结合物体先验和空域约束的室内空域布局推理

姚拓中 [1] 左文辉 [2] 宋加涛 [1] 应宏微 [1]

摘 要 对结构化室内场景的空域布局结构进行估计是计算机视觉领域的研究热点之一. 然而, 对于内部堆放了众多杂乱物体的室内场景, 现有的大多数方法容易受到各种物体遮挡的影响而无法对这一类场景的布局结构进行准确推理. 为此, 本文方法充分考虑了房间和物体之间的几何和语义关联性, 参数化地对房间和内部物体的三维体积分别进行描述, 并且提出利用多种高层图像语义来获取物体的先验信息. 此外, 还在此基础上加入了空域排他性和空域位置等多种空域约束, 进而在改进室内场景空域布局估计的同时为物体的识别和定位提供关键信息. 本文方法不仅具有较低的求解复杂度, 而且通过试验表明相比于现有的经典方法在杂乱的室内场景中能够取得更为鲁棒的空域布局推理结果.

关键词 空域布局推理, 物体先验, 空域约束, 组合优化

引用格式 姚拓中, 左文辉, 宋加涛, 应宏微. 结合物体先验和空域约束的室内空域布局推理. 自动化学报, 2017, **43**(8): 1402−1411

DOI 10.16383/j.aas.2017.c160043

Estimating Spatial Layout of Cluttered Rooms by Using Object Prior and Spatial Constraints

YAO Tuo-Zhong[1] ZUO Wen-Hui[2] SONG Jia-Tao[1] YING Hong-Wei[1]

Abstract Estimating spatial layout of a structural indoor scene is one of the research hotspots in computer vision. However, most of the current solutions cannot work robustly in a cluttered room due to occlusion of different objects inside. In this paper, a new algorithm which integrates geometric and semantic relations between room and objects is proposed to recover the spatial layout of a cluttered room. This algorithm parametrically represents the 3D volume of both room and objects and uses multiple high-level image semantics to obtain object priors. Furthermore, several spatial constraints such as spatial exclusion and containment are used which simultaneously optimize spatial layout estimation of the room and provide significant information for object recognition and localization. One advantage of the algorithm is its low computational complexity, and experimental results also demonstrate that it can work more robustly in cluttered rooms than several classic algorithms.

Key words Spatial layout estimation, object prior, spatial constraint, combinational optimization

Citation Yao Tuo-Zhong, Zuo Wen-Hui, Song Jia-Tao, Ying Hong-Wei. Estimating spatial layout of cluttered rooms by using object prior and spatial constraints. *Acta Automatica Sinica*, 2017, **43**(8): 1402−1411

室内场景的三维空域布局推理在计算机视觉的诸多领域均具有非常重要的价值, 例如机器人的自主导航以及自动物体识别和安放等. 人类通常通过空域推理能力对室内场景中存在的各个平面和物体的尺寸和位置等信息进行理解, 例如, 能够识别桌子和沙发等家具并对其结构进行描绘, 或者发现沙发的某部分遮挡了床并且两者之间存在一定的间距等. 然而, 使计算机具备人类具有的上述空域布局理解能力对于计算机视觉而言是一个具有挑战性的工作.

收稿日期 2016-01-21 录用日期 2016-07-28
Manuscript received January 21, 2016; accepted July 28, 2016
浙江省自然科学基金 (LQ15F020004), 浙江省公益类技术研究项目 (2016C33255), 宁波市自然科学基金 (2015A610132, 2013A610113) 资助
Supported by Zhejiang Provincial Natural Science Foundation (LQ15F020004), Zhejiang Provincial Public Welfare Technology Research Project (2016C33255), and Ningbo Natural Science Foundation (2015A610132, 2013A610113)
本文责任编委 贾云得
Recommended by Associate Editor JIA Yun-De
1. 宁波工程学院电信学院 宁波 315016 2. 浙江大学信息与电子工程学院 杭州 310027
1. School of Electronic and Information Engineering, Ningbo University of Technology, Ningbo 315016 2. College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027

1 相关工作

迄今为止, 已有不少基于参数化场景空间的方法用于从诸如 “曼哈顿世界” (Manhattan world)[1] 等受约束的室内场景中恢复出相应的三维结构模型[2−3]. 这些基于单幅图像的方法通常采用诸如消失点估计[4−5] 以及几何结构预测[6−7] 等经典解决思路. 然而, 上述方法只关注室内场景的三个主方向估计, 并没有尝试提取房间结构以及物体尺寸等更为详细的三维描述信息, 因而仅能用于没有杂乱物体堆放的空房间. 相比之下, 由于物体遮挡造成的房间

形状结构模糊化更具挑战性.

目前, 已有一部分工作尝试了从单幅图像中对杂乱堆放了众多物体的场景进行三维结构恢复. 受路径规划应用的启发, Nabbe 等使用基于图像的外观模型将室外场景标注成地平面、垂直地面区域以及天空区域三个大类[8], 但其并没有在室内场景中进行相关试验. Micusik 等在对比试验中采用类似的场景几何和语义标注方法[9] 对场景的布局结构进行描述, 被应用到室内场景中并取得了一定的效果[10]. 此外, 还有一些方法致力于推理场景的深度[11−12] 和几何语义属性[13]. 不过, 此类方法在杂乱的室内场景中往往仅能实现较为粗略的空域布局推理结果, 无法准确勾勒出房间的空间结构.

最近几年的一些研究采用不同的参数化表示法对物体和房间之间的上下文关系进行建模并取得了良好的效果. Gupta 等根据 "积木世界" 概念对室内外场景进行解析, 并在基于立方体表示的房间地图基础上对位于其中的物体三维结构进行估计[14]; Lee 等同样利用了积木世界中的约束规则来将物体建模成与墙壁和地板对应的轴相平行的立方体[3, 15]; Hedau 等从图像中恢复杂乱堆放物体的标注并且使用简单的先验信息获取物体在三维场景中的空域位置[2, 16]; Wang 等提出的类似方法则不需要杂乱场景的人工标注[17]. 然而, 上述方法均将房间和物体的空域结构分开进行分析, 没有考虑到两者之间存在密切的几何和语义关联性, 进而影响了最终结果的鲁棒性. 值得注意的是, 目前已有一小部分工作开始致力于对室内场景中房间和物体的空域布局实现同步推理并取得了一定的成效[18−19], 但是上述方法主要通过构建复杂的图模型进行参数求解, 由于假设空间巨大造成算法的复杂度过大, 进而影响了算法的效率和可靠性.

2　算法描述

相比于将场景中的物体以积木分块形式进行建模实现场景空域布局定性推理的方法[14−15], 本文采用更为简化的参数化模型, 即在立方体表示法的基础上同时对房间的空域结构及其内部物体的分布进行联合推理, 基本流程如图 1 所示.

1) 本文算法提取房间内的直线段并估计相互正交的三个主消失点, 上述消失点定义了房间中各个平面 (例如不同朝向的墙壁、天花板和地板等) 的主方向并为房间内部的地板, 墙面以及天花板等提供了空域约束.

2) 结合上述几何信息和多种高层图像语义分别生成房间和物体的初始结构假设 (均用立方体表示).

3) 在房间和物体结构假设的基础上, 生成一系列候选的场景配置假设 (房间假设 + 物体假设).

4) 由于并非所有房间和物体的结构假设都满足场景配置假设的约束, 为此本文使用简单的三维空域推理对上述约束进行强化, 并对每个 "房间 − 物体" 假设对以及 "物体 − 物体" 假设对进行空域兼容性测试并挑选出满足要求的场景配置.

5) 在最终的场景配置假设推理中, 为了有效减少场景配置假设搜索的计算复杂度, 本文利用基于经典的组合优化法来采样出最优的场景配置.

3　房间结构假设的生成

与文献 [2] 类似, 本文通过两个步骤生成房间的

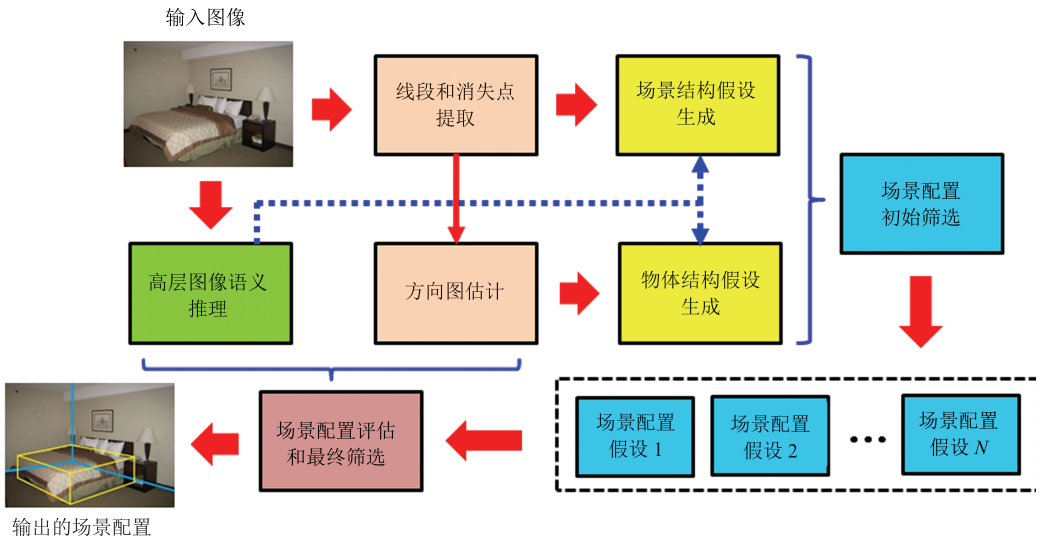
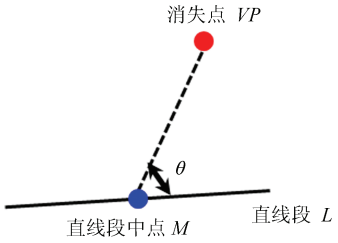


图 1　本文算法的基本流程

Fig. 1　The flowchart of our algorithm

结构假设. 1) 使用基于几何的方法对三个相互正交的主消失点进行估计以获取盒子的朝向信息, 如图 2 所示. 其中, 直线段到消失点的角距离定义为该直线段与其中点到该消失点连线之间的夹角, 如图 2 (a) 所示. 2) 通过对与消失点方向相一致的直线段对进行采样, 获取具有朝向一致性的墙面对应的参数化表达和尺度信息. 为了选择最优的房间结构假设, 采用结构化学习对每个候选的房间结构假设进行评估, 进而得到对应的置信度估计.



(a) 角距离
(a) Angle distance

(b) 直线段组
(b) Straight line groups

图 2 角距离和直线段组的定义
Fig. 2 The definitions of the angle distance and straight line groups

3.1 房间结构的朝向估计和参数化表达

本文用立方体结构对每个房间进行建模, 并且绝大多数室内平面的朝向均与该立方体的三个主方向相一致. 目前, 已有不少方法可实现对图像中相互正交的消失点集进行估计. 例如 Hedau 等提出的经典方法[2], 使用基于指数方式的投票策略得到直线段和消失点之间角距离微分的评分, 并通过基于随机一致性采样 (Random sample consensus, RANSAC) 的搜索策略对所有的主消失点根据可靠性进行排序. 其中, 候选的消失点为所有检测得到的直线段两两相交得到的交点, 而消失点集则从上述交点中选取. 基于指数方式的投票策略好处在于可以使得消失点的投票空间具有多峰的特性, 从而有助于将最优消失点与其他候选的消失点进行有效区分. 在本文中, 长度超过 30 个像素的直线段将被保留用于消失点的估计. 当确定最优的主消失点后, 图像中提取的每条直线段将根据朝向被分别分配给相应的消失点, 从而构成不同的直线段组. 在图 2 (b) 中, 归属于不同消失点的直线段被赋予不同的颜色, 而投票值低于设定阈值的直线段则被赋予蓝绿色.

基于立方体结构表述的房间朝向信息对于其各个角的投影施加了严格的几何约束, 如图 3 所示. 在图像平面中, 最多可以看到房间结构假设的 5 个平面, 分别对应于 3 个墙面、1 个天花板和 1 个地板. 房间结构假设中处于正面视点的四个角被分别定义为 A、B、C 和 D, 它们在二维图像中对应于 a、b、c 和 d. 三个相互正交的消失点分别为 VP_1、VP_2 和 VP_3, 它们满足以下三个条件: 1) 线段 ab 和 cd 与消失点 VP_1 共线; 2) 线段 ad 和 dc 与消失点 VP_2 共线; 3) 消失点 VP_3 位于矩形 $abcd$ 的内部.

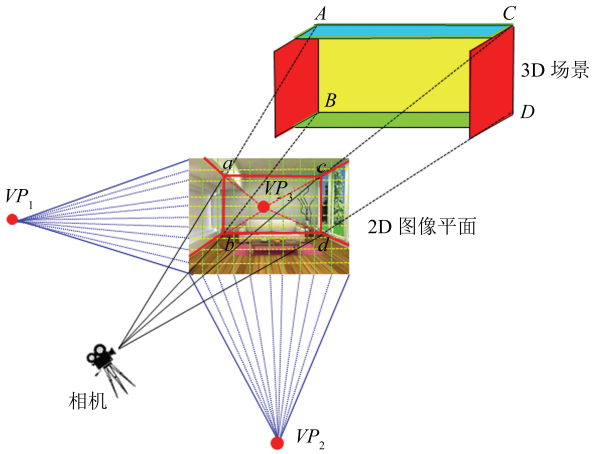


图 3 基于立方体描述的房间结构假设
Fig. 3 The cubic based room hypothesis

为了生成候选的房间结构假设集, 选取距离图像中心最远的两个消失点 VP_1 和 VP_2, 并通过设置一定的采样间隔从上述 VP_1 和 VP_2 朝着 VP_3 分别生成一定数量的直线段对. 上述归属于 VP_1 和 VP_2 的直线段相交形成了房间结构假设中处于正面视点的四个角 a、b、c 和 d, 而其余的可通过上述 4 个角到 VP_3 的连线生成. 当房间结构假设中可见的平面数目少于 5 个时, 房间结构假设中的四个角将位于图像外部.

如图 3 所示, 从 VP_1 和 VP_1 分别发射出 10 条射线以在图像平面中生成候选的房间结构假设集. 图 4 给出了部分候选的房间结构假设, 每个房间假设由分别从 VP_1 和 VP_2 发射的两条蓝色直线段所

构成, 进而生成描述房间三维结构的立方体所对应的 4 个角和 4 条边, 而立方体剩余的边则通过与 VP_3 进行连接得到.

图 4　候选的房间结构假设集

Fig. 4　Candidate room hypothesis set

3.2　候选房间结构假设的置信度估计

本文根据与训练集中人工标注的房间三维结构进行对比, 实现对房间结构假设进行排序. 假设室内训练图像集由 n 幅图像构成, $\{x_1, x_2, \cdots, x_n\} \in X$, 它们相应的房间结构假设 $\{y_1, y_2, \cdots, y_n\} \in Y$, 目的是学习映射关系 $f: X, Y \rightarrow R$, 使其能够赋予每个候选的房间结构假设相应的置信度评分. 在这里, 每个房间结构假设均被参数化为由五个平面构成的空间结构 $y = \{S_1, S_2, \cdots, S_5\}$. 映射关系函数 f 需满足: 输入图像 x_i 对应的房间结构假设 y_i 与真实假设 y 越接近, $f(x_i, y)$ 的值越高, 反之 $f(x_i, y)$ 的值下降. 那么, 房间结构假设的最优估计 y^* 可通过下式求解

$$y^* = \arg\max_y f(x, y, w) \tag{1}$$

式 (1) 是一个典型的结构化回归求解问题, 其输出为一个立方体结构的房间结构假设. 为了对其进行求解, 可采用文献 [20] 方法中的结构化学习框架, 通过利用训练集对输入空间中不同输出之间的关系进行建模, 通过经典的二次规划算法进行求解. 其中, $f(x, y) = w^{\mathrm{T}} F(x, y)$, 可利用式 (2) 对权重 w 进行学习:

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i \\ &\text{s.t.} \quad \xi_i \geq 0, \qquad \forall i \\ &\qquad w^{\mathrm{T}} F(x_i, y_i) - w^{\mathrm{T}} F(x_i, y) \geq D(y_i, y) - \xi_i, \\ &\qquad\qquad \forall i, \ \forall y \in Y \backslash y_i \end{aligned} \tag{2}$$

其中, y_i 为人工标注的 Ground truth 房间结构假设, ξ_i 为松弛变量, $D(y_i, y) = D_1(y_i, y) + D_2(y_i, y) + D_3(y_i, y)$ 为描述不同房间结构假设之间差异的代价函数. $D_1(y_i, y) = \sum_{j\in[1,5]} \delta(S_{ij}, S_j)$ 惩罚了某一房间结构假设中的某个平面 S_j 缺失而在另一房间结构假设中出现的情况; $D_2(y_i, y)$ 度量了两个房间结构假设之间的平面中心 c_j 的位置偏移; $D_3(y_i, y) = \sum_{j\in[1,5]} (1 - A(S_{ij} \cap S_j)/A(S_{ij} \cup S_j))$ 为两个房间结构假设之间各个平面之间的像素误差之和, 度量了房间结构假设之间的重合度. 其中 $A(\cdot)$ 为平面的面积, 当满足 $A(S_{ij}) > 0$, $A(S_j) = 0$ 或者满足 $A(S_{ij}) = 0$, $A(S_j) > 0$ 时, $\delta(S_{ij}, S_j) = 1$, 否则 $\delta(S_{ij}, S_j) = 0$.

$F(x_i, y)$ 为从房间结构假设 y 中提取的特征向量, 可通过与主消失点方向一致的直线段组进行计算得到. 在本文中, $F(x_i, y)$ 由基于几何的低层特征 F_g 和基于语义的高层特征 F_s 两部分组成. 对于每个平面 S_j, 基于几何的直线段组非加权性特征 f_l 定义如式 (3) 所示. 其中, L_j 为位于 S_j 中的直线段集, R_j 为位于 S_j 中与两个消失点 VP_1 和 VP_2 朝向一致的直线段集, $|l|$ 表示直线段 l 的长度. 最终, $F_g = \{f_l(S_1), f_l(S_2), f_l(S_3), f_l(S_4), f_l(S_5)\}$.

$$f_l(S_j) = \frac{\sum\limits_{l_i \in R_j} |l_i|}{\sum\limits_{l_i \in L_j} |l_i|} \tag{3}$$

当房间结构假设中的每个平面通过消失点 VP_1 和 VP_2 进行参数化后, 每个平面中的绝大多数直线段根据朝向将归属于上述两类消失点. 然而, 位于物体上的部分直线段并不满足上述情况, 例如图 2 (b) 中位于沙发的部分蓝色直线段应对应于水平消失点, 但是其朝向却显然与水平方向并不一致. 为此, 本文同样将直线段未落入物体区域中的置信度估计 $p(l_i)$ 作为权重来计算直线段组, 其可通过高层图像语义推理得到. 最终, 基于语义的直线段组加权性特征 f_s 定义如式 (4) 所示. 其中, $F_s = \{f_s(S_1), f_s(S_2), f_s(S_3), f_s(S_4), f_s(S_5)\}$.

$$f_s(S_j) = \frac{\sum\limits_{l_i \in L_j} p(l_i) \times |l_i|}{\sum\limits_{l_i \in L_j} |l_i|} \tag{4}$$

4　物体结构假设的生成

4.1　基于高层图像语义的物体位置估计

在杂乱的房间里通常堆放着桌子、椅子、沙发等物体, 它们的存在模糊了房间各个平面的边界. 而且, 使用的某些用于确定房间结构假设的特征往往会位于上述物体中, 从而对房间结构假设的准确推理造成困难. 如果能够得到上述物体所在的准确位置估计, 将有助于对先前预测得到的房间结构假设进行优化. 同样, 一个较为准确的房间结构假设同样将对房间中各个平面和物体实现更为准确的定位.

为了对物体的位置进行估计, 本文采用两种经典算法生成高层图像语义特征. 1) 场景的表面布局估计 (Surface layout estimation, SLE)[9]; 2) 基于

全体前景和背景假设排序的物体识别模型 (Object recognition model, ORM)[21].

在 SLE 中, 对算法[9] 进行相应的改进以适用本文的应用. 将平面的类别分为地板 (Floor)、左侧墙面 (Left wall)、中侧墙面 (Front wall)、右侧墙面 (Right wall)、天花板 (Ceiling) 和物体 (Object) 六大类. 在提取房间结构假设的特征时, 将分割块中每种平面类别的面积百分比以及彼此之间的重合度作为主要特征进行学习, 目的是提高没有物体放置时不同房间平面之间的区分度. 在 ORM 中, 在多尺度分割的基础上利用上述特征对六种平面类别进行学习, 实现对房间中杂乱堆放物体的检测和定位. 图 5 给出了通过挖掘不同高层图像语义得到的物体位置估计结果. 在图 5 (a) 中, 不同的平面类别通过不同的颜色表示, 红色、蓝色、黄色分别表示左侧墙面、中间墙面、右侧墙面, 绿色和紫色分别表示地板和物体. 在图 5 (b) 中, 高亮度区域为物体区域的定位结果.

(a) SLE 物体分布图
(a) SLE object map

(b) ORM 物体分布图
(b) ORM object map

图 5 基于不同高层图像语义的物体位置估计
Fig. 5 Different high-level image semantic based object localization

对于基于语义的直线段组特征而言, 将直线段上各个像素不属于平面类别 Object 对应的置信度作为 $p(l_i)$, 对式 (5) 进行计算. 其中, $p(l_i)$ 通过 SLE 和 ORM 方法分别得到的置信度加权获得. 与文献 [2] 不同, 不通过递归的方式直接筛选出最优的房间结构假设, 而是赋予每一个候选的房间结构假设相应的置信度估计, 用于最优场景配置假设的筛选.

4.2 候选物体结构假设的置信度估计

本文将物体进行基于立方体的参数化, 从而较好地描述其在房间中占据的空间大小, 并采用一种较为简单的方法生成物体结构假设. 在已知三个相互正交的消失点 VP_1、VP_2 和 VP_3 的基础上, 通过文献 [5] 方法估计相机的内参矩阵 K 以及对应于房间的旋转矩阵 R.

假设三维坐标系的零点位于相机的光心, x 轴、y 轴和 z 轴的朝向分别与房间的宽度、高度以及深度方向一致. 那么, 坐标系中的点 $\bar{X}$ 以及与其图像平面上对应点 $\bar{x}$ 之间的关系可通过如下投影关系描述. 为了生成物体结构假设, 假设相机高度 h_c 为一个随机值. 在物体结构假设中, 每个位于地板上的角点 $\bar{X}$ 需满足 $\bar{X}^{\mathrm{T}} + h_c = 0$, 其中 $n = (0, 1, 0)$ 为地板平面的法线. 利用上述约束可以确定物体结构假设的参考角点, 其他的角点可根据物体的三维尺寸推算. 上述角点在图像上的投影可通过式 (5) 得到

$$c\bar{x} = KR\bar{X} \tag{5}$$

这里通过对不同的相机高度和物体三维尺寸进行采样, 生成候选的物体结构假设, 其平面与垂直墙壁平行, 底部与地板平面重合. 对于相机高度而言, 地板平面的大小范围通过水平线以及连接两个水平消失点的消失线界定, 可利用上述约束限制生成的物体结构假设数量. 最终通过上述方法在每幅图像中生成 100 个物体假设, 如图 6 中不同颜色的立方体所示.

本文将每个候选的物体结构假设建模为不同朝向平面的集合 $\bar{c} = \{f_i\}_{i=1}^{F}$, 其中 F 为平面的个数. 事实上, 物体结构假设的某些平面被遮挡, 为此定义平面可见性变量 $\bar{v} = \{v_i\}_{i=1}^{F}$, 其中 $v_i = 1$ 表示该平面可见, $v_i = 0$ 表示不可见. 由于物体中每个平面的投影畸变是已知的, 可在矫正畸变后在图像中提取每个平面的特征, 其定义为 $G = \{\bar{g}_i\}_{i=1}^{F}$. 这里主要统计每个平面的 HOG 直方图, 此外加入每个平面中通过 SLE 得到的平均物体标注置信度作为特征 $\bar{g}$.

本文使用线性函数 $s(f_i) = \bar{w}_i^t \bar{g}_i$ 对物体结构假设的每个平面独立评分: 其中 $\bar{w}_i$ 为线性 SVM 的权重向量. 为了应对物体三维尺寸的变化并获得更优

的物体定位, 利用最近邻物体结构假设上具有最高评分的平面 $f_j \in N(f_i)$ 实现对每个平面的评分进行改进, 并且结合 ORM 的输出结果, 得到物体结构假设评分 $\bar{c}$.

$$scr(\bar{c}) = w_1 \frac{\sum\limits_i v_i \max\limits_{f_i \in N(f_i)} \times s(f_i)}{\sum\limits_i v_i} + w_2 v(\bar{c}) \quad (6)$$

其中, w_1 和 w_2 为归一化权重, $v(\bar{c})$ 为 ORM 输出的物体置信度.

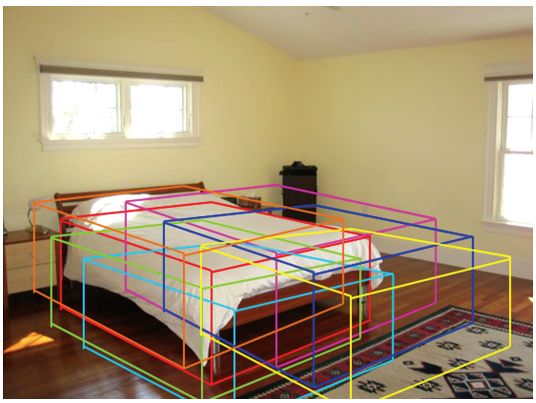


图 6　候选物体结构假设的生成

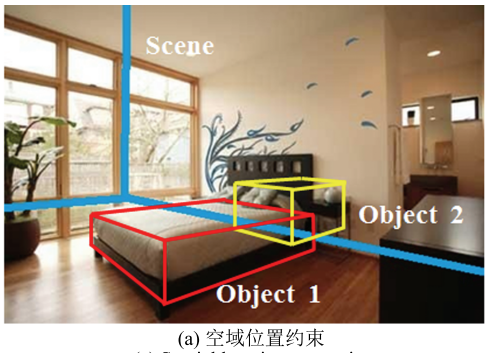
Fig. 6　Candidate object hypothesis generation

5　初始场景配置假设的生成

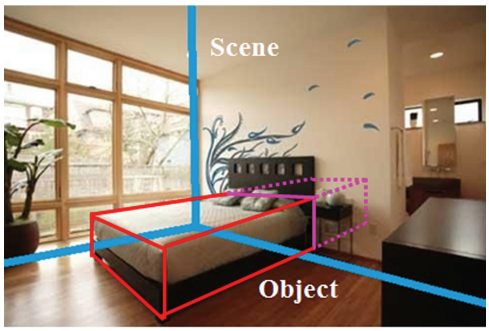
在获得房间和物体结构的初始假设后, 本文对室内场景的配置空间进行搜索, 选择与基于图像信息估计得到的局部场景几何最为匹配且最满足物理世界空域约束的配置. 为此, 采用了以下三种空域和语义相结合的场景配置约束条件, 如图 7 所示: 1) 空域排他性约束. 假设物体是彼此无法重合的固体, 那么不同物体占据的空间具有排他性, 即两个物体占据的空间不能相交; 2) 空域位置约束. 每个物体的所有部分必须处于房间之内, 不能位于墙壁之外; 3) 语义约束. 房间假设和物体假设均需要满足一定的置信度约束, 例如基于式 (2) 得到的房间假设置信度 $f(x_i, y)$ 或基于式 (5) 得到的物体假设置信度 $scr(\bar{c})$ 低于设定的阈值时, 将该场景配置假设丢弃.

5.1　场景配置假设的空域约束

评价房间假设和物体假设的组合是否满足以上三个场景配置约束是最关键的一个环节, 目的是丢弃部分不符合实际的场景配置. 在单幅图像中进行场景的三维结构推理困难的一个原因是无法获取房间的尺度信息. 为了测试 "房间 – 物体" 假设对以及 "物体 – 物体" 假设对的空域兼容性, 假设所有物体均位于地板之上, 其确定了房间和物体结构假设的尺度模糊程度并允许对它们的三维空域位置进行推理.



(a) 空域位置约束
(a) Spatial location constraint



(b) 空域排他性约束
(b) Spatial exclusion constraint

图 7　场景配置约束描述

Fig. 7　Scene configuration constraint

5.2　场景配置假设的推理

假设输入图像为 x, 房间结构假设集为 $\{r_1, r_2, \cdots, r_n\}$, 物体结构假设集为 $\{o_1, o_2, \cdots, o_n\}$, 目的是找到最优的场景配置 $y = \{y_r, y_o\}$, 其中, $y_r = \{y_r^1, y_r^2, \cdots, y_r^n\}$, $y_o = \{y_o^1, y_o^2, \cdots, y_o^n\}$. 当房间结构假设 r_i 在场景配置中被使用时, $y_r^i = 1$, 否则 $y_r^i = 0$; 当物体结构假设 o_i 在场景配置中被使用时, $y_o^i = 1$, 否则 $y_o^i = 0$. 当只有一个房间假设被用于定义场景配置时, $\sum_i y_r^i = 1$.

与房间结构假设的置信度估计方式类似, 通过如下最小化方式实现场景配置的最优估计 $y^* = \arg\max_y f(x, y, w)$. 本文将评分函数定义为: $f(x, y) = w^{\mathrm{T}} g(x, y) + w_\rho^{\mathrm{T}} l(y)$. 其中, $g(x, y)$ 为图像 x 中场景配置 y 对应的特征向量, $l(y)$ 用于对违反空域约束的房间和物体结构假设进行惩罚. 这里同样使用结构化 SVM 技术来对权重向量 w 进行学习

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$
$$\text{s.t.} \quad \xi_i \geq 0, \quad \forall i$$

$$w^{\mathrm{T}}g(x_i, y_i) - w^{\mathrm{T}}g(x_i, y) - w_\rho^{\mathrm{T}}l(y) \geq \delta(y_i, y) - \xi \tag{7}$$

其中, y_i 为人工标注的 Ground truth 场景配置, ξ_i 为松弛变量. 代价函数 $\delta(y_i, y)$ 定义了整幅图像中具有正确标注的像素比例. 例如, 被标注为左侧墙壁的像素实际属于正面墙壁或者被标注为物体的像素实际属于地板等情况均会被判断为错误标注的像素. 特征向量 $g(x, y)$ 度量了特征信息对场景配置 y 中各个平面的匹配程度. 这里首先通过 SLE 方法获取地板、左侧墙面、中间墙面、右侧墙面、天花板和物体六种平面类别的标注, 接着分别统计正确描述各个平面属性的像素比例, 并用一个 6 维的特征向量表示 $g(x, y)$. 式 (7) 中的惩罚项 $l(y) = l(y_r, y_o) + \sum_{i,j} l(y_o^i, y_o^j)$ 度量了空域约束被违反的程度. $l(y_r, y_o)$ 度量了房间的墙壁和物体之间的空域重合度, 惩罚了物体结构假设位于房间结构假设空间之外的配置, 与位于房间之外的体积大小成正比. $\sum_{i,j} l(y_o^i, y_o^j)$ 度量了两个物体 i 和 j 的空域重合度, 与两者投影到地板上的重合体积大小成正比.

通过求解式 (7) 寻找最优评分对应的场景配置 y^* 需要遍历全部可能的场景配置 $n \times 2^m$, 具有很高的计算复杂度. 目前, 组合优化技术在基于计算机视觉的物体识别等领域已经得到了广泛应用[22−24], 能够从大规模候选集中根据特定的需求实现高效采样. 本文采用经典的光束搜索法 (Beam search)[25] 以避免对全部场景配置进行评估. 光束搜索法的具体流程如下: 在搜索树的第一层中, 对仅具有一个房间结构假设 (无物体结构假设) 的场景配置进行评分. 在剩余的层中, 物体结构假设作为子节点被加入到基于场景配置的父节点中并对其进行评分. 那么, 具有最高评分的那个顶层节点将被加入到搜索树中作为子节点, 其中 d_l 即为第 l 层的光束宽度. 本文建立具有 $l = 4$ 层的搜索树, 每层的光束宽度为 $d_l = \{50, 5, 2, 1\}$, 光束搜索法将遍历所有的层或者直到没有与现有的场景配置相兼容的假设被加入为止. 最后, 搜索树中具有最优评分的节点即为求解得到的最优场景配置.

6 实验结果与分析

6.1 试验图像集

本文从 LabelMe 图像集[26] 中挑选了 308 幅室内图像, 其中 204 幅组成了训练集, 并人工标注了 Ground truth 立方体空域布局, 以及基于多边形边界的地板、墙面和天花板、平面几何描述和前景物体的位置等信息, 剩余的 104 幅组成了测试集.

6.2 场景空域布局推理实验

图 8 通过定性的方式给出了不同室内场景空域布局的评价结果. 其中, 各图第 1 列上面为原始图像, 下面为通过图像的几何信息得到的直线段提取结果; 第 2 列上面和下面分别为通过 SLE 以及 ORM 得到的物体位置估计; 第 3 列上面为仅通过图像几何信息得到的具有最高置信度的初始房间结构假设, 下面为本文提出的结合物体结构假设信息推理得到的最终房间结构假设结果, 其中的黄色立方体为估计得到的物体结构假设.

从图 8 可以看到, 当房间结构假设仅通过空域几何信息进行估计时, 基于置信度排序得到的结果容易导致不同程度的估计误差, 例如, 图 8 (a) 中没有找到两个相邻墙面之间正确的垂直分割线; 图 8 (f) 中由于床的存在使得两个相邻墙面底部的边界线距离实际的地板有较大的距离; 图 8 (j) 中同样由于桌椅的遮挡使得一侧墙面底部的边界线错误地定位在了桌椅与地板的交界线上. 当利用高层图像语义对房间中杂乱堆放物体的位置进行估计时, 可以看到两种不同的高层图像语义对于物体的定位具有各自的贡献, 例如在图 8 (e) 中通过 SLE 得到的物体位置要比 ORM 得到结果更为准确, 后者错误地将大片地板区域也判别为了物体, 而在图 8 (c) 中当背景相对简单时则是 ORM 取得了更为准确的物体定位结果, 而基于 SLE 得到的物体区域则错误地包含了部分墙壁. 通过将上述两种高层图像语义进行合理结合后, 不难发现本文算法估计得到的物体结构假设通常能够更为鲁棒地描述房间中实际的物体摆放位置以及它们的真实尺寸, 而上述物体结构假设同样对最终的房间结构假设的选择起到了进一步的优化作用, 例如, 图 8 (a)、图 8 (k) 和图 8 (m) 等, 在结合了物体位置和尺寸信息以后得到了更为接近实际描述的房间结构假设估计结果. 可见, 基于高层图像语义的物体先验和多元化空域约束对于房间结构假设推理的改进作用是显著的.

6.3 房间结构假设分析

为了对房间结构假设的结果进行定量评价, 将本文方法 (A4) 分别与三种经典的室内场景空域布局推理方法 (Hedau 等的方法 (A1)[2]、Lee 等的方法 (A2)[3] 和 Schwing 等的方法 (A3)[18]) 进行比较. 表 1 利用文献 [2] 中定义的像素误差 (Pixel error) 和角误差 (Corner error) 给出上述三种方法的定量评价结果. 其中, 像素误差为立方体各个平面上与 Ground truth 标注不同的像素百分比, 角误差为房间结构假设中各角所在位置与 Ground truth 标注之间的均方根 (Root mean square, RMS) 误差.

从表 1 可以看到, 本文方法在低层的图像几何信息基础上, 合理加权了多种高层图像语义特征, 取得了显著的改进. 其中, 与 A1 方法相比, 像素误差和角误差分别降低了 4.3 % 和 1.3 %, 与较新的方法

图 8　室内场景的空域布局推理结果

Fig. 8　Spatial layout estimation of indoor scenes

表 1　房间结构假设误差

Table 1　Room hypothesis error

方法 (OPP)	A1	A2	A3	A4
Pixel error	21.2	26.8	17.0	15.9
Corner error	6.3	11.4	5.5	5.0

A3 相比具有更低的误差, 进一步证明了本文方法的优势.

图 9 给出了上述三种房间结构假设估计方法之间的定性比较. 其中, 各图第 1 列为原始图像, 第 2 ~ 5 列分别为 A1、A2、A3 和 A4 方法得到的房间结构假设结果对比. 通过对比我们不难发现, A2 的结果最不稳定, A3 和 A4 的结果比 A1 更好一些. A3 和 A4 相比, 性能上较为相似, 例如第 4 行、第 8 行和第 12 行场景对应的结果. 不过在更多具有较强物体遮挡或者空间结构模糊的场景中, 例如第 2 行、第 7 行、第 9 行和第 10 行, 本文方法 A4 可以得到较为准确的房间空域结构描述, 而 A3 方法勾勒的立方体与真实的房间空域结构具有更大的偏差.

6.4　房间结构假设分析

为了对本文应用的两种高层图像语义在物体结构假设推理中起到的作用进行评价, 将 SLE 算法 (B1)、ORM 算法 (B2) 与本文提出的两者线性加权的方法 (B3) 进行比较. 图 10 给出了上述方法以像素误差和物体识别率 (Detection rate) 为度量的定量评价. 从图 10 中可以看到, 在像素误差方面, 尽管 B2 比 B1 具有更高的像素误差, 但是通过合理的线性加权, 本文方法 B3 取得了最低的像素误差, 与 B1 和 B2 相比分别下降了 4.1 % 和 13.5 %. 在物体识别率方面, B3 同样取得了最高的识别精度, 与 B1 和 B2 相比分别提高了 6.8 % 和 2.9 %, 进一步验证了本文线性加权方式的合理性.

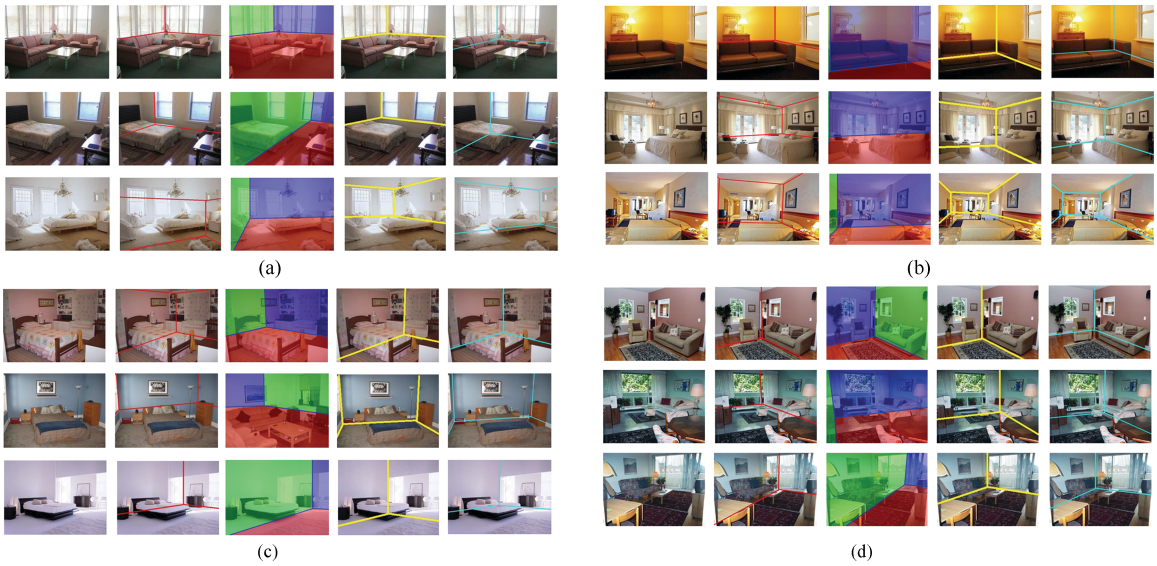


图 9 不同房间结构假设估计方法的比较

Fig. 9 Comparisons of different room hypothesis approaches

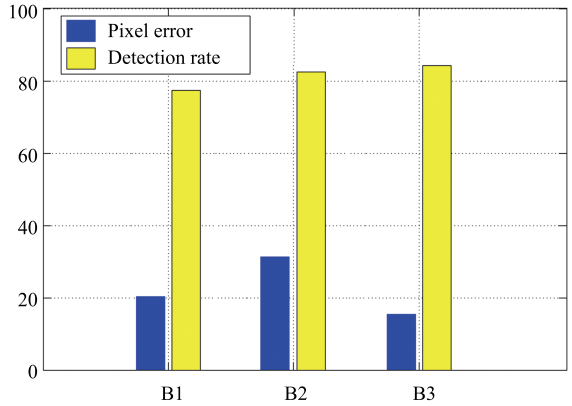


图 10 不同高层图像语义在物体结构假设中的像素误差和物体识别率

Fig. 10 The pixel error and object recognition rate of different high-level image semantics in object structure hypothesis

7 结论

本文提出一种简单快速的方法以实现对杂乱堆放了各种物体的室内场景的空域布局进行推理. 为了参数化地对房间和物体的三维体积进行描述, 提出在算法中加入空域排他性和空域位置等几何约束, 将多种高层图像语义加入到算法框架中, 改进房间和物体的结构假设估计, 最终通过基于组合优化的结构化学习策略实现快速的最优场景配置假设筛选. 实验证明, 与现有的多种经典方法相比, 本文算法在杂乱的室内场景中能够获得更为准确的房间和物体空域结构描述.

References

1 Coughlan J M, Yuille A L. Manhattan world: compass direction from a single image by Bayesian inference. In: Proceedings of the 7th IEEE International Conference on Computer Vision. Kerkyra, Greece: IEEE, 1999. 941−947

2 Hedau V, Hoiem D, Forsyth D. Recovering the spatial layout of cluttered rooms. In: Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009. 1849−1856

3 Lee D C, Hebert M, Kanade T. Geometric reasoning for single image structure recovery. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009. 2136−2143

4 Košecká J, Zhang W. Video compass. In: Proceedings of the 7th European Conference on Computer Vision. Copenhagen, Denmark: Springer, 2002. 476−490

5 Rother C. A new approach to vanishing point detection in architectural environments. *Image and Vision Computing*, 2002, **20**(9−10): 647−655

6 Barinova O, Konushin V, Yakubenko A, Lee K, Lim H, Konushin A. Fast automatic single-view 3-D reconstruction of urban scenes. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 100−113

7 Yu S X, Zhang H, Malik J. Inferring spatial layout from a single image via depth-ordered grouping. In: Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Anchorage, AK, USA: IEEE, 2008. 1−7

8 Nabbe B, Hoiem D, Efros A A A, Hebert M. Opportunistic use of vision to push back the path-planning horizon. In: Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. Beijing, China: IEEE, 2006. 2388−2393

9 Hoiem D, Efros A A, Hebert M. Recovering surface layout from an image. *International Journal of Computer Vision*, 2007, **75**(1): 151−172

10 Micusik B, Wildenauer H, Kosecka J. Detection and matching of rectilinear structures. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA, 2008. 1−7

11 Saxena A, Schulte J, Ng A Y. Depth estimation using monocular and stereo cues. In: Proceedings of the 20th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. 2197−2203

12 Liu B Y, Gould S, Koller D. Single image depth estimation from predicted semantic labels. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 1253−1260

13 Liu M M, Salzmann M, He X M. Discrete-continuous depth estimation from a single image. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014. 716−723

14 Gupta A, Efros A A, Hebert M. Blocks world revisited: image understanding using qualitative geometry and mechanics. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010. 482−496

15 Lee D C, Gupta A, Hebert M, Kanade T. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In: Proceedings of the 2010 Advances in Neural Information Processing Systems 23. Vancouver, British Columbia, Canada: Curran Associates, Inc., 2010. 1288−1296

16 Hedau V, Hoiem D, Forsyth D. Thinking inside the box: using appearance models and context based on room geometry. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010. 224−237

17 Wang H Y, Gould S, Koller D. Discriminative learning with latent variables for cluttered indoor scene understanding. In: Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: Springer, 2010. 497−510

18 Schwing A G, Fidler S, Pollefeys M, Urtasun R. Box in the box: joint 3D layout and object reasoning from single images. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, VIC, Australia: IEEE, 2013. 353 −360

19 Choi W, Chao Y W, Pantofaru C, Savarese S. Understanding indoor scenes using 3D geometric phrases. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013. 33−40

20 Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. *The Journal of Machine Learning Research*, 2005, **6**: 1453−1484

21 Li F X, Carreira J, Sminchisescu C. Object recognition as ranking holistic figure-ground hypotheses. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 1712− 1719

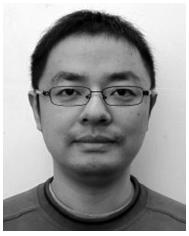
22 Lampert C H, Blaschko M B, Hofmann T. Efficient subwindow search: a branch and bound framework for object localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(12): 2129−2142

23 Russakovsky O, Ng A Y. A Steiner tree approach to efficient object detection. In: Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 1070−1077

24 Vijayanarasimhan S, Grauman K. Efficient region search for object detection. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2011. 1401−1408

25 Russell S, Norvig P. *Artificial Intelligence: A Modern Approach* (3rd edition). New Jersey: Pearson, 2009.

26 Russell B C, Torralba A, Murphy K P, Freeman W T. LabelMe: a database and web-based tool for image annotation. *International Journal of Computer Vision*, 2008, **77**(1−3): 157−173



姚拓中 宁波工程学院电信学院讲师. 2011 年获得浙江大学博士学位. 主要研究方向为计算机视觉, 机器学习. 本文通信作者. E-mail: thomasyao@zju.edu.cn

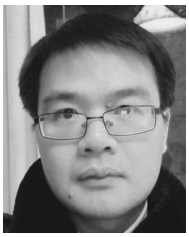
(**YAO Tuo-Zhong** Lecturer at the School of Electronic and Information Engineering, Ningbo University of Technology. He received his Ph. D. degree from Zhejiang University in 2011. His research interest covers computer vision and machine learning. Corresponding author of this paper.)

左文辉 浙江大学信息与电子工程学院博士研究生. 2007 年获得浙江大学学士学位. 主要研究方向为计算机视觉, 机器学习. E-mail: wenhuizuo@126.com

(**ZUO Wen-Hui** Ph. D. candidate at the College of Information Science and Electronic Engineering, Zhejiang University. He received his bachelor degree from Zhejiang University in 2007. His research interest covers computer vision and machine learning.)

宋加涛 宁波工程学院电信学院教授. 2003 年获得浙江大学博士学位. 主要研究方向为图像处理, 模式识别.
E-mail: sjt6612@163.com

(**SONG Jia-Tao** Professor at the School of Electronic and Information Engineering, Ningbo University of Technology. He received his Ph. D. degree from Zhejiang University in 2003. His research interest covers image processing and pattern recognition.)



应宏微 宁波工程学院电信学院讲师. 2004 年获得浙江工业大学硕士学位. 主要研究方向为图像处理, 视频压缩.
E-mail: yinghongwei@163.com

(**YING Hong-Wei** Lecturer at the School of Electronic and Information Engineering, Ningbo University of Technology. He received his master degree from Zhejiang University of Technology in 2004. His research interest covers image processing and video compressing.)