While $(C_-^*<C^*)\|(C_+^*<C^*)$

步骤 5. $(\boldsymbol{w}_k|_{k=1}^3,b,\boldsymbol{\xi},\boldsymbol{\xi}^*):=solve\_stm((y_i,\boldsymbol{y}_i)|_{i=1}^l,(y_j^*,\boldsymbol{y}_j)|_{j=1}^u,C,C_+^*,C_-^*);$

步骤 6. While $(\exists p,q:(y_p^*\cdot y_q^*<0)\,\&\,(\xi_p^*>0)\,\&\,(\xi_q^*>0)\,\&\,(\xi_p^*+\xi_q^*>2))$

步骤 7. 取当前标记为正的测试样本中判别函数输出值最小的 S 个,与当前标记为负的测试样本中判别函数输出值最大的 S 个相应地组成 S 个样本对,交换它们的标记值;

步骤 8. 对交换过标记值后的样本集再进行学习,即

$$\left(\boldsymbol{w}_k|_{k=1}^3,b,\boldsymbol{\xi},\boldsymbol{\xi}^*\right):=solve\_stm((y_i,\boldsymbol{y}_i)|_{i=1}^l,(y_j^*,\boldsymbol{y}_j)|_{j=1}^u,C,C_+^*,C_-^*);$$

步骤 9. End;

步骤 10. $C_+^* := \min(C_+^* \cdot 2, C^*)$;

步骤 11. $C_-^* := \min(C_-^* \cdot 2, C^*)$;

步骤 12. End;

步骤 13. 返回测试样本的类别标记值,$y_j^*|_{j=1}^u \in \{+1,-1\}$;

步骤 14. End.

3 实验与讨论

3.1 实验数据

为了验证本文提出的视频语义概念检测方法的有效性,我们采用 TRECVID2005 提供的真实视频数据和标注信息进行测试[30].语义概念检测的目标就是为了发现一个视频镜头中是否存在某个语义概念,实质上就是分类问题.Snoek 等人定义了一个由 101 个多媒体语义概念组成的“词典”[31](如图 1 所示),本文从中选取了“爆炸(Explosion)”“体育(Sports)”“军事(Military)”“飞机(Airplane)”“大厦(Building)”“沙漠(Desert)”“会议(Meeting)”“娱乐(Entertainment)”“囚犯(Prisoner)”“政府官员(Government Leader)”“汽车(Car)”“天气预报(Whether)”等语义概念来测试.这些概念基本上涵盖了新闻视频中人们感兴趣的范围.

Fig.1 Illustrative images of semantic concepts

图 1 语义概念示例图像

TRECVID2005 的视频数据分为 develop 和 test 两部分,其中,develop 数据集合中镜头的类别信息已被标注好,而 test 集合则是没有类别信息的未标注数据.本文将 develop 数据集合划分为已标记(训练)样本集合和交叉验证样本集合.对于训练数据,从训练集合中选取所有正标记的样本,同时以正、负标记 1:5 的比例选取负样本,这样选取是为了在分类器性能与运行效率之间寻找平衡,最终平均为每个语义概念选了 1 200 个训练样本和 300 个交叉验证样本.另外,本文从 test 数据集合中选取了 2 000 个镜头作为未标记(测试)样本数据.

如前文所述,PCA,ISOMAP 及 TensorShot 均为基于流形的降维方法,为了更直观地比较它们对视频数据降维的结果,我们选取了 3 类视频镜头数据,将其降维结果进行了可视化表示,如图 2 所示.

从图 2 中可以看出,与其他两种方法相比,PCA 的结果较差,未能将不同类的数据分开,而是混合到了一起.对于 ISOMAP,虽然 3 类数据能够基本分开,但是仍有部分数据点混杂在一起,类别信息不够明显.而 TensorShot 方法能够有效区分不同类别的数据.

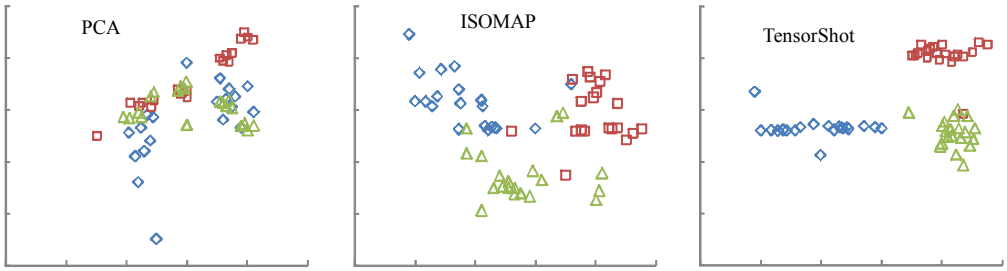

Fig.2 2D visualization of video shot set

图 2 视频镜头数据集的二维可视化表示

3.2 算法评价准则

为了评估检测算法的准确性,本文根据 ROC 曲线(receiver operating characteristic curve)[32]加以说明.ROC 曲线是分类和检索系统中常用的度量参数,能够反映检测算法的准确性和有效性.

在运用分类器对测试集进行分类时,有些样本被正确分类,有些样本被错误分类,这些信息可以通过构造混合矩阵反映出来.这个矩阵是查全率(recall)、查准率(precision)等很多评价标准以及 ROC 分析的基础.如表 1 所示,实际正例数 $P=TP+FN$,实际负例数 $N=FP+TN$,实例总数 $C=P+N$,则有查准率($Precision=TP/(TP+FP)$),查全率 $Recall=TP/P$.ROC 曲线中定义两个概念:错误的正例率(false positive rate,简称 FPR)$FPR=FP/N$;正确的正例率(true positive rate,简称 TPR)$TPR=TP/P$.ROC 曲线是以 FPR 为横轴和 TPR 为纵轴的二维图.在 ROC 曲线图中,通常如果曲线 X 始终位于曲线 Y 的左上方,则曲线 X 优于曲线 Y,也就是说,一个优秀的分类器对应的 ROC 曲线应该尽量靠近单位方形的左上角.同时,为了能够更加直接地比较分类器的性能,希望将 ROC 曲线描述的分类器性能转换为一个数值来表示,目前为止,一种通用的方法是计算 ROC 曲线与横轴间所围成的面积(area under the ROC,简称 AUC).AUC 表示分类器在所有代价比上的平均性能优劣,面积越大表示分类器的预测性能越好.

Table 1 Confusion matrix of two classes classification

表 1 二元分类问题的混合矩阵

		True class	
		+	–
Predicted class	+	True positive (TP)	False positive (FP)
	–	False negative (FN)	True negative (TN)

3.3 实验结果分析

为了验证张量镜头和直推式支持张量机算法的有性,本文分别设计了以下实验,并对实验结果采用文献[33]中的 ROC Analysis 来进行分析、比较.

3.3.1 基于向量表达与基于张量表达降维方法的比较

- PCA+SVM:采用主成分分析 PCA 对特征向量进行降维后,利用 SVM 训练分类器.
- ISOMAP+SVM:采用 ISOMAP 对特征向量进行降维后,利用 SVM 训练分类器.
- TensorShot+STM:对张量镜头进行降维后,采用 STM 训练分类器并对测试数据进行检测.

图 3 中列出了 4 个语义概念:“爆炸”“天气预报”“体育”“军事”,分别采用 PCA+SVM,ISOMAP+SVM 和 TensorShot+STM 三种方法进行分类检测的 ROC 曲线.从图 2 中可以看出,经 PCA 降维后再进行 SVM 训练得到的分类器的结果较差,ISOMAP 降维的结果虽稍好于 PCA,但在性能上远不如本文提出的直推式支持张量机算法对张量镜头的分类检测.因为 PCA 仅仅适用于满足线性分布的数据集,而视频镜头的分布在本质上并不是线性空间.张量镜头方法充分利用了视频中多种模态的特征以及它们之间的时序关联特性,不仅弥补了传统向量拼接带来的“维度灾难”问题,而且能够发现镜头集合所在流形空间的本征结构,是一种有效、快速的线性降维方法.

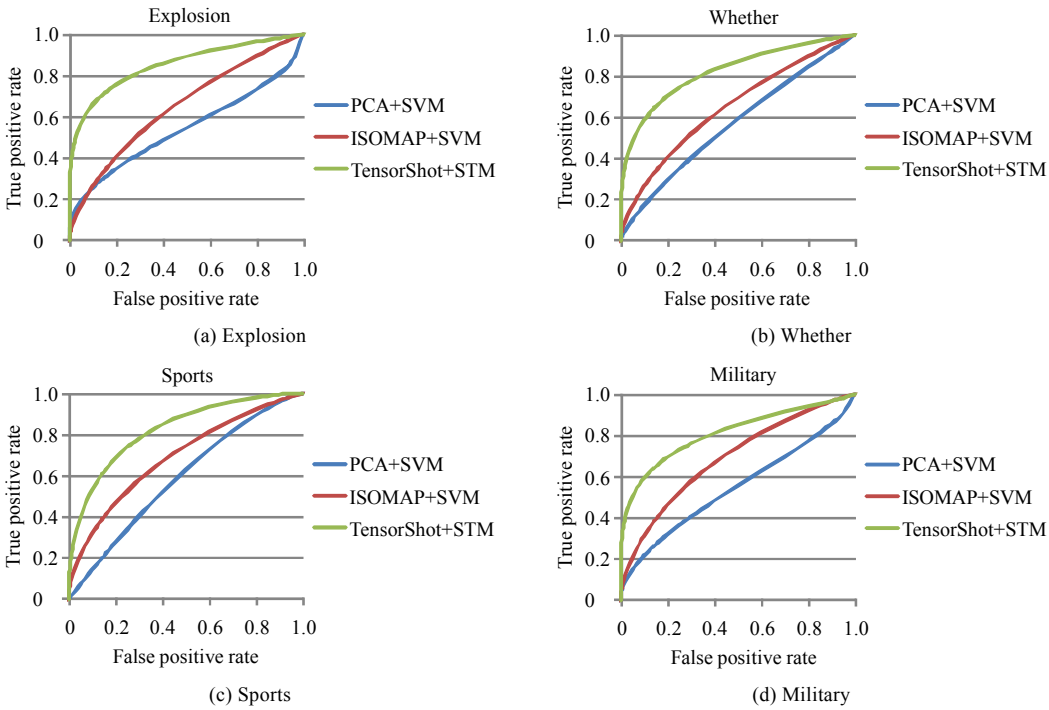

Fig.3 ROC curve of three different dimension reduction approaches for four semantic concepts

图 3 4 个语义概念分别采用 3 种降维方法的 ROC 曲线

3.3.2 基于向量表达与基于张量表达半监督学习方法的比较

- PCA+TSVM:采用主成分分析 PCA 对特征向量进行降维后,利用 TSVM 训练分类器.
- ISOMAP+TSVM:采用 ISOMAP 对特征向量进行降维后,利用 TSVM 训练分类器.
- 本文的算法 TensorShot+TSTM:对张量镜头进行降维后,采用直推式支持向量机训练分类器,并对测试数据进行检测.

图 4 中列出了 4 个语义概念,“爆炸”“天气预报”“体育”“军事”,分别采用 PCA+TSVM,ISOMAP+TSVM 及本文提出的 TensorShot+TSTM 算法进行分类检测的 ROC 曲线.可以看出,虽然基于向量的 TSVM 在一定程度上能够提高分类性能,但是仍低于基于张量的 TSTM 的检测结果.TSTM 充分利用了未标记数据在学习过程中所提供的分布信息,改善学习机和分类器的性能,TSTM 能够良好地对张量镜头进行学习和训练,是 TSVM 对于张量的有效扩展,并且比 TSVM 具有更高的优化和分类能力.

3.3.3 6 种方法的 AUC 值比较

为了更准确地进行比较,表 2 列出了实验中全部语义概念分别用 6 种方法进行分类所得到的 ROC 曲线的 AUC 值.例如,对于概念“爆炸”,TensorShot+TSTM 方法的 AUC 值为 0.892,TensorShot+STM 的 AUC 值为 0.784,ISOMPA+TSVM 为 0.772,ISOMPA+SVM 为 0.659,而 PCA+TSVM 为 0.613,PCA+SVM 为 0.529.可以看出,不仅 TensorShot+STM 算法比基于向量 PCA+SVM 和 ISOMPA+SVM 的 AUC 值都要高,而且 TensorShot+TSTM 算法比 PCA+TSVM 及 ISOMPA+TSVM 也分别高 0.279 和 0.12,同时,TensorShot+TSTM 算法仍然比 TensorShot+STM 高 0.108,效果远高于其他算法,这表明 TensorShot 降维方法与 TSTM 学习方法的结合能够带来更好的分类和检测效果.其他语义概念的 AUC 值也体现了本文所提出算法的优越性.

但是,我们在实验中也发现,不同语义概念所得到的检测结果有很大的区别.某些概念,如“爆炸”“飞机”“大厦”等都得到了很高的正确率,而“囚犯”的检测率则比较低.因此,如何能够针对不同的语义概念“自适应”地选取特征及降维,成为一个很有意义的研究问题.

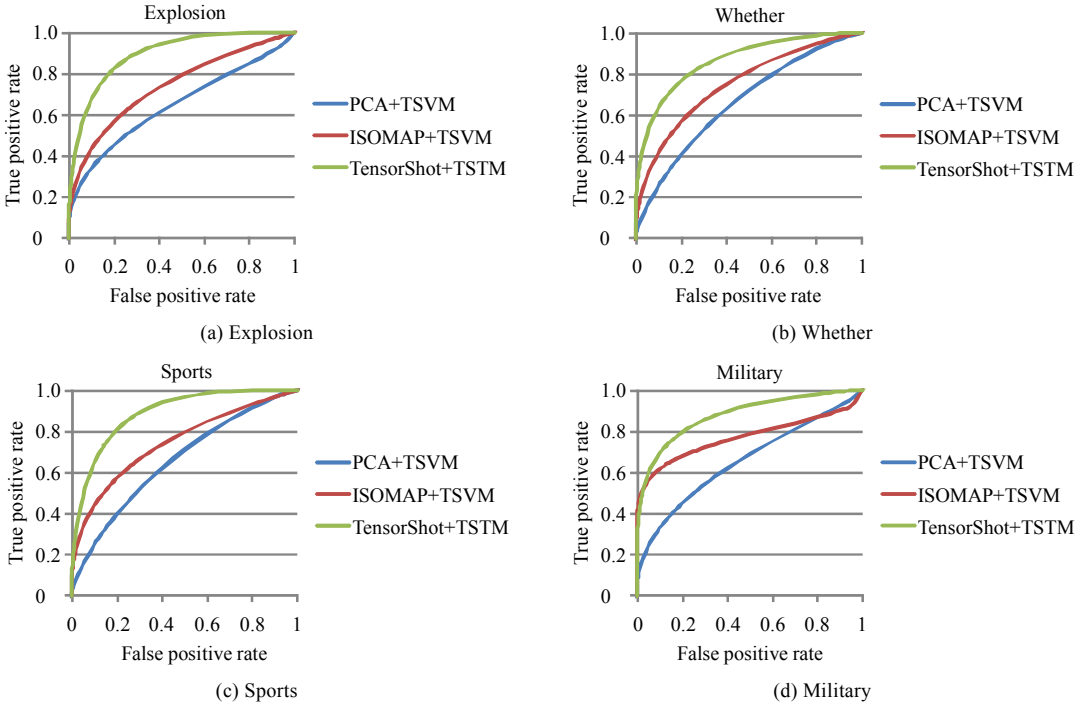

Fig.4 ROC curve of three semi-supervised learning algorithms for four semantic concepts

图 4 4 个语义概念分别采用 3 种半监督学习方法的 ROC 曲线

Table 2 AUC of six different approaches for twelve concepts

表 2 12 个语义概念分别采用 6 种方法的 AUC 值

Semantic concepts	PCA+ SVM	ISOMAP+ SVM	TensorShot+ STM	PCA+ TSVM	ISOMAP+ TSVM	TensorShot+ TSTM
Explosion	0.529	0.659	0.784	0.613	0.772	0.892
Sports	0.522	0.646	0.735	0.605	0.715	0.824
Military	0.507	0.635	0.747	0.596	0.739	0.837
Airplane	0.594	0.702	0.759	0.678	0.768	0.871
Building	0.583	0.71	0.772	0.656	0.784	0.856
Desert	0.524	0.647	0.768	0.627	0.761	0.839
Meeting	0.513	0.572	0.691	0.581	0.675	0.766
Entertainment	0.508	0.587	0.637	0.574	0.633	0.718
Prisoner	0.502	0.541	0.605	0.566	0.598	0.685
Govern_Leader	0.511	0.568	0.679	0.547	0.658	0.736
Car	0.578	0.675	0.755	0.659	0.721	0.845
Weather	0.568	0.736	0.783	0.691	0.769	0.841

4 结 论

本文提出了一个基于张量的视频镜头表达和处理框架.与传统的基于向量的学习方法,如 PCA 和 ISOMAP 不同,该框架将视频所包含的图像、音频和文本等多模态特征进行融合,表达为三阶张量;并在考虑视频多模态的时序关联共生特性及保持张量镜头空间本征结构的基础上,通过流形学习得到低维语义子空间.同时,半监督学习利用了大量的未标记样本来改善学习性能.本文也提出了一种新的直推式支持张量机算法训练分类器来实现对视频镜头的语义概念检测,对于未标记样本的分析和利用使得分类器的性能有了较大的提高.通过对真实的视频数据进行实验并与其他算法比较,其结果表明,本文的算法优于传统的向量表达方法以及仅用已标记数据的监督学习算法.

通过构建语义 Ontology 来缩小“语义鸿沟”,以建立底层特征与高层语义之间的联系是今后的研究重点.同时,如何更加准确地利用未标记数据的分布信息来改进学习性能,也是今后值得研究的问题.

References:

[1] Zhuang YT, Yang Y, Wu F. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. IEEE Trans. on Multimedia, 2008,10(2):221−229.

[2] Yang Y, Zhuang YT, Wu F, Pan YH. Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Trans. on Multimedia, 2008,10(3):437−446.

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

[4] Babaguchi N, Kawai Y, Kitahashi T. Event based indexing of broadcast sports video by intermodal collaboration. IEEE Trans. on Multimedia, 2002,4(1):68−75.

[5] Snoek CGM, Worring M. Multimedia event-based video indexing using time intervals. IEEE Trans. on Multimedia, 2005,7(4): 638−647.

[6] Hu N, Wang YW, Lü N. Study on multimodel retrieval method of content-based video. Journal of Jilin University (Information Science Edition), 2006,24(3):265−270 (in Chinese with English abstract).

[7] Liu YN, Wu F. Video semantic concept detection using multi-modality subspace correlation propagation. In: Proc. of the 13th Int’l Multimedia Modeling Conf. (MMM 2007). Berlin, Heidelberg: Springer-Verlag, 2006. 527−534.

[8] Yu HC, Bennamoun M. 1D-PCA, 2D-PCA to nD-PCA. In: Proc. of the 18th Int’l Conf. on Pattern Recognition. New York: IEEE Computer Society, 2006. 181−184.

[9] Vasilescu MAO, Terzopoulos D. Multilinear analysis of image ensembles: TensorFaces. In: Proc. of the 7th European Conf. on Computer Vision. Berlin, Heidelberg: Springer-Verlag, 2002. 447−460.

[10] Jolliffe IT. Principal Component Analysis. 2nd ed., New York: Springer-Verlag, 2002.

[11] Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000,290(5500):2323−2326.

[12] Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500):2319−2323.

[13] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Proc. of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002. 585−591.

[14] He XF, Niyogi P. Locality preserving projections. In: Proc. of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2003.

[15] Turk MA, Pentland AP. Face recognition using eigenfaces. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. New York: IEEE Computer Society, 1991. 586−591.

[16] Matusik W, Pfister H, Brand M, McMillan L. A data-driven reflectance model. In: Proc. of the SIGGRAPH. New York: ACM, 2003. 759−769.

[17] He XF, Ma WY, Zhang HJ. Learning an image manifold for retrieval. In: Proc. of the ACM Conf. on Multimedia. New York: ACM, 2004. 17−23.

[18] Tang JH, Hua XS, Qi GJ, Wang M, Mei T, Wu XQ. Structure-Sensitive manifold ranking for video concept detection. In: Proc. of the ACM Conf. on Multimedia. New York: ACM, 2007. 852−861.

[19] Hoi SCH, Lyu MR. A multimodal and multilevel ranking scheme for large-scale video retrieval. IEEE Trans. on Multimedia, 2008, 10(4):607−619.

[20] He XF, Cai D, Liu HF, Han JW. Image clustering with tensor representation. In: Proc. of the ACM Conf. on Multimedia. New York: ACM, 2005. 132−140.

[21] de Lathauwer L, Moor BD, Vandewalle J. A multilinear singular value decomposition. SIAM Journal of Matrix Analysis and Applications, 2000,21(4):1253−1278.

[22] Tao DC, Li XL, Wu XD, Hu WM, Maybank SJ. Supervised tensor learning. Knowledge and Information Systems, 2007,13(1): 1−42.

[23] Burges CJC. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998,2(2): 121−167.

[24] Zhu XJ. Semi-Supervised learning literature survey. Technical Report, 1530, Department of Computer Science, University of Wisconsin-Madison, 2005.

[25] Joachims T. Transductive inference for text classification using support vector machines. In: Bratko I, Dzeroski S, eds. Proc. of the 16th Int'l Conf. on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 1999. 200−209.

[26] Lathauwer LD. Signal processing based on multilinear algebra [Ph.D. Thesis]. Belgium: Katholieke Universiteit Leuven, 1997.

[27] Chen YS, Wang GP, Dong SH. A progressive transductive inference algorithm based on support vector machine. Journal of Software, 2003,14(3):451−460 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/451.htm

[28] He XF, Cai D, Niyogi P. Tensor subspace analysis. In: Proc. of the Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2005. 499−506.

[29] Chung FRK. Spectral graph theory. 2nd ed., Providence: American Mathematical Society, 1997. 2−14.

[30] TREVID: TREC video retrieval evaluation. http://www-nlpir.nist.gov/projects/trecvid

[31] Snoek CGM, Worring M, van Gemert JC, Geusebroek JM, Smeulders AWM. The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proc. of the ACM Int'l Conf. on Multimedia. New York: ACM, 2006. 421−430.

[32] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006,27(8):861−874.

[33] Eng J. ROC analysis: Web-Based calculator for ROC curves. Baltimore: Johns Hopkins University. http://www.jrocfit.org

附中文参考文献:

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

[6] 胡楠,王英武,吕凝.基于内容的视频多模式检索方法.吉林大学学报(信息科学版),2006,24(3):265−270.

[27] 陈毅松,汪国平,董士海.基于支持向量机的渐进直推式分类学习算法.软件学报,2003,14(3):451−460. http://www.jos.org.cn/1000-9825/14/451.htm

吴飞(1973－),男,湖南冷水江人,博士,副教授,CCF 高级会员,主要研究领域为多媒体分析与检索,计算机动画,统计学习理论.

刘亚楠(1982－),女,博士生,主要研究领域为多媒体分析与检索,视频分析与理解,机器学习.

庄越挺(1965－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为多媒体检索,计算机动画,人工智能,计算机图形学,数字图书馆.