# 对象定位处理中分类信息融合技术研究*

钱 怡, 林 莹, 武港山
(南京大学 计算机软件新技术国家重点实验室, 南京 210093)

**摘 要**: 为提高图像中对象定位技术的处理效果,对对象定位技术和分类技术的融合进行了研究。针对大规模、多对象类别的图像对象定位问题,提出了先进行快速分类,再精确定位的处理方案。通过 MIMLSVM + 多类别分类算法预判出包含对象的图像,利用 ESS 方法在上述图像中定位对象;针对高精度对象定位需求,提出了融入全局分类信息的最优框打分机制,将 MIMLSVM + 算法对于图像的分类信息融入 ESS 方法中最优框的打分信息中。在 PASCAL 2006 数据集上相应的实验结果表明,前者在缩短处理时间的同时取得了不错的定位平均精度,而后者对最优框得分的改进也在多个类别上带来了定位效果的提高。实验结果表明,分类信息融入对象定位处理中能提升处理效果。

**关键词**: 信息融合; 对象定位; 多类别分类; 多示例多标记学习框架; 快速子窗口搜索方法; 最优框
**中图分类号**: TP391 **文献标志码**: A **文章编号**: 1001-3695(2013)12-3844-06
doi:10.3969/j.issn.1001-3695.2013.12.086

## Research of fusing image classification into object localization

QIAN Yi, LIN Ying, WU Gang-shan
(*State Key Laboratory for Novel Software Technology*, *Nanjing University*, *Nanjing* 210093, *China*)

**Abstract**: In order to improve the performance of the object localization technology, this paper researched the approach of fusing image classification into object localization. According to the object localization task in the large-scaled image set containing multi-class objects, this paper proposed an effective scheme that a fast image classification was carried out before a precise object localization. The MIMLSVM + algorithm predicted which images contained the objects, then the ESS method localized the objects just in those images. And aiming at a higher precision in the object localization task, the paper proposed a new scoring mechanism of the optimal box which included the global classified information in the multi-label classification task. The corresponding experimental results on the PASCAL 2006 data set show that the former method shortens the processing time while obtaing a good average positioning accuracy; the latter method brings an improvement in the localization performance on many categories as it make the scoring mechanism of the optimal box better. The experimental results prove that fusing image classification into object localization can really improve the performance of object localization.

**Key words**: information fusing; object localization; multi-label classification; multi-instance and multi-label learning framework; ESS(efficient subwindow search); optimal box

## 0 引言

在信息时代,随着多种数码产品的普及,人们的生活处处涉及图像、视频等多媒体资源,如何让计算机辅助人们从中提取有用的信息具有重大的意义。现实生活中,物体的位置信息有巨大的应用价值,如自动识别车辆行人来辅助驾驶、数码相机自动人脸识别等。物体定位相关的研究也越来越受到重视,然而相关的方法往往拘泥于定位技术本身,没有考虑实际应用中图像资源庞大、对象类别丰富的现实问题。针对现实中的图像资源,需要高效有力的方法进行对象定位处理,从而为后续更进一步的语义认知和资源运用打下基础。

分类和定位一直是媒体内容分析领域的热点,分类能够给出图像的类别信息,定位则给出图像中物体的位置信息。不难发现,对于包含了不完整物体的图像,分类算法往往仍能正确判别,而定位算法会给出比较低的定位得分;对于那些物体很小不明显的图像,分类算法往往会给出比较低的类别得分,而定位算法仍能有效定位,可见分类和定位是对图像信息的不同解读。Harzallah 等人在文献[1]中的实验表明对于包含了某类物体的正例图像集,大部分的情况是分类和定位算法中只有一个能给出比较正确的结果,两者的结合能互相提升效果。分类和定位的结合引起学者们的兴趣,一些方法通过将图像分割为已标记的区域,并将这些标记用于周围物体的定位的结合方法取得了不错的定位效果[2,3]。文献[1]中提出了单类别分类和对象定位结合的概率模型,实验结果也再次证明了分类和定位技术对图像的认知具有一定独立性,两者结合能带来双方面的改进。

本文针对现实应用中图像资源庞大、类别多的特点,提出了两种将分类信息融入对象定位处理的方法。实际图像资源中,某类对象往往只存在于小部分的图像中,对所有图像进行该类对象的定位只会浪费时间和计算力。对此,本文提出先快

**收稿日期**: 2013-02-03; **修回日期**: 2013-03-29 **基金项目**: 国家自然科学基金资助项目(61021062); 国家"863"计划资助项目(2011AA01A202)

**作者简介**:钱怡(1991-),女,安徽人,硕士研究生,主要研究方向为媒体内容分析(qianayi@gmail.com);林莹(1986-),女,吉林人,硕士研究生,主要研究方向为对象拷贝检测;武港山(1967-),男,江苏人,教授,博导,主要研究方向为计算机视觉计算、多媒体信息检索、Web 智能信息处理.

速定位,再精确定位的处理方法,即先使用分类技术从原始图像资源中找出包含感兴趣对象的图像,再在后者中使用对象定位技术找出具体物体的位置。另外,由于对象定位的最终效果不仅与最终框定(最优框)的位置信息有关,还与定位方法对最优框的打分有关。相对于定位方法中通常只是用最优框的局部信息作为该框定的打分,本文提出将分类算法学习得到的全局信息融入最优框的打分机制中。相对于定位算法根据局部信息得出的打分,融入了分类全局信息后的新得分能够使得正确定位的得分在大量的图像集中脱颖而出,从而改进定位任务的效果。不同于其他相关研究,本文是将全局多类别分类信息而不是分割后的局部单类别的分类信息融入定位技术。本文选择目前主流的对象定位和分类方法来进行相关研究。

目前主流物体定位的方法有滑动窗口[4]和霍夫投票[5]方法,这两种方法差异很大。滑动窗口方法从图像所有可能的窗口中选择物体最可能存在的位置;霍夫投票方法则由霍夫决策树中多个节点投票决定出物体中心的位置,再确定物体所占区域。本文使用直接用矩形框框定出对象位置的滑动窗口方法。目前基于滑动窗口的物体定位方法在多个公开数据集均取得了不错的对象定位效果,然而巨大的窗口判定次数严重增加了定位时间。Lampert 等人[6]基于分支定界理论提出了快速子窗口搜索方法 ESS,在保证得到与原滑动窗口方法同样的最优解的情况下,极大地减少了窗口的判定次数,将定位速度提高到单幅图像单个物体定位时间低于 40 ms(2.4 GHz 的 PC 下处理一幅 500×333 像素的图像)。本文选择 ESS 方法来进行对象定位。

现实场景中的图像往往同时包含了多个类别的对象,蕴涵了多种简单语义。Zhou 等人[7]提出多示例多标记学习框架(MIML)来解决包含多种对象的图像分类和标注问题,这种新的学习框架用多个示例来描述一个对象,而一个对象对应多个标记,更好地诠释了对象与语义之间的复杂关系。诸多实验表明,基于 MIML 框架的算法比传统监督学习、多示例学习、多标记学习的算法更好地解决了自然场景和文本领域中的分类问题。MIML 学习框架下的算法作为一种多类分类方法,可以同时给出图像所属于的多个类别,其中的 MIMLSVM + 算法更适合大规模图像的分类问题,并且有不错的精度和效率。

本文对 ESS 和 MIMLSVM + 技术进行了深入的研究,并把它们应用到了对象定位处理中。实验结果表明,它们的有机结合能够提高对象定位的效果。

## 1　对象定位技术

滑动窗口方法用于物体定位由来已久,它根据分类器对图像中所有可能的子区域的打分,返回分值最高的框形区域作为对象的位置。由于滑动窗口需要遍历的窗口数目巨大,十分费时费力,研究者们提出了一些基于启发式搜索的方法来加快搜索速度,然而这些方法往往只能得到局部最优解,不能提供准确可靠的定位。Lampert 等人[8]基于分支定界理论提出 ESS 算法,在保证得到全局最优解的前提下极大地减少了窗口判定的次数。

分支定界方法是由图灵奖获得者“三栖学者”查理德·卡普在 20 世纪 60 年代发明,分支是从初始解空间开始将一组解分为几组子解,定界是建立这些子解空间的目标函数 $f$ 的边界。如果某一组子解的目标函数值在预先设定的边界之外,就将这一组子解舍弃(剪枝),从而不断逼近最优解。

ESS 对象定位方法基于分支定界理论,从对象定位的初始解空间(整个图像中所有可能的矩形区域)开始,将其分为两个子解空间,分别计算它们的上界,再选择拥有最大上界的子空间分解成两个子解空间并计算上界,不断重复选择拥有最大上界的子空间、分解、计算上界的过程,直到最大上界对应的子空间中只有一个矩形框,即得到了最优解。

ESS 为每个子解空间$\mathcal{R}$都计算一个上界 $\hat{f}(\mathcal{R})$,假设 $f(R)$ 为矩形区域 $R$ 的分值,这个上界需要满足:

a) $\hat{f}(\mathcal{R}) \geqslant \max_{R\in\mathcal{R}} f(R)$;

b) $\hat{f}(\mathcal{R}) = f(R)$,如果子解空间$\mathcal{R}$中只有区域 $R$。

ESS 中目标函数 $f$ 至关重要,而与其对应的求上界函数 $\hat{f}$ 需要保证快速收敛,即快速得到只有一个矩形区域的子解空间。ESS 采用词袋(bag of words)表示图像,对每类对象分别训练分类器,依据分类器对图像 $I$ 的打分得到目标函数 $f(I)$。词袋的字典根据提取所有图像的局部特征聚类得到。对于图像 $I$,根据字典生成其词袋表示 $h$,SVM 分类器对其判别函数为 $f(I) = \beta + \sum_i \alpha_i \langle h, h^i \rangle$,$h^i$ 为训练样本的词袋表示,$\alpha_i$ 和 $\beta$ 为 SVM 训练过程中学习得到的权重向量和偏移量。令 $\omega_j = \sum_i \alpha_i h_j^i$,可将上述函数转换为 $f(I) = \beta + \sum_{j=1}^{n} w_{c_j}$(具体转换可参考文献[8]),这里 $n$ 为 $I$ 中局部特征点的数目,$c_j$ 为第 $j$ 个特征点对应的聚类中心编号。这样以和的形式来代替线性内积,为计算某个区域 $R$ 的打分提供了可能。分类器对图像 $I$ 的打分 $f(I)$,为图像 $I$ 中所有特征点($j$)对应聚类中心($c_j$)所对应的权重($w_{c_j}$)的累加。同理,图像 $I$ 中某个区域 $R$ 的打分 $f(R)$ 可通过累加 $R$ 中的特征点对应权重之和来计算。聚类中心所对应的权重有正值和负值,因而 $f(R)$ 的累加过程中有或正或负的被加数。这里,用 $f^+(R)$ 表示其中所有正值的累加结果,$f^-(R)$ 表示其中所有负值的累加结果。

针对上面的打分函数 $f(I)$,解空间$\mathcal{R}$的上界函数定义如下:

$$\hat{f}(\mathcal{R}) := f^+(R_{\max}) + f^-(R_{\min}) \tag{1}$$

这里,$R_{\max}$ 和 $R_{\min}$ 分别是解空间$\mathcal{R}$中的最大和最小矩形区域;$f^+$ 和 $f^-$ 分别为 $f(R)$ 中正和负的被加数。最终 $\hat{f}$ 的计算在常量时间内,这正是 ESS 定位迅速的关键。

相对于简单地用词袋来描述图像对象,还有表现力更好的分层空间金字塔表示方法,它根据空间金字塔核[9]计算不同的目标函数 $f(I)$ 和相应的上界函数 $\hat{f}(\mathcal{R})$。

ESS 作为主流的对象定位方法,能够有效地对所有图像均返回一个或者多个最优窗口,并且针对每个窗口都有相应的打分。然而 ESS 作为纯定位算法,对完全不包含对象的图像也会经过一系列判定返回最优窗口。对于大规模、包含多类别对象的图像集而言,使用 ESS 对所有图像进行对象定位,肯定会耗时耗力(其中的大量图像可能根本不包含某类对象)。ESS 对每幅图像均返回一个最优框定及针对这个框定的打分,理想情况是正确的框定能有比较高的打分,有偏差或者错误的框定得分偏低,最优框的打分机制很大程度地影响着定位技术的效果,然而 ESS 对于最优窗口的打分完全基于窗口的局部信息,忽略了整幅图像包含的全局信息,存在一些不足。

## 2　图像标注方法

多个类别分类的问题向来是研究者们致力的热点。近年

来 Zhou 等人[7]提出的 MIML 有效地解决了自然场景分类、文本分类、果蝇基因标注等问题,极大地推动了多类别分类的研究。基于 MIML 的算法前期训练一个多类别分类模型,对于一张图像,可以一次性列出其可能属于的所有类别,完全不同于二值分类方法。不同的 MIML 算法在前期训练时花费的时间不一,然而判定图像类别的速度均极快。诸多文章中实验结果表明,MIML 算法对于现实场景的多类别分类往往具有极好的召回和精度。

多示例多标记学习与以往的传统监督学习、多示例学习[10]、多标记学习[11]有明显的不同。在这一框架下,一幅图像由多个示例描述,一个对象有多个标记,这些标记对应整个对象而不是某个示例,学习的目的在于预测整个示例包对应的类别标记。监督学习、多示例学习、多标记学习可以看做多示例多标记学习的特例。在对真实世界中的学习问题求解时,MIML 框架一方面更好地表示了对象,为后面的学习提供了关键的描述,另一方面它充分考虑了真实对象的多义性特点。尽管 MIML 是一种多对多的映射,但它为了解一个对象为什么具有某个类别标记提供了一种可能,还有助于涉及复杂概念的单标记学习问题的解决[12]。

MIML 框架下的算法有基于退化机制的 MIMLBOOST 和 MIMLSVM[7],基于正则化机制的 D-MIMLSVM[12] 和 M3MIML[13]。另外 Zha 等人[14]基于隐条件随机场提出 MLMIL 算法,Zhang 等人[15]提出 MIMLRBF 算法,利用 RBF 神经网络来进行学习;Nguyen[16]的 SISL-MIML 利用最大边学习框架;Zhou 等人提出的 MIMLSVM + 算法[17]则有效解决了果蝇基因标注问题。大规模图像集往往图像数量较多,且类别的正反例失衡,某些类别的图像可能只占比较小的一部分。MIMLSVM + 算法相对于其他的 MIML 算法而言,不仅适用于大规模学习问题,而且考虑到了类别不平衡问题[17],本文基于MIMLSVM + 算法来训练各类的分类器。

MIML 的形式化定义如下:$X \subseteq R^d$ 为所有示例的集合,$Y = \{1, 2, \cdots, Q\}$ 为类别标记的集合,MIML 的任务就是从样本集 $\{(X_i, Y_i) \mid 1 \leqslant i \leqslant N\}$ 中学习得到函数 $f_{\text{MIML}}: 2^X \rightarrow 2^Y$,其中 $X_i \subseteq X$,代表一组示例 $\{x_{1i}, x_{2i}, \cdots, x_{ni}\}$,其中每个示例 $x_{ni}$ 由一个一维特征向量来表示,而 $Y_i \subseteq Y$,表示与 $X_i$ 关联的标记集合 $\{y_{1i}, y_{2i}, \cdots, y_{li}\}$。这里 $n_i$ 是 $X_i$ 中包含的示例个数,$l_i$ 是 $Y_i$ 中所含标记的个数。参照文献[17],按照规则网格选择图像的关键点,所有关键点处的 SIFT 特征向量构成这幅图像的多示例包。

如同 MIMLBOOST、MIMLSVM 算法,MIMLSVM + 也是一种基于支持向量机的方法,它选用多示例核来作为支持向量机中的核函数。假设 $n$ 是训练集中图像的个数,$y \in Y$ 代表一个标记词,$X_i$ 为训练集中第 $i$ 幅图像的示例包(局部特征集),对每个类别 $y$,定义指示函数 $\varphi(X_i, y)$ 如下:如果第 $i$ 幅图像属于类别 $y$ 则 $\varphi(X_i, y) = 1$,否则 $\varphi(X_i, y) = -1$。最终的 SVM 分类模型主要解决下面的优化问题:

$$\min_{\omega_y, b_y, \varepsilon_{iy}} \frac{1}{2} \| \omega_y \|^2 + C \sum_{i=1}^{n} \varepsilon_{iy} \tau_y \tag{2}$$

$$\text{s.t.} \quad \varphi(X_i, y)(\langle \omega_y, \Phi(X_i) \rangle + b_y) \geqslant 1 - \varepsilon_{iy}$$

$$\varepsilon_{iy} \geqslant 0 \quad (i = 1, 2, \cdots, n)$$

这里,$\langle \cdot, \cdot \rangle$ 代表内积;$\Phi(X_i)$ 函数将示例包 $X_i$ 映射到高维空间;$\omega_y$ 和 $b_y$ 为高维空间中线性判别函数的参数;$\varepsilon_{iy}$ 为非负松弛变量,允许少量训练示例包被误分类;$\| \omega_y \|$ 反映了模型复杂度;$C$ 用于权衡模型复杂度和训练示例的累积误差;$\tau_y$ 为放大系数,用于解决类别不平衡问题。支持向量机中核函数的选择至关重要,MIMLSVM + 使用的是众所周知的多示例核[18],定义如下:

$$K_{\text{MI}}(X_i, X_j) = \frac{1}{n_i n_j} \sum_{x_{is} \subseteq X_i} \sum_{x_{js} \subseteq X_j} e^{-\| x_{is} - x_{js} \|^2} \tag{3}$$

这里:$X_i$ 为第 $i$ 幅图像的示例包;$n_i$ 为其中示例的个数;$\| x_{is} - x_{js} \|$ 为两个特征向量的相似度。多示例核 $K_{\text{MI}}(X_i, X_j)$ 充分利用示例包之间的关系来描述一个对象,相对于用特征本身,具有更多的有效信息,有助于学习任务的解决。

MIMLSVM + 作为多示例多标记新框架下的分类算法,能够同时给出图像中包含多个类别对象的信息,并且针对每个类别有不同的分类打分,所需要的判定时间短,具有针对大规模图像的类别不平衡处理机制,且在 HammingLoss、RankingLoss、OneError、Coverage、Average_Precision 等多类别分类指标上均有不错的表现,十分适用于本文的多类别分类任务。

## 3 基于 MIMLSVM + 预分类的对象定位处理

定位和分类作为图像领域的两大研究热点,它们对图像的认知结果不同,认知方法也具有一定的独立性。定位是根据图像中局部信息来判定出对象的位置,分类则综合上下文等全局信息给出图像的类别,两者相互裨益。对于大规模、包含多种对象的图像集的定位问题,如果对于其中每一幅图像进行某类对象的定位,其时间计算的消耗将十分巨大。另外,某种对象往往只存在于大规模图像集的一个小范围中,不需要对所有图像进行该类别的对象定位处理。因而不仅需要优秀的对象定位算法,还需要分类算法的预处理来缩小对象定位的图像范围。针对大规模、包含多种对象的图像集定位,本文提出了基于 MIMLSVM + 预分类的对象定位处理,先基于高分类精度的 MIMLSVM + 算法得到每类对象存在的目标图像,再使用 ESS 算法逐一进行对象定位。

本文对大规模、包含多种对象的原始图像集进行分类,得到多个子图像集,每个子图像集对应一类对象,而其规模往往远少于原始数据集,同时 MIML 算法的高召回率和精度能保证分类的效果。接下来本文只在这些子图像集上进行相应类别的对象定位,利用 ESS 方法最终用矩形框框出物体最可能存在的位置。

图 1 展示了基于 MIMLSVM + 预分类的对象定位处理流程。首先统一对训练集中所有图像进行预处理:按照均匀网格选取关键点并提取 SIFT 特征,即提取 dense SIFT 特征[19];然后基于用 dense SIFT 特征描述的训练集分别训练 MIML 中的多类别分类模型和 ESS 需要的简单分类器,两种分类器的正反例形式完全不同,如图 2 所示。在 MIML 框架下,学习的过程如下:训练集中的每一幅图像作为一个样本,用这幅图像的所有 dense SIFT 特征构成的矩阵来描述这个样本,并且每个样本都对应着一个一维向量作为标记,向量的长度为所考虑的对象类别的个数,上述描述信息和标记信息作为 MIML 学习算法的输入,经过相关学习得到多个类别的分类模型。ESS 所需要的简单分类器的训练则有很大不同,需要依次为每个类别训练一个分类器。对于某个类别,首先需要框出这个类别的正反例图像,如图 2,狗本身的截图作为正例,反例从狗的背景或其他不包含狗的图像中随机截出,然后依据所有正反例图像包含的关键点处的 dense SIFT 特征生成词袋分别来描述正反例样本,

训练得到一个简单的二值分类器,满足 ESS 后续的需求。可以看出,分类技术和定位技术中的分类器正反例形式、训练机制各不相同,有必要分别进行分类器的训练。

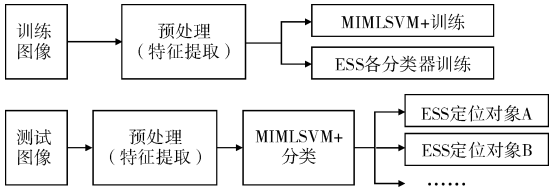



图 1　基于 MIMLSVM + 预分类的对象定位处理

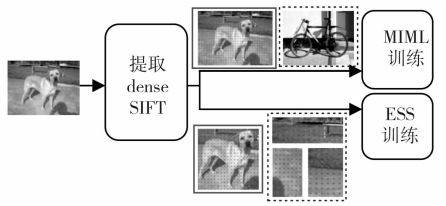



图 2　两个不同分类器的正反例

（实线图片框中为正例示例,虚线图片框中为反例示例）

对于测试集,同样需要预处理步骤提取相关特征等过程,得到 MIMLSVM + 算法所需要的多示例表示和 ESS 方法需要的词袋表示。对于测试图像,在 MIML 框架下学习到的分类模型可以直接判定出其中包含了哪几类对象,而 ESS 进行定位还需要事先生成测试图像的 clst 文件[8],clst 文件和分类器学习到的模型权重将一起作为 ESS 进行物体定位的依据。本文逐一进行不同类别的对象定位,得到相应的定位结果。定位任务的结果包含测试图像中物体最可能存在的位置和对这个位置的置信度打分,这里的打分将会在计算平均精度时作为排序的依据,影响着定位任务的效果评估。

## 4　利用分类信息提高 ESS 定位效果

ESS 方法对于每幅图像返回最优框的坐标信息和最优框的得分,对象定位任务的评估不仅要验证窗口框定的正确性,还考量窗口打分的合理性。理想情况是对于正确的框定给出较高的得分,反之给出较低的得分。通过前面的介绍可知,ESS 方法中最优框的得分主要由最优框中的关键点决定,虽然 ESS 定位取得了不错的效果,但是这里最优框的打分只利用了框内的局部信息。当利用 ESS 对图像集中的所有图像进行对象定位,结果发现如果按照最优窗口打分高低排序,那些打分比较高的窗口所在图像可能根本就不包含该类别的物体,这些图像对应的最优框虽然得分很高,却没有任何意义,这不仅和 ESS 对象定位方法中使用简单分类器带来的局限性有关,也与最优框打分机制只使用了局部信息有关。因而需要更合理的最优框打分机制。本文提出将分类方法学习得到的图像全局信息融入最优框的打分中,相对于之前 ESS 基于框内局部信息的打分,本文的分类信息融入着力于提高分类结果中正例图像(在高分类精度的情况,也就是确实包含了对象的图像)中最优框的打分,相对降低反例图像中最优框的打分,这样使得按打分排序时,正例图像的最优框尽可能地排在前面。MIML 学习框架下的算法在分类方面有令人满意的效果,其给出图像分类结果的同时,也给出了相应的一个得分,用于评定分类器认为图像属于某一类别的可信度。本文尝试将这一包含了 MIMLSVM + 分类算法学习得到的图像全局信息的得分融合到 ESS 得到的图像最优窗口的得分中去。

MIMLSVM + 对图像的多类别分类结果为一维向量 $V$,向量的长度等于类别的个数,如果 $V(i)>0$,则判定图像包含对象 $i$,反之,则不包含。假定对于对象 $i$,ESS 判定图像 $I$ 的最优窗口为 $R$,其得分为 $S_{\mathrm{ESS}}(R)$,而 MIML 算法对图像 $I$ 在第 $i$ 类对象上的打分为 $S_{\mathrm{MIML}}(I)=V_I(i)$,则计算这个窗口的最终打分为

$$S(R)=S_{\mathrm{ESS}}(R)\times \mathrm{e}^{S_{\mathrm{MIML}}(I)} \tag{4}$$

首先基于 ESS 方法对所有的图像进行物体定位,得到 ESS 对象定位结果,包含框的位置和框的打分;然后按照式(4)为所有的最优框均计算一个新的打分,而框的位置信息不作改变。将新的打分和对应框的位置作为最终结果。式(4)是一种取对数并作乘的结合方法,使得 MIML 算法给分比较高的图像可以得到更好的总得分。笔者还尝试过其他结合方法,如线性加权、线性乘积等,最终发现式(4)能够最大限度地带来定位效果的提高。

## 5　实验

针对基于 MIMLSVM + 预分类的对象定位处理和利用分类信息提高 ESS 定位效果这两种融合方法,在 PASCAL 2006 数据集上分别进行了两个实验。实验 1 针对 PASCAL 2006 的测试集,分别进行了基于 MIMLSVM + 预分类的对象定位和直接对所有测试集进行 ESS 对象定位,对比了从整个测试集而言,两者各自所需要的处理时间以及对象定位的平均精度。实验 2 在同样的数据集上,首先进行了整个测试集上的 ESS 对象定位,然后分别根据原始 ESS 最优框打分和融合了分类信息的新最优框打分来评估图像的对象定位任务。本实验在平均精度和精度召回曲线两方面均进行了比较。

PASCAL 2006 数据集作为对象定位领域常用的公开数据集,包含 5 304 张图像,其中 2 618 张图像用于训练和调整参数,2 686张图像用于测试和评估,涵盖了 10 个类别共 9 507 个物体。其中超过 23% 的图像同时属于多个类别,如表 1 所示。

表 1　PASCAL 2006 数据集

| 数据集 | 图像个数 | | 类别个数 | | 特征维度 |
|---|---|---|---|---|---|
| PASCAL 2006 | 5 304 | | 10 | | 128 |
| 每个样本示例个数 | | | 每个样本标记个数 | | |
| 最小 | 最大 | 平均 | $K=1$ | $K=2$ | $K\geqslant 3$ |
| 133 | 1 404 | 802 | 4 062 | 1 084 | 155 |

### 5.1　实验 1

本文在 PASCAL 2006 的测试集分别进行了基于 MIMLSVM + 预分类的对象定位和直接对所有测试集进行 ESS 对象定位。首先对 PASCAL 2006 中的所有图像进行预处理,按照规则网格选取关键点来提取 SIFT 特征[20],在 MIMLSVM + 分类算法中,一幅图像中所有关键点的特征组成多个示例来表示这幅图像。用 VLFeat[19] 工具包可以方便地提取 dense SIFT,其中 bin 大小和采样步长分别设为 8 个像素和 16 个像素。根据图像大小不同,一幅图像最终最少包含 133 个示例,最多包含 1 404 个示例,每个示例均为 128 维的 SIFT 特征。对于 ESS 算法中,正反例是图像中的一个区域,由于 dense SIFT 不仅包含了 SIFT 特征还保存了关键特征点在图像中的位置信息,可以很容易地据此得到 ESS 方法中正反例的特征描述。

MIMLSVM + 算法采用多示例核为每个类别训练一个 SVM

分类器。基于 PASCAL 2006 数据集的验证子集来选择正则化参数 $C$。实验显示,在 2.4 GHz PC 上 MIMLSVM + 判定 2 686 幅测试图像仅需要 1.5 s。对于自行车、小汽车、猫、狗这四类,MIMLSVM + 分别判定测试集中有 236、512、318、275 幅该类别的图像,平均返回 335 幅图像,召回率和精度如表 2 所示。

表 2 四个类别上 MIML 的召回率、精度和定位平均精度

| class | precison | recall | AP( MIML-ESS) | AP( ESS) |
|---|---|---|---|---|
| bicycle | 0.716 | 0.631 | 0.337 | 0.289 |
| car | 0.803 | 0.756 | 0.229 | 0.223 |
| cat | 0.642 | 0.526 | 0.254 | 0.166 |
| dog | 0.491 | 0.365 | 0.159 | 0.126 |

假定 MIMLSVM + 判定为正例的图像一定包含该类别的物体,并且只对这部分图像集进行后续的对象定位。考虑到进行所有十个类别的定位工作量比较大,选择了自行车、小汽车、猫、狗这四类进行后续定位。

对于每一类对象,分别训练了一个 SVM 分类器。正例来自于 PASCAL 2006 数据集本身提供的真实物体框定,反例分为两种情况:a)在包含此类物体的图像中非物体部分随机框选;b)在不包含此类物体的图像中随机框选。正反例比例约为 1∶10。对每一类,本文对正反例图像上的 dense SIFT 特征进行 K-means 聚类得到大小为 1 000 的码本,用对应词袋表示正反例,进行训练得到分类模型。从此模型中可以导出 ESS 定位所需要的权重文件,这是后续定位工作的一个重要依据。ESS 方法还需要测试图像的一些信息,例如高度、宽度、clst 信息( clst 信息包含了测试图像中所有关键点的横向、纵向坐标和此点处的 SIFT 特征所在的聚类中心的编号)。

针对自行车、小汽车、猫、狗这四个类别,本文在 Linux 下对测试集图像进行了批量的 ESS 来定位各类别物体。假设返回的是每幅图像中 ESS 打分最大的矩形框 $r'$,真实框定为 $r$,如果满足式(5)则认为矩形框 $r'$定位正确。

$$\frac{\text{area}(r\cap r')}{\text{area}(r\cup r')}\geqslant 0.5 \tag{5}$$

ESS 在 2.4 GHz PC 上为一幅图像定位一个物体所需要的时间平均约为 40 ms,在大小为 2 686 幅图像的测试集上,比较直接地对全部测试图像定位和先分类再在小的范围定位所需要的时间(四个类别的平均值),结果如表 3 所示。可以看出后者能节省大量的时间,可以预知,对于更大规模数据集的定位问题,后者将更明显地提高物体定位的效率,而且 MIML 框架一次性完成对多个类别的分类,类别越多,越能减少整体的定位时间。本文将 MIML 判为反例的图像的得分均设为 0,正例图像打分为 ESS 所得分数,根据 PASCAL VOC 数据集提供的评估软件,在 MIML 筛选后的子集上各类别的平均精度如表 2 所示。可见四个类别上平均精度均高于 ESS 在所有测试图像上的定位平均精度,这不仅与测试集的范围不一样有关(前者是在测试子集上,后者是整个测试集上),更离不开 MIMLSVM + 分类算法高的召回率和精度。

表 3 本文方法和 ESS 的时间对比

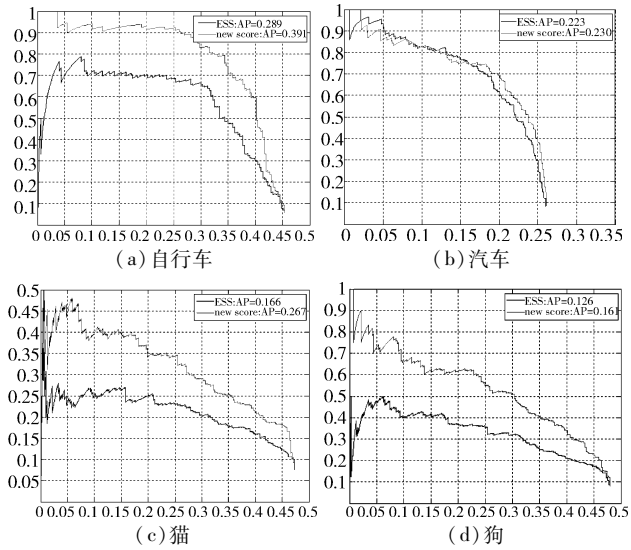
| method | time/s |
|---|---|
| ESS | 107.44 |
| proposed method | 15 |

### 5.2 实验 2

本文对比了两种不同的最优框打分机制。首先对 PASCAL 2006 的所有测试图像均进行了 MIMLSVM + 分类和 ESS 对象定位处理,具体预处理提取特征和分类、定位的做法参照实验 1。

根据式(4)为测试集中所有图像的 ESS 返回框的得分进行了优化,并根据 PASCAL VOC 数据集提供的评估软件得到平均精度表格,如表 4 所示。可以看出新的打分一定程度上提高了物体定位任务的平均精度,在自行车、猫和狗这三个类别上均有所提高,特别是前两个类别上均提高了 10% 以上。然而在汽车类别上的提升较小,本文分析了相关原因。汽车类图像同其他类有不同,同一幅图像中含有很多个汽车的比例非常大,本文进行的 ESS 实验中,只考虑了对一幅图像返回一个最优矩形框的情况,这样其他包含汽车的区域都不可能再被框出,受此限制,新的打分带来的效果提高略逊于其他类别。

根据 PASCAL 2006 数据集提供的工具画出这四个类别上的精度—召回率曲线,如图 3 所示。每条曲线最右边的点对应着所有图像各返回一个物体时的召回率和精度情况,如果设定物体框得分的下限值,只有超过这个值的框才被考虑。可以通过不断提高这个下限值得到左边的曲线,越往左边,更高得分的最优框才被考虑,越少的框被用于计算召回率和精度。如图,浅色线代表新的最优框打分,黑线代表 ESS 原本的最优框打分,可以看出前者更准确地反映了最优框的框定效果。



(a)自行车 (b)汽车 (c)猫 (d)狗

图 3 优化后的打分对比原始打分的 PR 图

(横坐标为召回率,纵坐标为精度)

表 4 为两种打分对应的平均精度,新的打分机制在四个类别上均带来了提高。可见利用 MIML 带来的全局信息对原始的 ESS 定位的优化,新的得分可以更好地拉开正确框定和错误框定的得分差距;随着下限值的提高,剩下来的更多的是正确框定,最终带来了平均精度的提升。

表 4 优化后的 ESS 和原始 ESS 的定位平均精度对比

| class | ESS | new score |
|---|---|---|
| bicycle | 0.289 | 0.391 |
| car | 0.223 | 0.230 |
| cat | 0.166 | 0.287 |
| dog | 0.126 | 0.161 |

## 6 结束语

为提高图像中对象定位技术的处理效果,本文对对象定位技术和分类技术的融合进行了研究。一方面针对实际应用中图像集规模大、所含对象类别多的特点提出基于 MIMLSVM +

预分类的对象定位处理方法,通过在多类别分类领域表现突出的 MIMLSVM + 来缩小需要定位的图像范围,然后再通过快速子窗口定位方法 ESS 用矩形框框出物体。在图像量庞大、所含对象类别繁杂的情况下,本文前期的分类算法剔除大部分不包含对象的图像,大大减少了后续 ESS 进行定位的次数,从而大大减少了判定时间,取得了不错的定位平均精度。另一方面,为了得到更好的对象定位效果,将分类算法学习得到的全局信息融入定位算法对最优窗口的打分中,综合了分类和定位技术的打分更好地反映了正确矩形框得分和错误矩形框得分的差别,更符合客观事实。实验表明,综合后的打分在不同类别上均带来了定位效果的提高,再次证明了分类和定位技术对图像的认知具有一定的独立性。

接下来的工作是更紧密地利用分类和定位技术的认知独立性,提出更好的二者融合的方法,最终以更快捷更有效的方法来解决多类别物体定位问题。

**参考文献:**


[1] HARZALLAH H, JURIE F, SCHMID C. Combining efficient object localization and image classification[C]//Proc of the 12th IEEE International Conference on Computer Vision. 2009:237-244.

[2] HEITZ G, KOLLER D. Learning spatial context: using stuff to find things[C]//Proc of the 10th IEEE International Conference on European Conference on Computer Vision. 2008:30-43.

[3] WOLF L, BILESCHI S. A critical view of context[J]. International Journal of Computer Vision, 2006, 69(2):251-261.

[4] DAKAK N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005:886-893.

[5] GALL J, LEMPITSKY V S. Class-specific hough forests for object detection[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2009:1022-1029.

[6] LAMPERT C H, BLASCHKOM B, HOFMANN T. Beyond sliding windows: object localization by efficient subwindow search[C]//Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2008:1-8.

[7] ZHOU Zhi-hua, ZHANG Min-ling. Multi-instance multi-label learning with application to scene classification[C]//Advances in Neural Information Processing Systems. 2007:1609-1616.

[8] LAMPERT C H, BLASCHKOM B, HOFMANN T. Efficient subwindow search: a branch and bound framework for object localization [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31(2):2129-2142.

[9] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories[C]// Proc of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006:2169-2178.

[10] MARON O, RATAN A L. Multiple-instance learning for natural scene classification[C]//Proc of the 15th International Conference on Machine Learning. 1998:341-349.

[11] BOUTELL M R, LUO J, SHEN X. Learning multi-label scene classification[J]. Pattern Recognition, 2004, 37(9):1757-1771.

[12] ZHOU Zhi-hua, ZHANG Min-ling, HUANG Shen-jun, *et al*. Multi-instance multi-label learning[J]. Artificial Intelligence, 2012, 176(1):2291-2320.

[13] ZHANG Min-ling, ZHOU Zhi-hua. M3MIML: a maximum margin method for multi-instance multi-label learning[C]//Proc of the 8th IEEE International Conference on Data Mining. 2008:688-697.

[14] ZHA Zheng-jun, HUA Xian-sheng, MEI Tao. Joint multi-label multi-instance learning for image classification[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2008.

[15] ZHANG Min-ling, WANG Zhi-jian. MIMLRBF: RBF neural networks for multi-instance multi-label learning [J]. Neurocomputing, 2009, 72(16-18):3951-3956.

[16] NGUYEN N. A new SVM approach to multi-instance multi-label learning[C]//Proc of the 10th IEEE International Conference on Data Mining. 2010:384-392.

[17] LI Ying-xin, JI Shui-wang, KUNAR S, *et al*. Drosophila gene expression pattern annotation through multi-instance multi-label learning [J]. ACM/IEEE Trans on Computational Biology and Bioinformatis, 2012, 9(1):98-112.

[18] GARTNER T, FLACHP A, KOWALCZYK A. Multi-instance kernels [C]//Proc of the 19th International Conference on Machine Learning. 2002:179-186.

[19] VEDALDI A, FULKERSON B. VLFeat: an open and portable library of computer vision algorithms[C]//Proc of International Conference on Multimedia. 2010:1469-1472.

[20] LOWE D G. Distinctive image features from scale invariant key points [J]. International Journal of Computer Vision, 2004, 60(2):91-110.


(上接第 3832 页)

## 4　结束语

本文介绍了等值线绘制的两种方法,即多直线叠加法和多像素绘制法的算法流程。其中多直线叠加法添加了直线的反走样绘制,解决了等值线走样问题,并且通过不同透明度的直线的叠加实现了等值线的渐变绘制。多像素绘制法将单个等值线看成具有一定面积的矩形,根据矩形中像素点到矩形主线的距离赋予像素点不同的透明度值,实现了等值线的渐变效果,符合人眼对亮度的色彩视觉特性,比多直线叠加绘制法绘制效果自然,易于接受。实验结果表明,用多直线绘制法和多像素绘制法绘制的等值线图,符合人眼的视觉特征,等值线图自然,具有艺术效果。

**参考文献:**


[1] 王鹏,周茂林,姚兴苗,等. 等值线绘制中的多重网格剖分快速搜索算法[J]. 计算机应用研究,2011,28(6):2346-2347,2351.

[2] 王学潮. 等值线绘制简易算法[J]. 北华航天工业学院学报,2012,22(4):21-23.

[3] 王鹏. 等值线快速绘制方法研究及系统设计与实现[D]. 成都:电子科技大学,2011.

[4] 爨莹,朱航洲. 一种新的等值线绘制方法研究[J]. 微型技术,2012,31(3):46-49.

[5] 周燕,金伟其. 人眼视觉的传递特性及模型[J]. 光学技术,2002,28(1): 57-59,62.

[6] 唐艳红. 海洋环境信息等值线自动绘制方法研究[D]. 哈尔滨:哈尔滨工程大学,2007.

[7] 刘伟奇,冯睿,周丰昆. 符合人眼视觉特性的颜色亮度模型[J]. 光学学报,1999,19(10):1426-1429.

[8] 白廷柱,金伟其. 光电成像原理与技术[M]. 北京:北京理工大学出版社,2006.

[9] 徐俊波. 基于 OpenGL 的科学可视化算法研究与实现[D]. 哈尔滨:哈尔滨工程大学,2005.

[10] 王美兰,赵继成,秦东华. OpenGL 及其在 VC ++ 下的开发应用[J]. 武汉大学学报,2006,39(4):62-65.