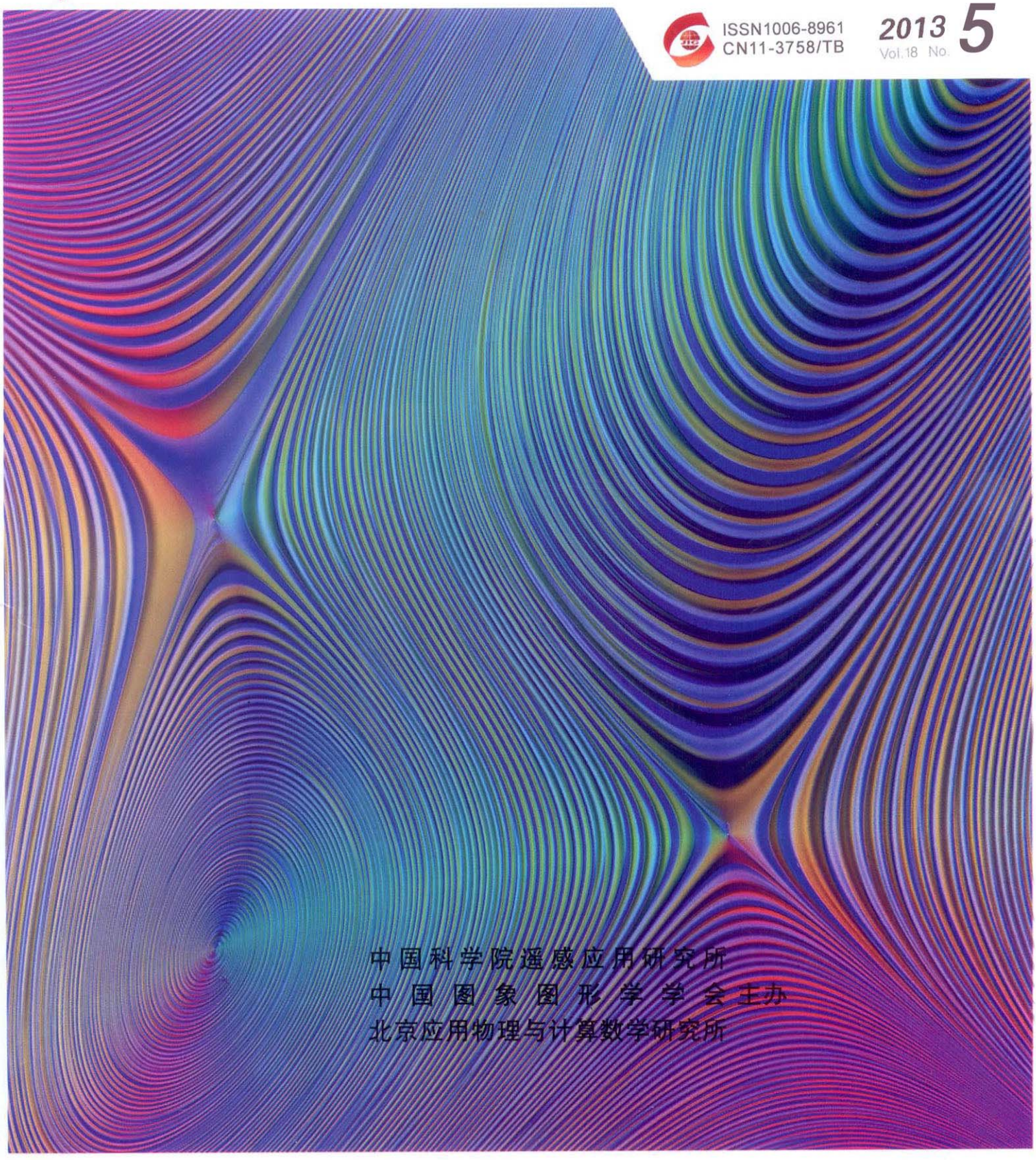
Journal of Image and Graphics

中国图象图形学报

ISSN1006-8961
CN11-3758/TB

2013 Vol.18 No. 5



中国科学院遥感应用研究所
中国图象图形学学会主办
北京应用物理与计算数学研究所

中国精品科技期刊

中国图象图形学报

Zhongguo Tuxiang Tuxing Xuebao

2013 年 5 月　第 18 卷　第 5 期(总第 205 期)

目　次

Journal of Image and Graphics

(**Monthly, Started in 1996**)

Vol. 18 No. 5 May 2013

Contents

中图法分类号: TP391 **文献标识码**: A **文章编号**: 1006-8961(2013)05-0529-08
论文引用格式: 刘咏梅,杨帆,于林森. 场景语义树图像标注方法[J]. 中国图象图形学报,2013,18(5):529-536.

场景语义树图像标注方法

刘咏梅[1], 杨帆[1], 于林森[2]
1. 哈尔滨工程大学 计算机科学与技术学院, 哈尔滨 150001;
2. 哈尔滨理工大学 计算机科学与技术学院, 哈尔滨 150080

摘 要: 自动图像标注是一项具有挑战性的工作,它对于图像分析理解和图像检索都有着重要的意义。在自动图像标注领域,通过对已标注图像集的学习,建立语义概念空间与视觉特征空间之间的关系模型,并用这个模型对未标注的图像集进行标注。由于低高级语义之间错综复杂的对应关系,使目前自动图像标注的精度仍然较低。而在场景约束条件下可以简化标注与视觉特征之间的映射关系,提高自动标注的可靠性。因此提出一种基于场景语义树的图像标注方法。首先对用于学习的标注图像进行自动的语义场景聚类,对每个场景语义类别生成视觉场景空间,然后对每个场景空间建立相应的语义树。对待标注图像,确定其语义类别后,通过相应的场景语义树,获得图像的最终标注。在 Corel5K 图像集上,获得了优于 TM(translation model)、CMRM(cross media relevance model)、CRM(continous-space relevance model)、PLSA-GMM(概率潜在语义分析-高期混合模型)等模型的标注结果。

关键词: 自动图像标注;图像场景;场景语义树;图像聚类

Automatic image annotation based on scene semantic trees

Liu Yongmei[1], Yang Fan[1], Yu Linsen[2]
1. *School of Computer Science and Technology, Harbin Engineering University, Harbin* 150001, *China*;
2. *School of Computer Science and Technology, Harbin University of Science and Technology, Harbin* 150080, *China*

Abstract: Automatic image annotation (AIA) is an important and challenging job for image analysis and understanding such as Content-Based Image Retrieval (CBIR). In AIA, a model for annotation which represents the relationship between the semantic concept space and the visual feature space, is constructed by learning from annotated image datasets. The performance of the AIA is still poor because the relationship between the key words and visual features is too complicated due to the semantic gap. However, with constrains under the scenes, the correlation between them becomes simpler and clearer for better annotation results. In this paper, a method of automatic image annotation based on scene semantic tree is proposed. Image scenes are obtained through the annotated words using PLSA. Scene semantic trees are constructed for each image scene. Un-annotated images are classified into certain scenes with visual features and are annotated using the corresponding scene semantic tree. Using the visual features provided by Duygulu, the experiments get the more effective results on Corel5K database than TM, CMRM, CRM and PLSA-GMM.

Key words: automatic image annotation; image scenes; scene semantic tree; image clustering

收稿日期:2012-07-05;**修回日期**:2012-10-25
基金项目:中央高校基本科研业务费专项(HEUCF100604)
第一作者简介:刘咏梅(1973—),女,副教授,2005 年于哈尔滨工程大学获计算机科学与技术专业博士学位,主要研究方向为图像理解、模式识别。E-mail:liuyongmei@ hrbeu. edu. cn

0 引 言

在基于内容的图像检索(CBIR)领域,由于图像的低级视觉特征和高级语义之间的鸿沟,使得检索效果很难达到人们满意。而且,人们实际上更加倾向于利用关键词或短语来给出查询条件,因此图像标注就显得十分重要了。图像标注共分为3种类型:手工标注、半自动标注和自动标注。虽然自动图像标注(AIA)的正确率尚不及手工标注和半自动标注,但其效率较高,节省了人力和时间。目前,自动图像标注已经成为图像理解领域的重要研究课题。

随着高质量图像标注信息的逐步增多,为探求视觉内容与标注字之间关联性提供了大量可靠的学习样本。到目前为止,已经出现了不少行之有效的标注方法。较早地,Mori等人[1]提出的共现方法通过统计学习来获得图像与标注字之间的关系。Duygulu等人[2]提出的机器翻译模型从物体识别角度出发,将标注图像分割成与物体相对应的图像区域,将连续的区域视觉特征量化为离散的视觉词汇表,采用了一类经典的统计翻译机模型,将构成图像的一组区域视觉特征翻译成相应的标注字。这种在物体视觉区域与代表高级语义的标注字之间建立对应关系的方法为自动图像标注提供了新的指导方法。在此之后,中外学者提出了许多图像标注的统计模型,如CMRM(cross modia relevance model)[3]、CRM(continous-space relevance model)[4]、最大熵模型[5]、扩展生成语言模型[6]和图学习模型[7]等。Monay和Sivic[8-9]利用自然语言理解中的LSA(latent semantic analysis)和PLSA(probabilistic latent semantic analysis)模型分析图像与标注字之间的关系,这种做法在目前较为流行。此外,语义上下文关系被用来在视觉特征和高级语义之间建立联系[10-11]。但是,与作为语义内容描述的标注字信息相比,毕竟视觉词汇与高级语义内容之间还存在巨大差距,因此仅依赖图像的视觉特征,必然会制约语义类别提取结果的有效性。同时也说明,如何更加充分地利用标注字信息还没有引起广泛的关注。

总体而言,目前自动图像标注的各类方法仍然无法提供一个令人十分满意的效果,原因主要在于视觉特征与标注字之间的关系过于复杂。标注字存在一词多义和多词同义现象,以及同一语义的物体可能在不同的场景中表现出不同的视觉效果,导致图像的分割区域和标注字之间的联系十分复杂。从本质上讲,难度仍然来自于语义鸿沟。由于描述能力的不足,图像视觉特征存在明显的歧义性,视觉相似的图像无法保证语义内容的一致性,因此仅依赖于图像的视觉特征显然无法取得良好的标注效果。

图像所能表达的语义内容十分丰富,一幅图像放在不同的环境下,可能呈现出不同层面的信息。如果能将学习样本进行分类,分成许许多多个不同的语义场景,那么在特定的场景下,标注字的歧义性会明显降低,图像区域的视觉特征所能表征的语义范畴也会显著缩小,从而视觉特征和标注字之间的联系也会变得更加清晰。这样,就可以将一个具有复杂联系的学习问题分解为许多简单联系的子问题,能够突破以往图像识别所受类别规模的限制,明显降低学习复杂程度。本文思想是采用一组图像来突出所要传递的语义内容,这组图像可以用来构建一个语义场景。利用代表图像高级语义的标注字信息,对学习图像首先进行场景分析,以确保在特定语义场景下获得比较完善的图像视觉描述,从而提高通过图像视觉特征进行自动语义标注的可靠性。另一方面,将学习图像按照场景分类后,就可以针对具有少量标注字的每个场景单独进行学习,还可以采取并行处理的方式,在不同机器上同时训练以加快学习的速度。

针对这一思想,借助于场景语义的过渡,提出了一种场景语义树图像标注方法(SSTM),采用标注字信息利用PLSA(probability latent semantic analysis)模型进行场景聚类,利用高斯混合模型(GMM)建立视觉场景空间,对特定场景的图像建立一种树型结构用于标注该场景下的待标注图像。本文方法主要分为两个步骤:1)在训练过程中,对学习用的标注图像库利用抽取到的语义特征进行自动的场景分类,获得每个类别的特征场景空间,对每个场景空间再利用视觉特征构建该场景的语义树;2)在标注阶段,对待标注图像先进行场景分类,然后调用该场景的语义树利用视觉特征进行标注。

1 自动场景描述

1.1 PLSA场景聚类

由于希望通过用一组图像来突出一个特定的场景,从而增加标注的语义内容一致性,首先对用于学习的图像进行PLSA场景聚类,每个场景得到一组

图像。与仅依靠视觉特征进行语义类别提取不同,以标注图像为研究对象,不仅包括视觉特征还包括标注字,而且标注字在语义聚类中显得尤为重要。为了保证聚类结果的语义一致性和 PLSA 模型的合理性,利用标注字描述方式,对学习样本图像进行自动场景语义聚类。

PLSA 模型将图像中的每个标注字看做一个混合模型的采样,这个混合模型的每个分量代表一个潜在的主题。每个标注字都从单个的主题中产生,而不同图像中的不同标注字从不同的潜在主题产生。图像的潜在主题可以理解为所包含的抽象目标类别如草地、天空、马等。PLSA 方法通过极大化观测数据和潜在主题变量的对数似然函数得到训练图像的潜在主题分布。而对于希望进行场景分类的测试图像则采用叠入方法,固定单词的潜在主题分布来计算测试图像的潜在语义表示。

采用图像处理语言的方式来描述模型。给定一组训练图像 $\boldsymbol{D}=(d_1,d_2,\cdots,d_M)$,每个图像均包含若干个局部区域,这些区域被量化为包含 N 个标注字的集合 $\boldsymbol{W}=(w_1,w_2,\cdots,w_N)$。因此训练图像的集合就可以由一个图像-单词的互共生矩阵 $\boldsymbol{N}=(n_{ij})$ 来表示,其中 $n_{ij}=n(d_i,w_j)$ 表示词 w_j 在图像 d_i 中出现的频率。用 $\boldsymbol{Z}=\{z_1,z_2,\cdots,z_K\}$ 表示 K 个潜在语义的集合。则每个图像产生的概率都可以看作是条件概率 $P(z_k|d_i)$ 的混合。而这个条件概率被称做因子模型。这样的因子模型可以看做是潜在变量产生时视觉词汇出现的概率 $P(w_j|z_k)$ 的组合。这个生成过程可以用下面的联合概率密度函数表示为

$$P(d_i,w_j)=P(d_i)\sum_{k=1}^{K}P(w_j|z_k)P(z_k|d_i) \quad (1)$$

式中,$P(w_j|z_k)$ 为潜在语义在视觉词汇上的分布概率,也表示为该词对潜在语义的贡献度;$P(z_k|d_i)$ 表示图像中具有相应潜在语义的分布概率。根据极大似然准则,可以通过极大化对数似然函数 L 来进行模型的参数估计,即

$$L=\sum_i\sum_j n(d_i,w_j)\log P(d_i,w_j) \quad (2)$$

这样的极大似然解问题可以通过标准的 EM(期望最大化)算法来估计参数。

1.2　场景语义视觉空间的学习

对于待标注的图像,仍然需要利用视觉特征对其进行场景归类,因此在对学习图像进行场景聚类后,需要对不同的语义场景图像的视觉空间进行非监督学习,然后按照待标注图像对不同场景视觉空间的符合程度,将其归属为不同的语义类别。利用 GMM 模型建立场景空间。

对服从混合模型分布的位于 d 维空间中的数据 $\boldsymbol{x}$,它的概率密度函数可表示为

$$f(\boldsymbol{x}|\boldsymbol{\Theta})=\sum_{k=1}^{K}\pi_k f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)\quad \forall x\in\mathbf{R}^d \quad (3)$$

式中,$\boldsymbol{\Theta}=(\pi_1,\cdots,\pi_K,\theta_1,\cdots,\theta_K)$ 为参数向量,K 为混合分量的个数,$\pi_k\in(0,1)$($\forall k=1,2,\cdots,K$)为混合分量的先验概率,满足 $\sum_{k=1}^{K}\pi_k=1$。对于高斯混合模型,每个分量 $f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)$ 是一个高斯分布函数,即

$$f_k(\boldsymbol{x}|\boldsymbol{\theta}_k)=\frac{1}{\sqrt{(2\pi)^d\det(\boldsymbol{\Sigma}_k)}}\times \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} \quad (4)$$

式中,$\boldsymbol{\theta}_k=(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ 为第 k 个混合分量的参数。

EM 算法是从不完全数据中求解模型参数的一类常用方法[9]。不完全数据是指观测到的数据 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_N\}$,而完全数据则是由观测到的数据 $\boldsymbol{X}$ 和指明数据来源的随机变量 $\boldsymbol{Z}=\{z_1,\cdots,z_N\}$ 组成,其中 z_i 用于指明观测数据 x_i 来源于那个混合分量。在无法确定数据来源的情况下,EM 算法通过对不完全数据的 log 似然函数的极大化,获取模型参数 $\boldsymbol{\Theta}$ 的估计值。对于独立同分布的数据,$L(\boldsymbol{\Theta})$ 可表示为

$$L(\boldsymbol{\Theta})=\sum_{i=1}^{N}\log f(x_i|\boldsymbol{\Theta}) \quad (5)$$

标准 EM 算法由 E 步骤和 M 步骤的迭代组成。设 $\boldsymbol{\Theta}^{(t)}$ 为第 i 次迭代后获得的参数估计值,则在第 $i+1$ 次迭代时,E 步骤计算完全数据 log 似然函数的期望

$$Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})=E_p(\log p(\boldsymbol{X},\boldsymbol{Z};\boldsymbol{\Theta}))=\sum_{k=1}^{K}\sum_{i=1}^{N}(\log\pi_k f_k(x_i|\boldsymbol{\theta}_k))p(k|x_i;\boldsymbol{\Theta}^{(t)}) \quad (6)$$

式中,$p(k|x_i;\boldsymbol{\Theta})=p(z_i=k|x_i;\boldsymbol{\Theta})=\frac{\pi_k f_k(x_i|\boldsymbol{\theta}_k)}{f(x_i|\boldsymbol{\Theta})}$ 为贝叶斯后验概率。

M 步骤求取模型参数的极大似然估计值

$$\boldsymbol{\Theta}^{(t+1)}=\arg\max_{\boldsymbol{\Theta}}Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) \quad (7)$$

2 场景语义树的构建

PLSA-GMM[12]的图像标注方法有其局限性,首先每个场景下图像所包含的标注字远大于欲标注的标注字个数,而且不同标注字与场景语义的关系也十分复杂,词频分布也较为分散,定长标注限制了候选标注字的范围,从而影响了标注精度。

为此,在 PLSA-GMM 基准模型的基础上,提出了利用场景语义树来进行图像标注的方法,该方法不直接使用场景的语义主题作为待标注图像的标注,而是对每个场景构建语义树,对场景语义进行合理的深入挖掘。

对于特定的场景,建立一个与之对应的语义树,其中根节点对应该场景出现频率最高的常用标注字,而叶子节点对应该场景内一些较为特别标注字。随着树的生长,相应叶子节点的标注字的出现频率逐渐降低。同时,该场景语义树的建立仍然通过视觉特征,因此,该树也有其视觉层面的含义,即在根节点处聚集了该场景的所有的训练图像,某节点处对应的图像汇集了其左子节点和右子节点处的图像,在树的叶子节点处图像最少。对待标注图像,通过判别其场景归属,再从该场景语义树的根部游走到某个叶子节点,得到该路径下相应的标注字。

SSTM 方法建立的场景语义树设为 $\boldsymbol{F}=(t_1,t_2,\cdots,t_K)$,其中 $t_k=\{t_v^k,t_{\text{word}}^k\}$ 是一个二元组,t_v^k 代表了第 k 个场景语义树中图像在各个节点的分布,相当于视觉层,t_{word}^k 代表第 k 个场景语义树的标注字分布,相当于语义层,先建立树的视觉层,再得到相应的语义层。对每个场景下的图像集 $\{c_{\text{img}}^i,i=1,2,\cdots,k\}$ 建立一个二叉树结构,得到一组各个节点上的图像序列 $\boldsymbol{C}^k=(I_1,I_2,\cdots,I_n)$,$n$ 为该二叉树的所有节点的个数。场景语义树构造算法具体如下:

输入模型语义层数据 $\boldsymbol{C}=(c_1,c_2,\cdots,c_k)$ 和二叉树的最大深度 h_{depth}。

1)开始周期为 k 的循环,取得 c_{img}^k。

2)设 $\boldsymbol{L}=\{\boldsymbol{C}^K,\boldsymbol{A}^K\}$,计算 $\boldsymbol{W}^K,\boldsymbol{D}^K$,将 $\boldsymbol{C}^K$ 和 $\boldsymbol{A}^K$ 分别存入 t_v^k 和 t_{word}^k。

3)开始建立二叉树,深度为 h,取值区间为[1,h_{depth}],递增步长为 1。

4)开始在二叉树深度为 h 的层中进行由左至右的遍历,取得每个节点的图像编号并设为 $\boldsymbol{L}_i$。

5)输入 $\boldsymbol{L}_i$、$\boldsymbol{W}_i$ 和 $\boldsymbol{D}_i$,使用 N-Cuts(规格化切分)算法[13]对深度为 h 层的节点图像进行二分,得到 $\boldsymbol{L}_{\text{new}}^h=(\boldsymbol{L}_1,\boldsymbol{L}_2,\cdots,\boldsymbol{L}_i)$,$i\in[1,2,\cdots,h]$,$\boldsymbol{L}_i=\{\{\boldsymbol{L}_i^l,\boldsymbol{L}_i^r\},\{\boldsymbol{A}_i^l,\boldsymbol{A}_i^r\}\}$;$h=h+1$,若 $h>h_{\text{depth}}$ 则执行步骤 6),否则执行步骤 4)。

6)将 $\boldsymbol{L}_{\text{new}}^h$ 中的 $\boldsymbol{L}^{l,r}$ 和 $\boldsymbol{A}^{l,r}$ 分别存入 t_v^k 和 t_{word}^k。

输出:$\boldsymbol{F}=(t_1,t_2,\cdots,t_K)$。

在建树算法的描述中,$\boldsymbol{W}^K$ 是由 w_{ij} 构成的 $n\times n$ 对称阵,w_{ij} 表示图像 I_i 和 I_j 的相似性度量结果,$\boldsymbol{D}^K$ 是由 d 构成的 $n\times n$ 对角阵,$d(i)=\sum_j w(i,j)$,$\boldsymbol{W}_i\subset\boldsymbol{W}^K$,$\boldsymbol{D}_i\subset\boldsymbol{D}^K$,$\boldsymbol{A}$ 为节点中剔除了父节点高频标注字的本节点剩余高频标注字,在满足该节点不同高频标注字的频率相等时,允许 $length(\boldsymbol{A})\geqslant 1$,$length(\boldsymbol{A})$ 表示 $\boldsymbol{A}$ 中包含标注字的数量,在使用 N-Cuts 算法对图像集合进行二分时,将单幅图像视为图的节点、图像之间的视觉距离视为边的权值。

3 SSTM 方法的整体框架

结合自动场景描述和场景语义树的构建,SSTM 方法的步骤如下:

1)采用 N-Cuts 分割算法对用于学习的标注图像进行分割,获得图像区域的视觉描述。

2)对用于学习的标注图像进行自动的语义场景聚类。

对用于学习的标注图像,计算其标注字权值向量,向量的维数为图像集所有标注字的个数,每一维代表 1 个标注字,每一维的元素值为该图像中如果有相应标注字时所占整个标注字的比重。例如,一幅图像如果含有 bear、water、ground 3 个标注字,那么标注字权值向量中,在 3 个标注字相应位置的权值为 1/3,其余为 0。然后利用各个图像的标注字权值向量进行 PLSA 聚类。

3)对每个场景语义类别生成特征场景空间。

通过步骤 2),标注类似的图像聚集在一起,就构成了一个语义场景。因为与之相应的标注字可以获得较高的权值,因此语义内容的重点也就随之凸显出来。对每个语义类别,将该类别的图像采用 GMM 模型对特征空间进行描述。

4)对每个特征场景空间建立相应的语义树。

对特定场景下的学习图像,利用推土机距离

(EMD)[14]和 TF-IDF 距离[15]的联合距离[16]计算两两图像间的距离[17],建立视觉近邻图。该图的节点代表图像,任意两点的边代表两个图像间的距离。对该视觉近邻图,采用 N-Cut 算法换为一棵语义二叉树。方法如下:首先树的根节点对应场景中所有图像,利用 N-Cuts 算法对视觉近邻图进行二分,这样就产生了根节点的左右子树所对应的根节点,以此类推,直到不可二分、只剩下 1 个图像或超出规定深度为止。该二叉树的另一层含义是:每个节点处都对应一些图像,通过这些图像可以凸现出该节点处显著的标注字,根节点处的标注字在该场景的所有图像中出现频度最高,代表整个场景的语义,该场景下出现频度越低的标注字出现在越靠近树的底端,叶子节点对应的标注字出现频度最低。

5)对待标注图像,利用与学习图像同样的方法进行图像分割,获得图像区域视觉描述。然后向各个特征场景空间投影,并采用投影后的视觉特征对混合模型的拟合程度确定该待标注图像的语义类别。确定了语义类别后,通过调用相应的场景语义树,获得图像的最终标注。

从根节点开始,计算与其左右子节点处图像的平均 EMD 视觉距离来判断该待标注图像是向左还是向右游走的方向,直到游走到任一叶子节点,这样,就得到了从根节点到该叶子节点的一条路径。然后将该路径下所有节点处对应的标注字赋予给该待标注图像。

SSTM 方法的训练和标注过程分别如图 1、图 2 所示。

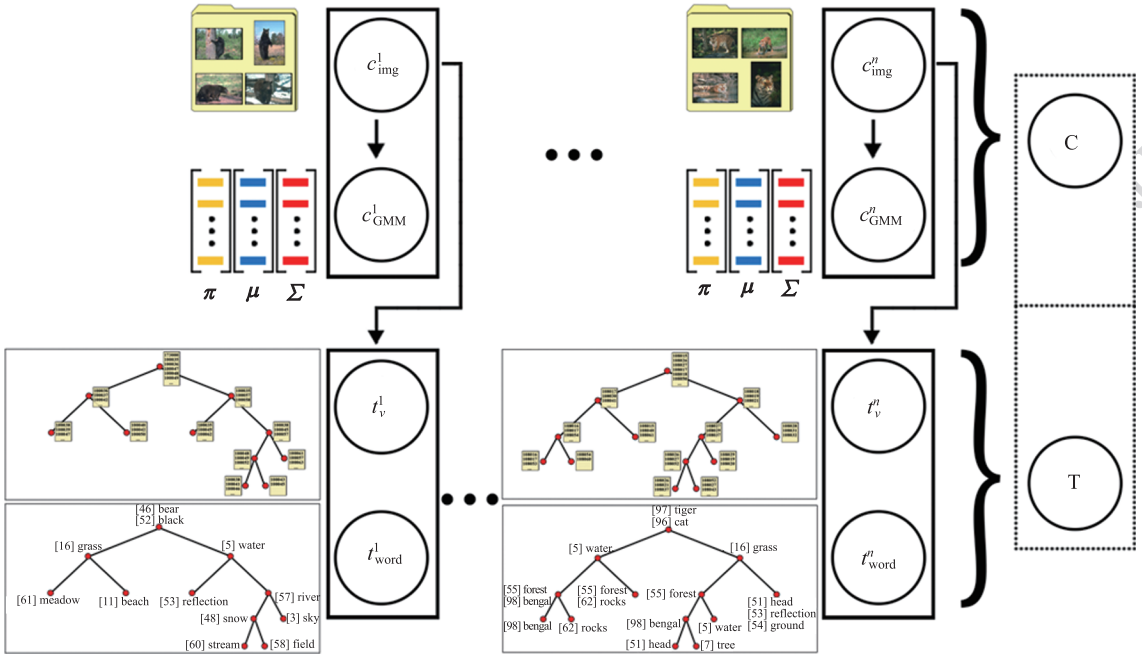


图 1　SSTM 方法的训练过程

Fig. 1　Training process of SSTM method

SSTM 方法的训练过程分为 C(clusters)步和 T(trees)步。在 C 步中,对用于学习的标注图像,通过 PLSA 方法利用标注字权值聚类场景语义,得到 n 个聚类。对所有的场景聚类结果 $\{c^i_{img}\}^n_{i=1}$,利用 GMM 模型建立特征场景空间 $\{c^i_{GMM}\}^n_{i=1}$,以便于对待标注图像的场景分类。在 T 步中,首先利用视觉相似性测度和 N-Cut 算法对图像进行不断二分建立场景二叉树的视觉层 $\{t^i_v\}^n_{i=1}$,然后由视觉层统计得到该场景二叉树的语义层 $\{t^i_{word}\}^n_{i=1}$,获得场景语义树。

SSTM 方法的标注过程仍然分为 C 步和 T 步。在 C 步中,对于 1 幅待标注图像 $I^{?,?}_{test}$,由于场景特征空间已经被表示为多个 GMM 模型,因此利用其视觉特征直接与各个 GMMs 进行比较,将该待标注图像归为某一特定场景 n。在 T 步中,调用相应场景的语义树,待标注图像 $I^{n,?}_{test}$ 利用视觉特征从根节点开始向下游走,直到叶子为止,同时获得该路径下的标注字,此时图像 $I^{n,annotation}_{test}$ 已经完成标注。

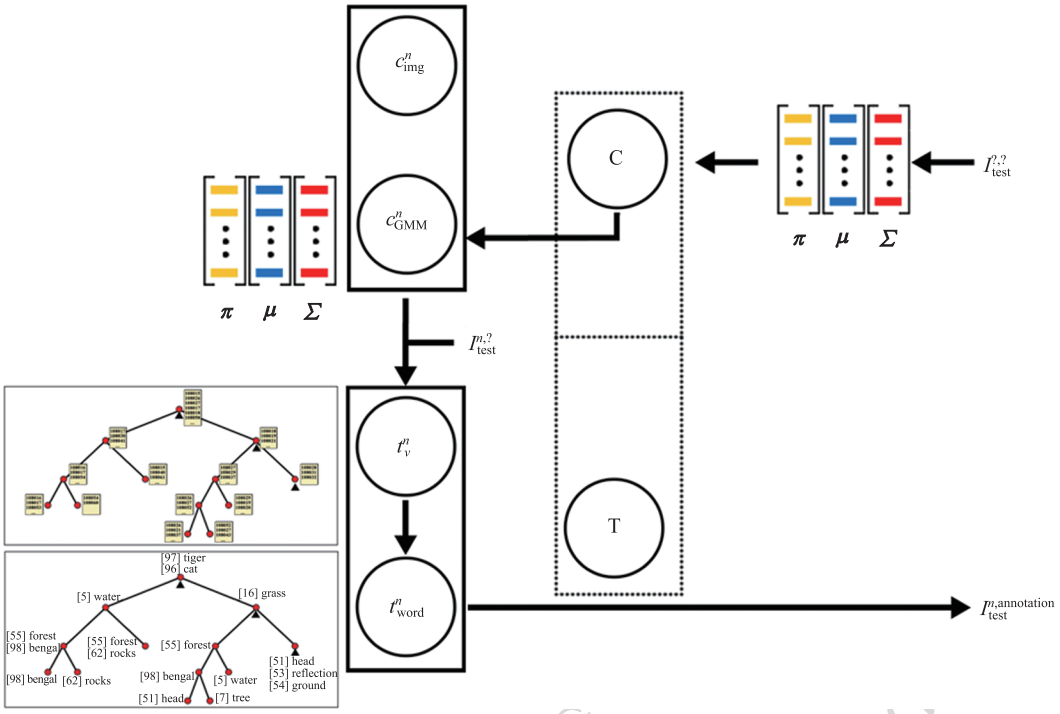


图 2 SSTM 方法的标注过程

Fig. 2 Annotation process of SSTM method

4 实验结果分析

为了测试本文方法的有效性和正确率，在 Corel5K 标注图像集上进行实验。Corel5K 标注图像集包括 50 张 CD 上的 5 000 幅标注图像，每张 CD 包括语义主题相同的 100 幅图像，其中 90 幅为训练图像，10 幅为测试图像。因此训练图像集包括4 500 幅图像，测试图像集包括 500 幅图像。全部标注字的数量为 374 个，每幅图像平均带有 3～5 个标注字。采用由 Duygulu 提供的图像视觉描述方式[2]，所有的 5 000 幅图像均 N-Cuts 方法分割，选出前 10 个面积最大的区域作为有效区域，对每个有效分割区域提取 6 维形状特征、18 维颜色特征、12 维纹理特征，共 36 维特征。

在对训练图像的场景聚类中，选取聚类数目 $n=50$。在场景视觉空间的建模中，GMM 模型的混合的个数由该场景下图像中出现的标注字的个数。

在对图像检索性能的评价中，查全率（recall）度量出对单个词查询的完整性，查准率（precision）反映出查询的精度。在对图像标注结果的评价中，对于给定的标注字 w，若在测试图像集的手工标注结果中包含 w 的图像个数为 N_{men}，使用自动标注模型的标注结果中包含该词的图像个数为 N_{auto}，其中参照手工标注结果有 $N_{correct}$ 个是正确的，则单个标注字的查全率和查准率分别为

$$R=\frac{N_{correct}}{N_{men}} \tag{8}$$

$$P=\frac{N_{correct}}{N_{auto}} \tag{9}$$

采用统计测试图像集中 500 幅图像所包含的全部 260 个标注字的平均查准率和平均查全率来考察本文方法的标注性能。

本文方法（SSTM）与 TM、CMRM、CRM 和 PLSA-GMM 模型进行了比较，这些方法也同样采用了 Duygulu 提供的图像视觉特征。模型的性能比较结果如表 1 和表 2 所示，表 3 是一些测试图像的标注结果实例。

表 1 模型的性能比较（性能最好的 49 个标注字）

Table 1 Performance comparison for models on best 49 words

	TM	CMRM	CRM	PLSA-GMM	SSTM
平均查全率	0.34	0.48	0.70	0.73	0.74
平均查准率	0.20	0.40	0.59	0.59	0.67

表 2　模型的性能比较(所有 260 个标注字)

Table 2　Performance comparison for models on all 260 words

	TM	CMRM	CRM	PLSA-GMM	SSTM
平均查全率	0.04	0.09	0.19	0.18	0.19
平均查准率	0.06	0.10	0.16	0.14	0.17

从实验结果不难看出,SSTM 比 TM、CMRM、CRM 和 PLSA-GMM 模型在标注性能上有所提高。SSTM 方法通过对每个语义场景建立的二叉树结构可以实现对待标注图像的变长标注,和 PLSA-GMM 模型进行比较,尤其在体现标注精度的查准率上有明显提高。

表 3　标注结果实例

Table 3　Examples for annotation results

标注方法					
手工	close-up, leaf, plants	clouds, sky, sun, tree	birds, nest, tree	frost, ice, sky, tree	bear, ice, polar, snow
SSTM	close-up, leaf, plants,stems, lily	clouds, sky, sun, tree,mountain	birds, nest, tree, grass	frost, ice, crystals, tree	bear, polar, snow, rocks
PLSA-GMM	Leaf, plants, texture, pepper, lily	clouds, sun, land, dock, storm	birds, nest, tree, branch, wood	frost, ice, fruit, crops, rapids	bear, polar, cubs, truck, vehicle

本文方法的优势不仅在于通过提取场景语义在视觉特征和高级语义之间建立了桥梁,而且与基准模型 PLSA-GMM 相比,克服了仅用前若干个(例如 5 个)最大后验概率的标注字进行的定长标注。本文方法的特色是建立了场景语义树,来对场景语义进行合理的组织和剪枝,使标注字与视觉特征的关系更加清晰,从而提高标注质量。

5　结　论

提出了一种基于场景语义树的图像标注方法。该方法以场景语义为桥梁,构造了一种语义二叉树,实现了对场景语义的逐层细化。随着路径的深入,语义二叉树节点所对应的高低级语义,即标注字与视觉特征描述之间的联系逐步清晰。目前该方法所构建的语义二叉树结构较为固定,缺乏一定的自适应性,因此将来工作的重点就是要提高语义树的数据自适应性。

参考文献(References)

[1] Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words [C]//Proceedings of the 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management. Orlando, Florida: MISRM, 1999, 1-9.

[2] Duygulu P, Barnard K, De Freitas J F G, et al. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary[C]// Proceedings of the 7th European Conference on Computer Vision-Part IV. Berlin: Springer-Verlag, 2002: 97-112.

[3] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures[C]// Proceedings of the 16th Annual Conference on Neural Information Processing Systems. Istanbul, Turkey: NIPS, 2003, 6: 307-315.

[4] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models[C]// Proceedings of the 26th Intl. ACM SIGIR Conference. Toronto, Canada: ACM, 2003, 119-126.

[5] Jeon J, Manmatha R. Using maximum entropy for automatic image annotation [C]//Proceedings of the 3rd International Con- ference on Image and Video Retrieval. Singapore: CIVR, 2004, 835-840.

[6] Wang M, Zhou X D, Zhang J Q, et al. Image auto-annotation via an extended generative language model[J]. Journal of Software, 2008, 19(9): 2449-2460. [王梅,周向东,张军旗,等. 基于扩展生成语言模型的图像自动标注方法[J]. 软件学报,2008,19(9): 2449-2460.]

[7] Lu H Q, Liu J. Image annotation based on graph learning[J]. Chinese Journal of Computers, 2008, 31(9): 1629-1639. [卢汉清,刘静. 基于图学习的自动图像标注[J]. 计算机学报, 2008,31(9): 1629-1639.]

[8] Monay F, Gatica-Perez D. PLSA-based image auto-annotation: constraining the latent space [C]// Proceedings of the 12th

annual ACM international Conference on Multimedia. New York, USA: ACM, 2004, 348-351.

[9] Sivic J, Russell B C, Efros A A, et al. Discovering objects and their location in images[C]//Proceedings of the Tenth IEEE International Conference on Computer Vision. Beijing, China: IEEE, 2005, 1:370-377.

[10] Gong T, Li S, Tan C L. A semantic similarity language model to improve automatic image annotation[C]//Proceedings of IEEE International Conference of Tools with Artificial Intelligence. Arras, France: ICTAI, 2010, 197-203.

[11] Yu X, Chua T S. Semantic context modeling with maximal margin conditional random fields for automatic image annotation[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, USA: CVPR, 2010, 3368-3375.

[12] Fu J. Research on semantic-based image annotation[DB]. Chinese Sciencepaper Online, Jan, 2011,1.[付杰. 基于语义的图像标注关键技术研究[DB]. 中国科技论文在线, 2011, 1.]

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

[14] Rubner Y, Tomasl C, Guibas L J. The earth mover's distance as a metric for image retrieval[J]. International Journal of Computer Vision, 2000, 40(2): 99-121.

[15] Huang C H, Yin J, Hou F. A text similarity measurement by combining semantic with TF-IDF model[J]. Chinese Journal of Computers, 2011,34(5):856-857. [黄承慧,印鉴,侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似性度量方法[J]. 计算机学报,2011,34(5):856-857.]

[16] Yang F, Liu Y M. Unified similarity measure methed based on visual and semantic information[EB/OL]. (2012-02-01)[2012-03-09]. http://www.paper.edu.cn/index.php/default/releasepaper/content/201202-572.[杨帆,刘咏梅. 基于视觉和语义信息的联合图像相似性度量方法[EB/OL]. (2012-02-01)[2012-03-09]. http://www.paper.edu.cn/index.php/default/releasepaper/content/201202-572.]

[17] Yu L S, Zhang T W. Image clustering based on correlation between visual features and annotations[J]. Acta Electronica Sinica, 2006, 34(7): 1265-1269. [于林森,张田文. 基于视觉与标注相关信息的图像聚类算法[J]. 电子学报,2006, 34(7):1265-1269.]