

# 基于图学习的自动图像标注

卢汉清 刘 静

(中国科学院自动化研究所 北京 100190)

**摘 要** 自动图像标注是图像检索任务中重要而具有挑战性的工作. 文中首先讨论并解释了自动图像标注问题, 通过总结现有的研究工作, 提出了一种基于图学习的图像标注框架. 在该框架下, 图像标注被分为两个阶段来完成, 即基本图像标注与图像标注改善. 其中, 前者是通过以图像间相似性为依据的图学习过程来提供图像的初始标注, 而后者是通过以词汇间语义相关性为依据的图学习过程来改善前者取得的标注结果. 该框架主要涉及到图像与文本词汇两种媒体的内部和相互之间的各种关系的估计问题. 基于此, 作者又给出了针对上述各子问题的改进方法, 并将它们综合起来实现了有效的图像标注. 最后, 通过 Corel 图像集与网络数据集上一系列实验结果, 验证了该模型框架及所提出解决方案的有效性.

**关键词** 图像标注; 图学习; 图像相似性; 词义相关性  
**中图法分类号** TP391

## Image Annotation Based on Graph Learning

LU Han-Qing LIU Jing

(Institute of Automation, Chinese Academy of Sciences, Beijing 100190)

**Abstract** Image annotation is an important and challenging task in image retrieval. This paper discusses the annotation process theoretically by reviewing some related work, and proposes a unified annotation framework via graph learning. The framework includes two sub-processes, i. e., basic image annotation and annotation refinement. In the basic annotation process, the image-based graph learning is utilized to obtain the candidate annotations. In the annotation refinement process, the word-based graph learning is used to refine those candidate annotations from the prior process. This paper also proposes some improvements on sub-problems involved in the framework and expect their combination to enhance the overall performance. Finally, experiments conducted on the Corel dataset and Web image dataset demonstrate the effectiveness of the unified framework and the proposed improvements.

**Keywords** image annotation; graph learning; image similarity; word correlation

## 1 引 言

随着数字影像技术与互联网技术的迅速发展, 用户可以轻松地获取大量网络图像. 为了有效地组织、查询与浏览如此大规模的图像资源, 图像检索技

术应运而生, 并且受到了广泛关注.

现有的图像检索方式主要分为两种: 基于内容的图像检索(Content-Based Image Retrieval, CBIR)和基于文本的图像检索(Text-Based Image Retrieval, TBIR). 通常, CBIR 要求用户提交一幅图像或简图作为查询, 采用图像的视觉特征(如颜色、纹理和形

状等)建立索引,然后根据图像与查询间的视觉相似性度量来实现检索.由于底层视觉特征与高层语义概念之间“语义鸿沟(semantic gap)”的存在,CRIR 的检索性能难以令人满意.而对 TBIR 来说,它要求用户提交文本作为查询,对图像需要事先建立文本索引,使得对图像的检索转化为对文本关键词的相关性匹配与查找.对用户而言,TBIR 在提交查询方面更加便捷,由此成为目前商业化图像搜索引擎的主要方式,如 Google<sup>①</sup>,Yahoo<sup>②</sup>等.但是,这种方式需要对图像建立文本索引,也就是说,实现图像的语义标注,这将是 TBIR 技术中极具挑战的一项工作.早期,人们是通过手工标注的方式来实现,但这项工作耗时费力,尤其面对大规模的网络图像时,显然它已经无法胜任.因此,如何快速、有效地实现对图像的自动语义标注,变得十分有必要.

目前,图像自动标注已经得到了广泛研究,人们提出了各种不同方法,并取得了相当多的成果.这些方法有着各自不同的出发点和解决方案.本文探讨了自动图像标注问题本身的特点及其可用信息,在此基础上,提出了一个统一的图像自动标注的算法框架.该框架不仅可以用来解释和分析许多相关的图像自动标注工作,而且能够指出图像标注技术可能的研究方向.它将原本复杂的图像自动标注问题分解为多个相对独立、易于解决的子问题.在这个框架中,图像标注过程可分为两个阶段来完成,即基本图像标注与图像标注改善,其中前者是通过以图像间相似性关系为依据的图学习过程来提供图像的初始标注,而后者是通过以词汇间语义关联关系为依据的图学习过程来改善前者取得的标注结果.该框架主要涉及到图像与文本词汇两种媒体的内部和相互之间的各种关系的估计问题.

基于此,本文还给出了针对上述各子问题的解决方法.首先,我们提出了一种最近邻生成链的方法来描述图像间的相似关系,它利用链式的统计信息来补充传统的基于数据点两两关系的计算方法,从而有效的融入了对数据结构分布的考虑.其次,我们结合训练数据中的标注信息通过多贝努利模型来估计图学习算法中的初始状态矩阵,取代了传统的(0 or 1)指示性函数.再次,我们给出了联合的词间关系的估计方法,主要考虑了训练集中标注文本的统计相关性和基于结构化词典 WordNet 的语义相关性.最后,将它们融合到整个标注框架中,有效地实现了图像的自动标注,并通过 Corel 图像集上一系列实验结果验证了该解决方案的良好性能.

## 2 框架的提出与分析

### 2.1 图像标注问题的分析

图像标注的基本目标是根据图像的视觉内容和可获得的指导信息来确定对应的文本语义描述.于是,在图像标注任务中会涉及到两种不同的媒体,其一是图像,其二是文本.这两种媒体具有良好的互补性,可以共同协作来传递信息,正所谓“图文并茂”.由这两种媒体可以产生 4 种关系,即图像间关系(Image-to-Image Relation, IIR)、词间关系(Word-to-Word Relation, WWR)以及图像与词之间两种不同的映射关系:由图像到词的关系(Image-to-Word Relation, IWR)和由词到图像的关系(Word-to-Image Relation, WIR),如图 1 所示.下面将对这 4 种关系分别进行介绍.

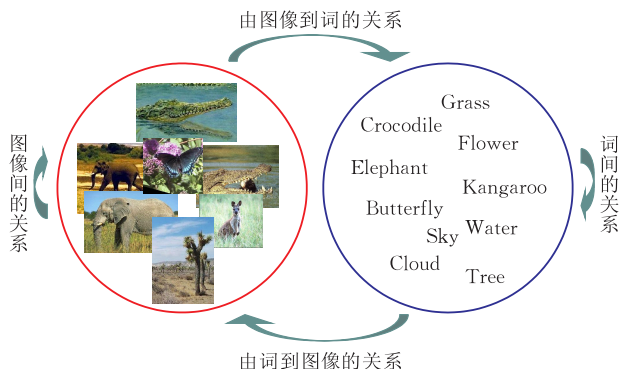


图 1 图像标注的简单图示

#### (1) 图像间关系(IIR)

它通常是指由视觉特性所决定的两幅图像之间的相关性.这里的视觉相关性可取决于许多方面,如颜色、纹理、形状与结构信息等.在图像分析、检索及计算机视觉方面,如何计算图像内容的表达形式以及对应的相似性是一个很重要的研究课题.

#### (2) 词间关系(WWR)

与低层的视觉特征相比,词汇能够反映更加明确、清晰的语义信息,是人们在日常生活中用来表达个人主观意愿的一种有效方式.在图像标注任务中,词间关系被广泛地理解为两词可同时作为一幅图像的语义标注的合适程度,通常体现在语义相关性与统计相关性上.

#### (3) 由图像到词的关系(IWR)

这种关系可以看作由图像产生词汇的可能性.

① <http://images.google.com>

② <http://images.search.yahoo.com>

一般情况下,它可以通过某种假设分布,在某一给定训练集上完成估计。例如,CRM 模型<sup>[1]</sup>中采用的是多项式分布,而在 MBRM 模型<sup>[2]</sup>中采用的是多伯努利分布。在多项式分布中,各个词汇被当作不同的、相互竞争的类别,将标注过程视为一种多分类问题(multi-class classification)。而图像标注问题更适合于采用多标记分类问题(multi-label classification)进行描述,各个类别之间并没有明显的互斥关系,即对应一幅图像的多个标注词汇可以联合存在。合理地考虑各类别之间的统计相关特性可成为提高性能的有效手段。

(4) 由词到图像的关系(WIR)

与 IWR 相反,WIR 的估计可看作:给定词汇后,求取生成图像的后验概率。现有的基于文本的图像搜索正符合由给定词汇来估计图像的过程。因而,它的估计可以结合检索领域的相关技术。有关这方面的研究将作为我们今后的工作内容而不作为本文的重点。

2.2 图学习算法的基本理论

基于图学习的算法是一种半监督算法,也就是说,已知类标的训练数据和未知类标的测试数据都将参与到算法的学习过程中。与传统的有监督学习和无监督学习相比,半监督学习可以在学习阶段利用更多的信息,如数据的分布特性等,它适用于总数据量较大、已标记训练数据量相对较小的情况。其主要思想是利用数据的总体空间分布特征和原始类标信息,使得最终得到的分类结果在数据空间总体上能够充分平滑(即相邻点的类标信息相似),同时保证尽可能地拟合训练数据。

文献[3-4]采用了图的方法来逼近流形,并通过图理论来学习流形空间中各数据点的分类情况。下面,我们将对这一图学习过程进行简单介绍。

设所有的数据可以表示为图  $G=(V,E)$ ,  $V$  表示顶点集合,由数据集中的各数据点  $x_1,x_2,\cdots,x_N$  组成,  $E$  为边的集合  $S=\{S_{ij},i,j=1,2,\cdots,N\}$ , 边的强度反映了各顶点间的相似性。对有类标的数据点  $x_i$ ,其对应的初始状态  $y_i=1$ ;而对未知类标的数据点  $x_j$ ,其对应的初始状态  $y_j=0$ 。于是,按照迭代方式进行的数据分类过程如下:

- (1) 计算邻接矩阵  $W^{N\times N}$ 。
- (2) 正则化  $W$  得到相似性矩阵:

$$S=D^{-1/2}WD^{-1/2} \tag{1}$$

其中,  $D$  为对角阵,  $D_{ii}$  为  $W$  中第  $i$  行各元素之和。

- (3) 迭代计算至收敛:

$$r(t+1)=\alpha S \cdot r(t)+(1-\alpha) y \tag{2}$$

其中,  $t$  为迭代次数,  $\alpha \in [0,1]$ ,  $r(0)=y$  为初始状态向量,即初始类标情况。

- (4) 按照最终状态  $r^*$ ,对各点进行分类。

可以证明,由于相似性矩阵的特征根满足  $-1 \leq e_k \leq 1$ ,该学习过程可以得到收敛解:

$$r^*=(1-\alpha)(1-\alpha S)^{-1} y \tag{3}$$

由式(2)或式(3)看出,在图学习过程中有两个重要组成部分:相似性矩阵  $S$  和初始状态向量  $y$ 。前者描述了数据点的分布情况,而后者提供了学习过程中用到的先验信息。这种图学习的过程可以简单地理解为基于数据点间的相似性关系进行类标传播的过程。

2.3 基于图学习的图像标注框架

基于前面的讨论,若我们将每幅图像(或每个标注关键词)作为图节点,以图像间(或标注词间)的相似关系作为边,通过图学习算法就可以实现标注信息从已标注图像到未知图像的传播,从而完成图像标注任务。

这里,我们将每个标注看作一类,当同时处理  $c$  个标注关键词时,初始状态向量  $y$  就变成了一个初始状态矩阵  $Y^{N\times c}$ ,即当图像  $i$  被标为  $w_j$  时,  $Y_{ij}=1$ , 否则,  $Y_{ij}=0$ 。相应地,描述标注状态的向量  $r$  也变为一个标注状态矩阵  $R^{N\times c}$ ,其中  $R_{ij}$  表示第  $i$  幅图像被标为第  $j$  个标注关键词的可能性。于是,式(2)与式(3)可改写为

$$R(t+1)=\alpha S \cdot R(t)+(1-\alpha) Y \tag{4}$$

$$R^*=(1-\alpha)(1-\alpha S)^{-1} Y \tag{5}$$

由此,我们提出了基于图学习的图像标注框架,如图 2 所示。这个框架由两部分组成:(1) 通过以图像为节点的图学习过程完成基本图像标注;(2) 通过以标注关键词为节点的图学习过程来实现对基本

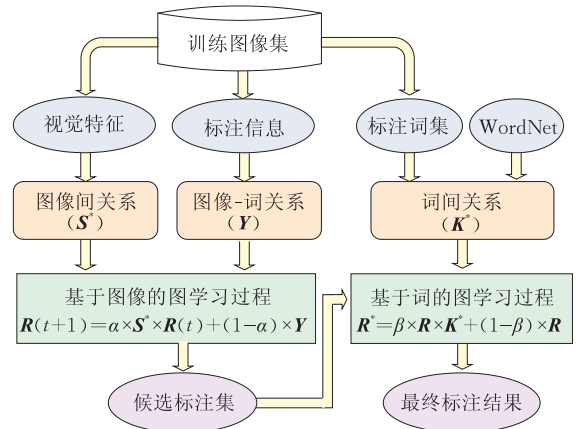


图 2 基于图学习的图像标注框架

图像标注后的标注结果的改善。

在基本图像标注阶段,是利用图像间的视觉相似性建立以图像为节点的图,通过图学习的算法将标注信息按照图像间的视觉相似关系从已标注图像传播到待标注图像。而在图像标注改善阶段,是利用标注关键词之间的相关性建立以词为节点的图,由基本标注阶段的标注结果来设定初始状态向量,通过图学习过程得到每幅图像的最终标注结果,从而较好地保持了标注结果的语义一致性。

## 2.4 该框架对其它相关工作的解释

前面的标注框架完整地考虑了自动图像标注任务中的主要问题,可以用来解释现有的多种相关工作。我们将通过以下几种算法的比较与分析,着重了解它们之间的共同点和差异之处,从而帮助理解其各自带来性能提高的原因。其中涉及到的公式及参数均保持与其出自的参考文献一致,在此不作具体介绍。

### (1) 跨媒体相关模型(CMRM)<sup>[5]</sup>

在该模型中,图像被表示为经过量化后分割区域的组合,每个量化区域被称作“blob”,它利用 blob 之间的关系计算图像相似性。基于此,一幅测试图像  $I$  的视觉信息  $\{b_i, i=1, 2, \dots, m\}$  与标注词汇( $w$ )的联合概率分布可表示为

$$P(w, b_1, b_2, \dots, b_m) = \sum_{J \in T_I} P(J) P(w | J) \prod_{i=1}^m P_v(b_i | J) \quad (6)$$

其中,  $P(w | J)$  与  $P(b | J)$  均假设为多项式分布,可通过最大似然估计得到。从中容易看出:

$$P(I | J) = \prod_{i=1}^m P_v(b_i | J) \quad (7)$$

上式给出了已标注图像  $J$  与待标图像  $I$  之间的相似关系,而  $P(w | J)$  是从训练集中已标注信息学习到的由图像到词的关系。因此,CMRM 的实现是基于图像间视觉相似性的类标信息的传播,即前面所提框架中的基本图像标注这一子过程。

### (2) 连续相关模型(CRM)<sup>[1]</sup>

与 CMRM 不同的是,该模型直接利用了图像区域的特征向量而没有再量化成“blob”,从而避免了量化误差可能带来的负面影响,然后通过 Gaussian 核函数来计算图像间的关系。CRM 仍然假设标注词汇满足多项式分布。于是,对于一幅测试图像  $I$  (被表示为一些区域的集合  $r_A$ ) 以及包含多个词汇的标注集合  $w_B$ ,其联合概率可表示为

$$P(r_A, w_B) =$$

$$\sum_{J \in T} P(J) \prod_{b=1}^{n_B} P_v(w_b | J) \prod_{a=1}^{n_A} \int_{R^K} P_R(r_a | g_a) P(g_a | J) dg_a \quad (8)$$

由此可以得到类似的形式:

$$P(W | J) = \prod_{b=1}^{n_B} P_v(w_b | J) \quad (9)$$

$$P(I | J) = \int_{R^K} P_R(r_a | g_a) P(g_a | J) dg_a \quad (10)$$

CRM 模型同 CMRM 相比只是采用了不同的图像间相似关系的估计方法,整体来看它完成的仍然是基本图像标注。

### (3) 多伯努利相关模型(MBRM)<sup>[2]</sup>

MBRM 同样采用了联合的区域特征来描述各幅图像,但它假设标注词汇服从多伯努利分布,其对应的联合概率形式为

$$\sum_{J \in T} P_T(J) \prod_{a=1}^{n_A} P(r_A | J) \prod_{v \in w_B} P_v(v | J) \prod_{v \notin w_B} (1 - P_v(v | J)) \quad (11)$$

由此可得

$$P(w | J) = \prod_{v \in w_B} P_v(v | J) \prod_{v \notin w_B} (1 - P_v(v | J)) \quad (12)$$

$$P(I | J) = \prod_{a=1}^{n_A} P(r_A | J) \quad (13)$$

该标注过程同 CMRM 非常类似,只是改进了由图像到词的关系估计。

### (4) 互相关传播模型(CLP)<sup>[6]</sup>

CLP 提出利用词汇之间的相关性来提高图像标注的性能,同时,在计算过程中考虑了词汇出现频率的平衡问题,以防高频词汇对低频词汇所对应的信息造成淹没,而这些低频词汇往往正是所标记物体的特性化描述。虽然文献[6]中的推导比较复杂,但是我们可以根据其将关键词  $w_k$  标注到图像  $I$  的推导结果得到

$$\begin{aligned} P(w_k | I) &= f(T_k) - f(T_{k-1}) \\ &= \sum_{J \in T} K(I, J) \Omega(t^T(T_k) t(J)) - \\ &\quad \sum_{J \in T} -K(I, J) \Omega(t^T(T_{k-1}) t(J)) \\ &= \sum_{J \in T} K(I, J) \{ \Omega(t^T(T_k) t(J)) - \\ &\quad \Omega(t^T(T_{k-1}) t(J)) \} \end{aligned} \quad (14)$$

其中,  $K(I, J)$  为图像  $I$  与图像  $J$  之间的相似性函数,  $T_k = 1, 2, \dots, k$ ,  $\Omega(\cdot)$  为一凹函数。显然,当满足

$$P(I, J) = K(I, J) \quad (15)$$

$$P(\omega_k | J) = \Omega(t^T(T_k)t(J)) - \Omega(t^T(T_{k-1})t(J))$$

(16)

时,该模型也可表示成一个基本图像标注的实现过程,但它在估计由图像到词关系时考虑到了词间的相互关系,这也是它相对前面几种方法的改进之处.

(5) 基于 WordNet 的方法(WNM)<sup>[7]</sup>

该模型在衡量词汇之间的关系时利用了来自 WordNet 的结构化语义信息. 它首先利用基于 TM 模型的算法得到一幅图像的候选标注词汇,然后综合使用多种基于 WordNet 的语义度量来计算每一词汇与其它所有词汇之间的语义相关程度,最后通过标注结果的语义一致性来确定标注结果,可表示为

$$T_R = S_{IWR} \cdot S_{WWR}$$

(17)

其中  $S_{WWR}$  是基于 WordNet 得到的语义相关矩阵,而  $S_{IWR}$  由 TM 算法的标注结果给定,当词  $\omega_j$  被确定为图像  $I_i$  的标注结果时,  $S_{IWR}(i, j) = 1$ , 否则为 0.

显然,通过上式(17)所实现的正是基于词间关系的标注改善过程. 它只采用了基于 WordNet 的语义相关度量,并且由于  $S_{IWR}$  的二进制设定方法,使得该法只能完成不相关标注的去除,而无法实现相关标注的补充.

由此看出,本文给出的图像标注框架可以用来解释多种相关的图像标注算法,但这些已有算法往

往专注于该框架下的某一部分进行改进. 实际上,为了得到更好的标注性能,针对该框架下的每一部分,包括三种关系(图像间、词汇间、图像/词汇间)的估计和图学习的算法等展开研究和改进,对自动图像标注的性能改善都是非常有意义的. 本文则着重在图像间关系及词间关系估计问题上提出了改进算法.

3 基本图像标注

3.1 以图像为节点的相似图的建立

通常,我们采用两两图像之间的视觉相似性(pairwise similarity)来建立以图像为节点的相似图. 但这种方式由于没有考虑到数据集或某个数据子集内的结构分布信息,往往效果不能令人满意. 例如,在图 3 中给出的两种情况,其中第 1 列是数据的原始分布,第 2 列是基于传统相似度求解方法得到的聚类结果,第 3 列是基于本节所提出的方法来建立相似图而得到的聚类结果. 其中,所采用的聚类方法是文献[8]中的谱聚类方法. 由此可以看出,数据点的类别信息与其周围各点的分布情况是十分相关的. 为了将这种结构分布信息引入到相似图的构造中,我们试图通过链式分布的统计信息来改善传统方法的不足,并将这种方法称为基于最近邻生成链(Nearest Spanning Chain, NSC)的构图方法.

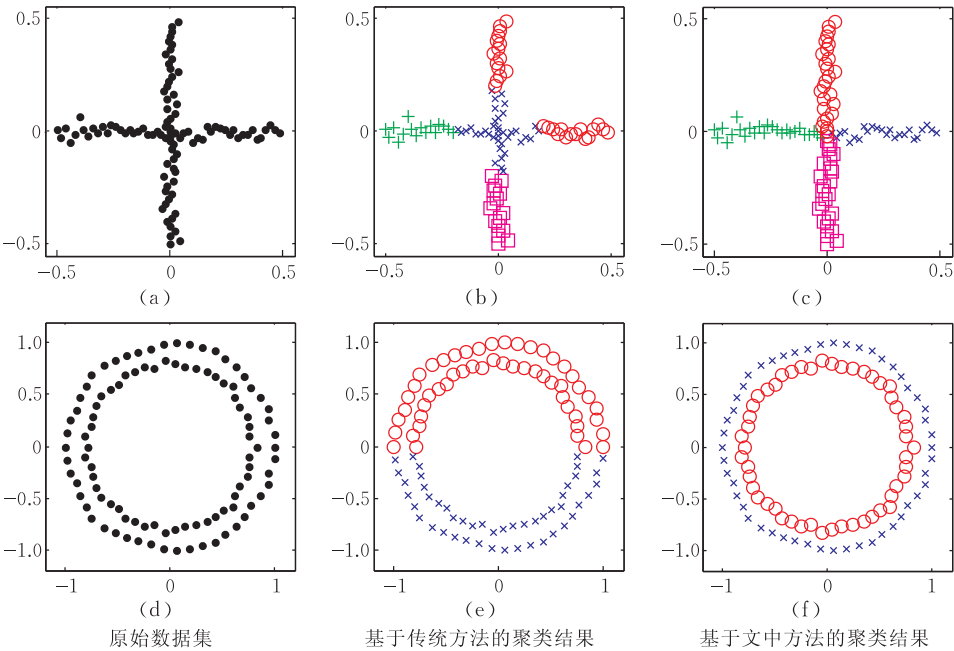


图 3 简单示例

所谓“最近邻生成链(NSC)”,它与图论中“最小生成树(Minimum Spanning Tree, MST)”的概念<sup>[9]</sup>

相类似,都要求数据集中的各点按照某种准则连接起来,构成一种图结构. 不同之处在于, NSC 是各点



按照一定次序顺次连接在一起的一条链,其中只包含一对一的连接,而不会出现一对多的分支式连接;而 MST 则允许存在分支式连接.可以说,NSC 是 MST 的极端简化形式.对由  $n$  个点组成的 NSC 来说,应具有如下性质.

**性质 1.** NSC 中各边是按照“当前最近邻”的准则来确定的,即某点总是在前面没有被连接过的点集中找到它的最近邻,并与之形成连接;

**性质 2.** 这  $n$  个点由  $(n-1)$  条边按照“当前最近邻”准则顺次连接;

**性质 3.** 首尾两个点只有一条边与之连接,其它  $(n-2)$  个点有两条边与之连接.

直观地讲,两个距离较近(相似)的点比两个距离较远(不相似)的点应该更有可能被连接在一起.于是,当我们选取多个不同的数据点作为起点生成一个 NSC 集合时,两个相似的数据点应该更多地被连接起来,而本来就不相似的点对被连接的可能性很小.因此,我们可以利用在该 NSC 集合中各点对间连接的统计信息来反应数据的结构分布.为了清楚起见,我们给出了一个非常简单的例子来说明这一点,如图 4 所示.其中,左图给出了沿单位圆分布的数据集,右图列出了由该数据集生成的所有可能的 NSC.以 1 点为例容易看出,1 点离 2 点要比离 9 点更近,它体现在给定的 NSC 集合中就是:(2-1)与(1-2)连接出现的次数大于(9-1)与(1-9).

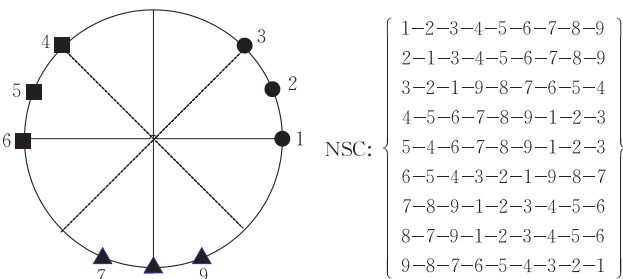


图 4 NSC 的简单示例

下面,我们将具体讨论如何表示存在于多个 NSC 中的统计信息,并将这种信息融合到点对间的相似性表示中.简单地说,在给定的 NSC 集合中,某一点对出现的频率可以作为该点对的统计相关性的描述,即出现次数越多,该点对的相关性越强,反之,相关度就弱.考虑到计算复杂度和生成 NSC 所依据的“当前最近邻”准则,我们从图像本身的特性出发,提出了一种加权式局部统计相关性度量来构造图像间的相似关系,主要包括以下几方面内容:

首先,图像的特征空间是一个局部紧密的稀疏

空间,因此基于图像数据集建立 NSC 时,没必要使得 NSC 中包含所有的数据点,而只需将分布得相对紧密的那些数据点连接组成一个 NSC.然后,分别选取数据集中的各点为起点来构造不同的 NSC,最终得到我们需要的 NSC 集合.

另一方面,由于“当前最近邻”准则的限制,出现在 NSC 中的各点对间连接的可信度是不一样的,即出现在前面的点对连接要比出现在后面的点对连接的可信度要高,因此在相关性表示时,需要根据每个点对连接出现的先后顺序进行适当的加权处理.

再者,对于出现在某个 NSC 中的各点,除了被直接连接的点对可认为是相关的,在一定邻域内被间接连接的点对也是具有相关性的.下面举例说明.如图 4 所示,点 1 与点 3 在数据分布上是相离较近的点,但在对应的 NSC 集合中这两点却一直没有被直接连接,而总是通过点 2 构成两点的间接连接,多次以(1-2-3)或(3-2-1)的链式连接出现.于是,为了全面的描述数据相关性,我们将间接连接的点对也考虑进来,但有一点需要注意,间接连接的中间链越长(点 2 可看作点 1 与点 3 间长度为 1 的中间链),该间接连接点对的可信度就越低.

基于上述三方面的内容,在给定含有  $N$  个 NSC 的集合后,任一点对  $(i, j)$  的统计相关性可表示为

$$C_{ij} = \sum_{n=1}^N seq\_w_{ij}^n \cdot f_{L_{ij}^n} \cdot \delta_{ij}^n \quad (18)$$

$$seq\_w_{ij}^n = \exp\left(\frac{\lambda_1}{idx_i + idx_j}\right) \quad (19)$$

$$f_{L_{ij}^n} = \exp\left(\frac{\lambda_2}{L_{ij}^n + 1}\right) \quad (20)$$

其中,  $seq\_w_{ij}^n$  表示在第  $n$  个 NSC 中,与该点对的出现位置有关加权系数;  $f_{L_{ij}^n}$  表示在第  $n$  个 NSC 中,与该点对的中间连接链长呈负相关的加权系数;当点对的中间链长小于等于给定的最大链长  $L_{\max}$  时,  $\delta_{ij}^n = 1$ , 否则,  $\delta_{ij}^n = 0$ .

我们将式(18)给出的加权式局部统计相关度量 ( $C_{ij}$ ) 结合传统的特征向量间的相似性 ( $W_{ij}$ ), 最终得到数据点对间相似性的度量:

$$W_{ij}^* = C_{ij} \cdot W_{ij} \quad (21)$$

下面,我们给出基于 NSC 的构图过程(假设已给定  $N$  幅图像):

(1) 分别以每幅图像对应的数据点为起点,按照“当前最近邻”准则,构造长度为  $M$  ( $M < N$ ) 的 NSC,由此得到一个 NSC 的集合,其容量为  $N$ ;

(2) 在 NSC 集合中,由式(18)计算各点对的加

权式局部统计相关性,从而得到表示整个图像集的统计相关矩阵  $\mathbf{C}^{N \times N}$ ;

(3) 计算基于图像特征向量的相似矩阵  $\mathbf{W}^{N \times N}$ ;

(4) 按照式(21),得到最终的相似矩阵  $\mathbf{W}^*$ ;

(5) 按照式(1),得到正则化的相似矩阵  $\mathbf{S}^*$ .

### 3.2 基本图像标注的实现

前一小节着重介绍了基于 NSC 的相似图的构造方法,而在图学习过程中还有一个重要部分就是初始状态矩阵的确定. 传统的图学习方法一般采用由(0 or 1)表示的指示函数来描述每个点的初始状态,即若图像  $i$  被标注为词  $w_j$ ,则对应的状态分量  $Y_{ij}=1$ ,否则  $Y_{ij}=0$ . 实际上,由于图像表达的语义十分丰富,每幅图像被某个词标注并不是绝对事件,而通常需要以某种概率来度量该事件发生的可能性. 另外,考虑到图像标注应该是一种多标记(Multi-label)而非多分类(Multi-class)问题,我们采用了多贝努利模型将标注信息表示为概率形式,以此来确定图学习的初始状态向量,表示如下:

$$Y_{ij} = P(w_j | I_i) = \frac{\mu \delta_{w_j, I_i} + N_{w_j}}{\mu + N_T} \quad (22)$$

其中,  $Y_{ij}$  表示给定图像  $I_i$  生成标注词汇  $w_j$  的条件概率;  $\mu$  是一个平滑参数;  $N_{w_j}$  表示标注图像集合中被标为  $w_j$  图像的数目;  $N_T$  表示已标注图像的总数; 当  $w_j$  是图像  $I_i$  的真实标注时,  $\delta_{w_j, I_i} = 1$ , 否则,  $\delta_{w_j, I_i} = 0$ .

接下来,我们结合已经得到的相似图  $\mathbf{S}^*$  和初始状态矩阵  $\mathbf{Y}$ ,按照式(4)进行迭代计算,最终得到稳定的状态矩阵  $\mathbf{R}^{N \times c}$ . 这一稳定解即作为第一阶段的基本图像标注的结果,它将为后面的图像标注改善提供参考.

## 4 图像标注改善

在前面以图像为节点的图学习过程中,主要利用了图像间的视觉相似性,但由于语义鸿沟的存在,对每幅图像的标注结果很难保证在语义上的一致性,为此我们引入了第二个以词为节点的图学习过程,用来实现图像标注的改善.

### 4.1 以词为节点的相似图的建立

图像标注的目标是要得到反映图像内容的一组词汇,而词汇之间存在着多种多样的语义关系. 最常见的是层次关系,如“汽车”是“交通工具”的一种. 除此,词汇之间还存在多种相关性,如“汽车”与“道路”之间有着很强的联系,这种相关性不依赖于特定数

据集,它是人们在生活中大量知识的积累和反映. 当一幅图像已被标为“汽车”、“人”等词汇后,“道路”作为该图像标注词汇的概率就会相应提升. 为了获取这种相关信息,一种方法是从训练数据集中利用已标注词汇间的共生概率来计算词汇间的关系. 该方法基于已标注信息,相对准确,但它不能反映更广义的人的知识. 于是,我们可以利用具有大量词汇的、包含了人的知识的结构化电子词典(如 WordNet)来计算词汇间的关系. 与统计方法相比,词典包括了更加完整的语义信息. 下面,我们将分别介绍这两种词间关系的估计方法.

#### (1) 基于共生关系的统计互相关度量

一般来讲,在训练集中有较高共生频率(同时出现在同一文档中)的词汇应该具有较强的语义相关性. 这是由于共生频率较高的词往往是代表了两个互相紧密关联的概念或物体,其中之一的出现也意味着另一概念或物体的出现,从而具有较大的可能性被同时用来标注一幅图像. 现实生活中有很多这样的例子,如“天空”与“云”,“大海”与“沙滩”等. 所以,词汇的共生关系可以有效地提供词汇间语义关系的信息. 共生词汇可被定义为同时出现在同一幅图像的标注中的两个词汇. 但是,简单地对共生关系进行频数统计,由于无法有效地考虑到各词汇的不同特性,因此度量效果不能令人满意. 这里,我们借用文本领域中的 TF-IDF<sup>[10]</sup>的思想对词汇的共生关系进行衡量:

$$K_{WC}(v_1, v_2) = K_c(v_1, v_2) \times \log\left(\frac{N_T}{n_1}\right) \quad (23)$$

其中,  $v_1, v_2$  为词汇,  $K_c(\cdot)$  为二者共生出现的次数,  $n_1$  为包含  $v_1$  作为标注的图像数目,  $N_T$  为训练集中的图像总数目. 需要说明的是,根据以上定义,  $K_{WC}(v_1, v_2)$  不一定等于  $K_{WC}(v_2, v_1)$ , 即不具有对称性. 我们认为这样的关系对于词汇之间来讲是合理的. 考虑  $v_1, v_2$  为两个出现频率差异较大的词汇. 其中  $v_1$  反映的是一种比较特定的概念或物体,出现次数较少,而  $v_2$  反映了比较宽泛的概念或物体,出现次数较多. 假设两者之间存在很强的共生关系,我们比较容易从  $v_1$  得到  $v_2$ , 但  $v_2$  可能由于其广泛性而与大量的词汇都有着比较强的联系,从而难以推断出  $v_1$  是否应当存在. 例如,对词汇“草原”与“狮子”,前者较宽泛而后者则较为特定. 我们可以比较容易地从“狮子”推断出“草原”的存在,但给定“草原”却很难推断出“狮子”的存在与否,因为“草原”可能与许多种动物紧密关联.

## (2) 基于 WordNet 的语义相关度量

WordNet 是在自然语言处理中得到广泛应用的一种结构化电子词典,它是 Princeton 大学的研究成果<sup>[11]</sup>. 在这个词典中,语义按照“is-a”的层次关系构造了多棵语义树,即语义树上的子节点是父节点的一个细化. 不同树上的节点之间也建有多种语义关系连接,如“is-made-of(由…组成)”、“is-an-attribute-of(是…的某种属性)”等. 所有的这些语义关系都可以用来衡量词汇之间的语义相关性. 在当前的 WordNet 2.1 版本中,总共收录了超过 11 万词汇. 每个词汇可能对应多种语义,所以一个词可能会出现在多棵语义树上的不同地方. 在 WordNet 的基础上,自然语言处理领域的研究提出了许多种定量计算词汇间相关性的度量,其中 Jiang 等人提出的 JNC 度量<sup>[12]</sup>被认为是最有效的一种度量方法之一<sup>[13]</sup>. 于是,我们就将该度量引入我们的解决方案.

在 JNC 中,需要给定一个已按照语义进行过良好标注的词库,由这个词库可以估计出每个概念  $c$  出现的概率:

$$Pr(c) = \frac{Freq(c)}{N} \quad (24)$$

式中  $Freq(c)$  指概念  $c$  出现的次数,而  $N$  是所有概念出现的总次数. 利用此概率可定义该概念对应的信息容量 (Information Content, IC):

$$IC(c) = -\log(Pr(c)) \quad (25)$$

而基于 WordNet 的语义相关性度量可理解为两个概念分别去除公共概念之后信息容量和的倒数,即

$$K_{WNC}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \cdot IC(lcs(c_1, c_2))} \quad (26)$$

式中的  $lcs(c_1, c_2)$  是概念  $c_1$  与  $c_2$  的最低公共父节点,可由 WordNet 的概念树结构来得到.

到此为止,我们已经介绍了两种词间关系的度量方法,各自从不同的角度进行衡量. 为了将这两种方法综合考虑,我们将它们的归一化结果进行线性融合,以此作为词间关系的联合表示:

$$S_{WWR} = \epsilon K_{WC} + (1 - \epsilon) K_{WNC}, \quad 0 \leq \epsilon \leq 1 \quad (27)$$

## 4.2 图像标注改善的实现

有了前面词间关系的估计,我们就能建立以标注词为节点的相似图,并将前面基本图像标注阶段得到的标注结果  $R$  作为本阶段图学习过程的初始状态矩阵,然后通过式(4)进行迭代计算,得到本阶段的标注结果矩阵  $R^*$ . 实际上,  $R^*$  中某个元素  $R_{ij}$  表示的是待标图像  $I_i$  可被标为词  $w_j$  的可能性. 于

是,我们通过对标注词集内的各关键词进行排序,然后取排在前面的几个关键词,就得到了最终的图像标注结果.

## 5 实验结果与分析

### 5.1 实验设置

为了验证本文方法的有效性,并同其它方法进行公平比较,我们采用了自动图像标注工作中普遍使用的 Corel 图像集<sup>[14]</sup>作为实验集. 该数据集共包含 5000 幅图像,其中 4500 幅作为训练图像,500 幅作为测试图像. 通过利用 GraphCut 方法<sup>[15]</sup>将每幅图像分割为 1~10 个区域,每一个区域被表示成一个 36 维的特征向量,其中包括颜色、形状、位置等特征,该特征表示同文献<sup>[14]</sup>中所用. 每幅图像对应 1~5 个真实标注,整个标注词集含有 371 个关键词,而出现在测试集中的关键词为 260 个.

图 5 给出在含有 4500 幅图像的训练集中各标注关键词出现的次数统计,从中可以看出:它们近似呈现 Zipf 分布的特性<sup>[16]</sup>. 其中,出现次数大于 20 次的词有 141 个,出现次数大于 10 次的也只有 209 个,也就是说有三分之一以上的词只给定了相当少的训练图像,因此,基于这样一个训练集来完成图像标注是个极具挑战的任务.

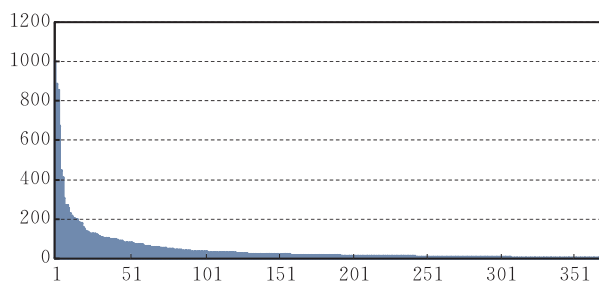


图 5 Corel 训练集中标注关键词出现的次数统计

为了评价图像标注方法的性能,我们采用了最常见的几个指标来衡量. 首先是查准率与查全率,它们的求取是以某单一关键词  $w$  作为查询,在标注好的测试图像集上进行检索,假设标注正确的图像数为  $N_c$ ,可检索到的所有图像数为  $N_s$ ,测试集中与该词相关的所有图像数为  $N_r$ ,于是可得

$$precision(w) = \frac{N_c}{N_s}, \quad recall(w) = \frac{N_c}{N_r} \quad (28)$$

出现在测试集中的每个关键词都将重复这个计算(出现在 Corel 测试集上的关键词共 260 个),最终把得到的查准率与查全率对所有关键词取平均,以此作为评价指标. 除此之外,我们还统计了被正确标



注的词汇的数量,即至少被正确标注一次的关键词的数量,这一数值反映了标注算法对词汇的覆盖程度,记为“NumWords”。在下面的性能比较图示中,我们采用了统一的图示方法,其中查准率与查全率的取值参考左坐标轴,而指标“NumWords”的取值参考右坐标轴。

## 5.2 实验 1: 参数设置对该算法的影响

下面,我们将通过实验来考察以图像为节点的构图过程中所涉及到的两个参数对性能的影响,它们分别是 NSC 的长度  $M$  和点对间连接的中间链最大长度  $L_{\max}$ 。为了清楚地了解这两个参数所起的作用,本组实验只实现了第一阶段的以基本图像标注为目的的图学习过程,而未加入基于词间关系的图像标注改善。

通过前面的讨论,我们需要首先构建  $N$  个长度为  $M$  的 NSC,然后从得到的 NSC 集合中发现各点对的统计相关信息。直观地说,如果 NSC 太长,将会增加时间和空间上的复杂度,同时也容易引入噪声的影响;相反,如果 NSC 的长度太短,那么可能出现的连接对数目将会变少,也就很难从统计意义上反映图像间的相关性。从图 6 中,我们可以得到类似的结论。随着  $M$  取值的增加,算法的性能是逐渐提高的,但当  $M > 250$  时,性能趋于平稳。之所以会趋于平稳,主要原因是在计算基于 NSC 的统计相关性时,我们引入了随点对连接发生位置而变化的不同权重,从一定程度上抑制了可能产生的噪声影响。在后面的实验中,我们设  $M = 300$ 。

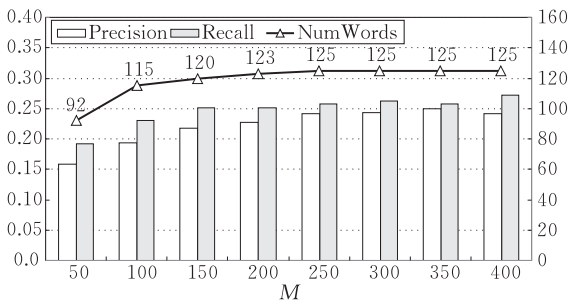


图 6 NSC 长度的变化对方法性能的影响

为了更好地描述各点对之间的相关关系,我们在构图过程中既考虑了点对间的直接连接,也考虑了它们之间的间接连接,从而引入了中间链的概念。我们是通过中间链的最大长度( $L_{\max}$ )这个参数来限制可允许的间接连接的选取范围。当  $L_{\max} = 0$  时,即只考虑直接连接的情况;而当  $L_{\max}$  的取值变大,表明可获取间接连接的点对就会增多。显然,两个极端情况都不会产生良好的标注性能,这一点从图 7 中也可以容易地看出。具体地说,随着  $L_{\max}$  的增加该算法

的性能在初始阶段是逐步上升的;当  $L_{\max}$  增加到 3,性能趋于平稳;而当  $L_{\max} > 6$  以后,性能开始变差。这是因为,太大的  $L_{\max}$  会引入太多的间接连接点对,使得带来的噪声随之增多;而当  $L_{\max}$  太小,可能会丢失一些本来比较相关的点对,以致于不能很好地反映数据分布信息。另外,由于每个连接点对的权值是与链长负相关的,于是求得的相关程度会在噪声和有用信息量之间寻求一种折衷,从而表现出在某段区域内相对平稳的性能。综合考虑,在后面的实验中设  $L_{\max} = 4$ 。

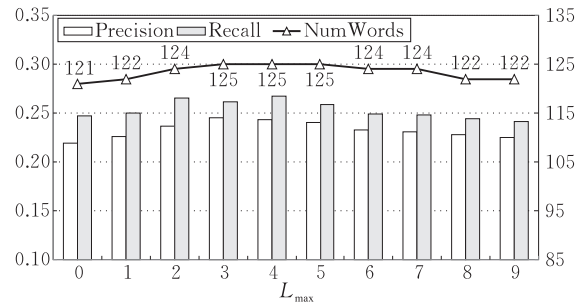


图 7 最大中间链长度的变化对方法性能的影响

除此之外,我们采用了交叉验证的方法对整个框架所涉及到的参数进行设置,具体为  $\lambda_1 = \lambda_2 = 1.00$ ,  $\epsilon = 0.70$ ,  $\alpha = 0.25$ 。

## 5.3 实验 2: 相关方法的比较与分析

在本组实验中,我们将通过与其它几种方法的比较,验证本文所提出的基于图学习的图像标注框架及其解决方案的有效性。这里,涉及到的方法包括:CMRM<sup>[5]</sup>、CRM<sup>[1]</sup>、MBRM<sup>[2]</sup>、GLM(采用传统建图方法的基本图像标注)<sup>[17]</sup>、CLP<sup>[6]</sup>、GLM+NSC(基于 NSC 的图学习方法的基本图像标注)、GLM+WWR(采用传统建图方法的基本图像标注联合词间关系的标注改善)、MGLM(本文提出的基于图学习的图像标注)。以上各种方法的性能比较如图 8 所示,由此可以得出:

(1) 基于标注信息描述由图到词的关系时,更适合按照多标记分类(multi-label classification)问题来解决,而不是多类别分类问题(multi-class classification)来假设词的先验分布。具体而言,CRM 假设标注词汇服从多项式分布,由此将图像标注归为多类别分类问题。MBRM 与 CLP 模型都采用了多标记学习的方法来实现图像标注,其中 MBRM 采用了相对固定的多伯努利分布,而 CLP 将词汇间的统计相关关系融入 to 标注过程中,通过更灵活的方式体现了这一点。由此带来的优势可以在三者的实验结果对比中体现出来,尤其是性能指标“NumWords”的显著提高,从 CRM 的 107 到 MBRM 的 121 再到

CLP 的 125.

(2) 基于图学习(Graph Learning Model, GLM)的各种方法较其它方法表现出更好的标注性能,其中 MGLM 取得了最好的效果. 具体地说, MGLM 较 MBRM 在查准率和查全率指标上分别提高了 31.8% 和 25.9%.

(3) GLM+WWR 通过引入词间关系实现图像标注性能的进一步提高,这种改善主要表现在查全率和“NumWords”两个性能指标上,由此可以看出基于词间关系的“标注性能改善”这一子过程,通过标注关键词的改善与补充有效地克服了标注词汇的多义性与同义性问题,对整个图像标注任务具有重要意义.

(4) GLM+NSC 作为传统构图方式的一种改进,使标注性能取得了显著的提高,尤其在查准率指标上的改善更为明显. 这证明了基于 NSC 的建图方法可以更好地描述图像之间的相似关系,为图学习过程中已标注信息的相似性传播提供了良好的指导,从而保证了较为准确的标注结果. 类似地,CRM 优于 CMRM 也证明了较好地估计图像间关系可以带来标注性能的提高,其中前者使用区域的特征值来计算图像相似性,而后者则通过区域聚类后得到的 blob 表示进行计算,由于聚类过程当中丢失了很多有用信息,所以 CRM 能够更好地估计图像之间的相似性.

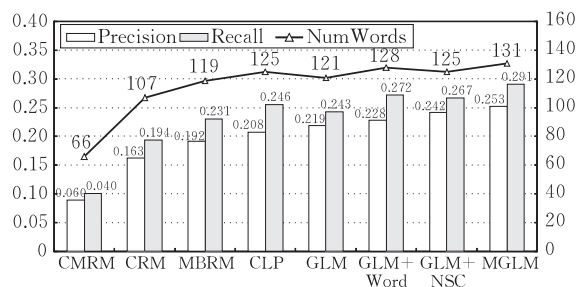


图 8 Corel 图像集上的性能比较

综上,我们分别针对在基于图学习的标注框架中所涉及的几个子问题进行了实验比较与分析,有力地验证了该框架在图像标注任务中所起到的重要作用,并为今后图像标注领域的研究工作提出了有价值的指导与启发.

## 6 结 论

本文提出了一种统一的基于图学习的图像标注框架,该框架能够对许多图像标注的研究工作进行

合理而有效的解释,并对图像标注问题给出了较为完整的概括,因此对今后的研究具有很大的指导意义. 在这个框架下,图像标注过程被分为两个阶段来完成,即基本图像标注与图像标注改善,其中涉及到图像与文本词汇两种媒体的内部和相互之间的各种关系的估计问题. 为此我们提出了针对框架中各子问题的解决方法,并将它们综合起来较好地完成了图像标注任务. 最后,通过 Corel 图像集上一系列实验结果验证了该模型框架及所提出解决方案的有效性.

## 参 考 文 献

- [1] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures//Proceedings of Advance in Neural Information Processing, 2003
- [2] Feng S L, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2004, 2: 1002-1009
- [3] Zhou D, Bousquet O, Lal T N, Weston J, Scholkopf B. Ranking on data manifolds//Proceedings of the 18th Annual Conference on Neural Information Processing System. 2003; 169-176
- [4] Zhou D, Bousquet O, Lal T N, Weston J, Scholkopf B. Learning with local and global consistency//Proceedings of the 18th Annual Conference on Neural Information Processing System. 2003; 237-244
- [5] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models//Proceedings of the 26th Annual International ACM SIGIR. 2003; 119-126
- [6] Kang F, Jin R, Sukthankar R. Correlated label propagation with application to multi-label learning//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006; 1719-1726
- [7] Jin Y, Khan L, Wang L. Image annotations by combining multiple evidence WordNet//Proceedings of the 13th Annual ACM International Conference on Multimedia. 2005; 706-715
- [8] Meila M, Shi J. A random walks view of spectral segmentation//Proceedings of the 8th International Workshop on Artificial Intelligence and Statistic. 2001
- [9] Thomas H C, Leiserson Charles E, Rivest Ronald L, Stein Clifford. Introduction to Algorithms, Chapter 23: Minimum Spanning Trees. MIT Press and McGraw-Hill, 2001; 561-579
- [10] Salton G, Buckley C. Term weighting approaches in automatic text retrieval. Information Processing and Management, 1988, 24(5): 513-523

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

[12] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy//Proceedings of the International Conference on Research in Computational Linguistics. 1997

[13] Pucher M. Performance evaluation of WordNet-based semantic relatedness measures for word prediction in conversational speech//Proceedings of the 6th International Workshop on Computational Semantics. 2005.

[14] Pinar Duygulu, Kobus Barnard, Nando de Freitas, David Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary//Proceedings of the 7th European Conference on Computer Vision. 2002

[15] Shi J B, Malik J. Normalized cuts and image segmentation//Proceedings of the IEEE Conference Computer Vision and Pattern Recognition. 1997

[16] Li Wen-Tian. Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on Information Theory, 1992, 38(6): 1842-1845

[17] Tong H, He J, Li M J, Zhang C, Ma W Y. Graph based multi-modality learning//Proceedings of the 13th Annual ACM International Conference on Multimedia, 2005: 862-871



**LU Han-Qing**, born in 1961, Ph.D., professor, Ph. D. supervisor. His research interests include analysis and understanding on multimedia information, medical image processing, pattern recognition, and computer vision.

**LIU Jing**, born in 1979, Ph. D. candidate, assistant researcher. Her research interests include multimedia analysis and retrieval, and the theoretical study on pattern recognition and machine learning.

Background

Image annotation has been an active research topic in recent years due to its potential impact on both image understanding and Web image search. All kinds of annotation approaches are proposed. As they seem to be different from each other, it is not easy to answer such questions as which models are better, what the connections among them are, and how they should be utilized. In this paper, the authors conduct a formal study on these issues and find that previous

research work can be explained in a unified framework. The framework offers some potential guidance on the study of image annotation, and some improvements under the framework can achieve positive effect.

The research was supported by the National Natural Science Foundation of China (60723005) and the National High Technology Research and Development Program (863 Program) of China Project (2006AA01Z315).