基于 LLOM 的单目图像深度图估计算法*

邓小玲[1,2], 倪江群[1†], 代 芬[1], 李 震[1]
(1. 华南农业大学 工程学院, 广州 510642; 2. 中山大学 信息科学与技术学院, 广州 510006)

摘 要: 针对计算机视觉理解单目图像立体结构的问题,进行了单目图像深度估计算法的研究。提出了一种基于监督学习方法的室外单目图像深度估计算法,其采用语义标注信息指导深度估计过程,融合绝对深度特征、相对深度特征以及位置特征作为深度特征向量,采用 LLOM 学习深度特征向量与深度值之间的关系。实验结果显示,该算法对路面、草地以及建筑物类等深度渐进变化的图像块,可获得较满意的深度估计结果。本算法为单目图像深度估计开辟了一个全新的有效途径。

关键词: 深度估计; 单目图像; 语义标注; 流形学习
中图分类号: TP391.41; TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2012)11-4357-03
doi:10.3969/j.issn.1001-3695.2012.11.092

Depth estimation algorithm from monocular image based on LLOM

DENG Xiao-ling[1,2], NI Jiang-qun[1†], DAI Fen[1], LI Zhen[1]
(1. *College of Engineering, South China Agriculture University, Guangzhou* 510642, *China*; 2. *School of Information Science & Technology, Sun Yat-sen University, Guangzhou* 510006, *China*)

Abstract: The key work of this paper was to estimate the scene depth for a single monocular image based on machine learning algorithm. Under the direction of semantic labels, an improved ASSOM algorithm named LLOM(locally liear online mapping) to depth estimation first time to learn the manifold of high dimension feature vectors, including absolute depth feature, relative depth feature and position feature. The experiments show that proposed method can obtain an acceptable result, especially for those blocks with gradual change of depth. This algorithm provides a new possibility to estimate depth from a single image.
Key words: depth estimation; monocular images; semantic label; manifold learning

当人类观看一幅图像时,可以毫不费劲地理解场景的 3 维结构。然而,对于当前的计算机视觉系统而言却存在着极大的挑战。要让计算机视觉具有类似人类的视觉感官能力,很明显需要计算机也具有类似人类的学习能力。因此,采用机器学习方法训练图像块特征信息与深度之间的关系,将是理解图像 3D 结构的可行方法。

目前,采用机器学习方法进行图像深度估计的研究报道相当少。文献[1]采用监督学习算法对无限制场景进行图像深度学习,该算法直接把深度值作为图像特征信息的函数,分别采用联合高斯 MRF(Markov random field)模型以及联合拉普拉斯 MRF 模型对给定图像深度的后验分布进行建模,根据图像块的深度特征信息估计图像块的深度值。该方法是目前唯一采用机器学习学习图像深度的典型方法。本文有别于文献[1]之处在于:a)对图像进行语义标注,把不同语义类别的图像块深度估计问题进行区别对待;b)在深度特征提取方法上,本文除了采用局部特征、全局特征、相对特征,还引入了图像块的位置特征;c)采用线性流形自组织子空间映射 LLOM[2] 方法对图像块特征与其深度的关系进行学习。

要想从单幅图像中估计深度信息,需要大量局部、邻域甚至全局的深度特征。这些高维数据对应着一个特定的 3D 深度结构,它们内部必然存在着特定的流形。学习非线性数据分布的一个可行办法是采用混合的局部线性模型。该办法在神经网络领域中被普遍采用[3~5]。自适应子空间自组织映射(ASSOM)通过竞争和协作学习过程,可以学习到从高维数据分布到一系列拓扑组织低维线性子空间的映射,这个过程被认为是数据抽取过程。ASSOM 相当于一个具有在线学习能力的主成分分析混合模型,该方法已被成功应用于语音和图像处理等领域。因此,本文针对非结构性室外环境图像,首次采用 ASSOM 的改进算法 LLOM 学习深度特征空间的局部线性流形。

1 深度特征信息的提取

从单目二维图像中恢复三维结构,主要依据单目提示信息。单目提示信息有很多种类,如纹理变化、纹理递变度、已知对象大小、光线和阴影、雾气、散焦度等。估计图像 3D 结构仅靠这些局部图像特征信息是不够的。因此,本文通过提取三种特征作为深度特征向量,即绝对深度特征、相对深度特征以及位置特征,以下分别介绍各特征的提取方法。

收稿日期: 2012-04-07; **修回日期**: 2012-05-18 **基金项目**: 广东省科技计划资助项目(2011B020308009);国家自然科学基金资助项目(31101077)

作者简介: 邓小玲(1978-),女,广东饶平人,讲师,博士,主要研究方向为多媒体信息处理;倪江群(1963-),男(通信作者),江西无锡人,教授,博士,主要研究方向为多媒体信息处理和通信、信息隐藏和数字取证(jqni_sysu@126.com);代芬(1978-),女,湖北天门人,副教授,博士,主要研究方向为电子与计算机技术应用;李震(1981-),男,吉林大安人,副教授,博士,主要研究方向为电子与计算机技术应用.

1.1 绝对深度特征

本文主要获取纹理变化、纹理递变度以及颜色三种信息作为绝对深度特征。通过在亮度通道上采用 Laws' masks[6] 去计算纹理的能量,从而获得纹理信息;在颜色通道上应用一个局部的平滑滤波来获得图像的雾气浓厚度;其纹理递变度通过在亮度通道与六个方向的边缘滤波器进行卷积获得。各卷积滤波器如图 1 所示。其中前九个是 3×3 的 Laws' masks,用于局部平滑、边缘检测和孔径检测;后六个是边缘导向检测器,每个方向间隔是 30°。

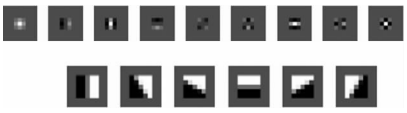

图1 计算纹理能量和纹理梯度的卷积滤波器

因此,对于图像 $I(x,y)$ 中每一个图像块 i,计算其绝对深度特征的方法如下:采用 17 个滤波器 $F_n(n=1,\cdots,17)$,其中 9 个 Laws' masks、6 个纹理梯度和 2 个颜色通道。这些滤波器分别与图像块进行卷积,计算其输出 $E_i(n)=\sum_{(x,y)\in \text{patch}(i)}|I*F_n|^k$,取 $k\in\{1,2\}$ 分别表示绝对值能量和及平方值能量和,得到一个初始维数为 34 的初始特征向量。

要估计一个图像块的绝对深度,仅靠图像块的局部特征是不够的,必须考虑更多的全局特性。因此本文尝试捕捉不同尺度空间(图像分辨率)以及同一列相邻图像块的信息,如图 2 所示。对于一个特定的图像块,需要统计的绝度深度特征包括 3 个空间尺度下其本身和 4 个邻域块的特征向量(3×5=15 个),以及同一列 4 个图像块的特征向量,因此一共有 19×34=646 维的绝对特征向量。

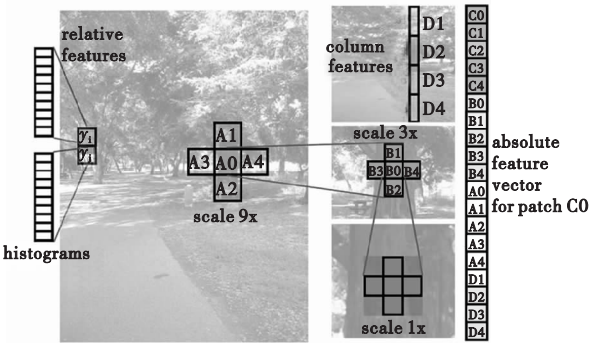

图2 绝对深度特征向量示意图

1.2 相对深度特征

相邻图像块之间存在着相依性,相对的特征向量可以表示这种相依性[1]。为此,计算 17 个滤波器输出 $|I*F_n|$ 的 10-bin 直方图,这样就可以得到共 170 个在尺度 s 下特定图像块 i 的相对特征 y_{is}。这些特征用于估计两个不同位置图像块的相互作用。两个相邻图像块 i 和 j 的相对特征向量是两个直方图的差值,即 $y_{ijs}=y_{is}-y_{js}$。

1.3 位置深度特征

由于涉及的图像都是摄像机在水平方向上拍摄的,那么图像上不同行具有的统计特性也不太相同。例如:一个蓝色的图像块出现在图像的上面,一般会认为这是天空的一部分区域;但如果图像的下方也出现一块蓝色区域,那很可能是水面的一部分。要让计算机视觉具有人类的这种先验知识,则需把图像块的位置也考虑进去,因此,本文把图像块中心点的位置(x,y)作为位置深度特征。

综上所述,用于训练学习的特征包含了绝对深度特征以及位置特征,不同语义类别的图像块根据深度值,将对应特征向量送往 LLOM 子空间映射网络进行训练学习;而相对特征不经学习训练过程,其使用目的在于为测试过程提供判断依据,以提高图像深度估计的精度。

2 基于 LLOM 的单目图像深度图估计

2.1 LLOM 算法简介[2]

LLOM 网络是基于 SOM 的一种混合模型,该网络由一个均值和基向量集描述一个局部线性流形。均值可看做学习数据的中心值,而基向量集则张成了一个局部的主元子空间,它表达了最重要的变化方向。

LLOM 学习算法步骤总结如下:

a)对每一个网络中的神经元,初始化其均值向量 m 以及基向量 $b_{qh}(h=1,2,\cdots,H)$,并使各个基向量彼此正交。

b)对于 t 时刻的输入短序列 $X^{(t)}$,根据式(1)计算获胜神经元。

$$c=\arg\min_{q\in Q} e(X^{(t)},L_q) \tag{1}$$

c)按照式(2)和(3)更新每个神经元 q 的均值 m 和基向量 b_{qh},然后使基向量正交化。

$$m_q^{(t+1)}=m_q^{(t)}+v^{(t)}(c,q)\lambda(t)\times\left(\sum_{s\in S(t)}\tilde{x}_{L_q}^{(t)}(s)+\alpha\sum_{s\in S(t)}\hat{x}_{L_q}^{(t)}(s)\right) \tag{2}$$

$$b_{qh}^{(t+1)}=b_{qh}^{(t)}+(1-\alpha)v^{(t)}(c,q)\lambda(t)\times\sum_{s\in S(t)}((x_{L_q}^{(t)}(s)-m_q^{(t)})^{\mathrm{T}}b_{qh}^{(t)})\tilde{x}_{L_q}^{(t)}(s) \tag{3}$$

其中:m_q 是第 q 个神经元的均值向量;b_{qh} 是第 q 个神经元的第 h 个基向量,$h=1,2,\cdots,H$。

d)进入下一个迭代步骤,重复步骤 b)c),直到满足特定的终止准则为止,如满足预先设定的学习步骤次数。

2.2 基于 LLOM 算法的图像深度估计

在文献[1]的研究中,图像块的深度估计完全忽略了图像场景的语义背景,把所有的图像块一视同仁,把场景深度作为图像特征的函数,采用 MRF 对其进行建模。在本文算法中,将采用语义信息指导深度信息的提取。采用语义信息有着以下优点:a)可以采用较简单的图像特征信息表征一种场景,避免了所有场景的共同特征信息;b)通过对图像进行语义标注,可以更容易判断每一个语义类的共面性和连接性。最后,图像中具有相同深度但不同语义类别的图像块千差万别,采用语义标注可更准确地建立深度与特征的对应关系。

对于图像的语义标注,本文不将其列入研究范围,可以采用任何一个多类的语义标注方法。文献[7~9]在基于语义的图像分割上都取得了很好的效果;文献[10]中,作者在获得较好语义标注的前提下,采用了 MRF 模型完成图像深度估计。

经过深度特征提取步骤后,便得到了不同语义类别各自深度值对应的训练特征集,将这些特征集送给各自的 LLOM 网络进行训练学习,便可得到各个语义类别图像的各个深度值所映射的线性子空间网络。特征集中每个深度对应的特征向量具有较高的维数,如 7×6 大小的图像块,所提取出的深度特征向

量维数为 648 维。将这些高维的特征向量映射到神经元数目为 $W \times W(W=3,\cdots,8)$ 的 LLOM 网络中进行自主学习。实验中,可以对神经元的数目以及维数进行调整测试,已达到最佳的映射。神经元的维数可以为 $H=1,2,3$。例如,如果选择 $W=4,H=2$,那么经训练后,648 维的特征向量就会被映射到 $4\times4\times3$的 LLOM 网络中(每个神经元有一个均值向量和两个正交基向量表示),从而达到高维数据的降维处理。另外,这种映射关系的结果也揭示了高维深度特征向量中隐藏的内在本质特征。

3　实验结果

本文的实验数据来源于斯坦福大学学者 Saxena 采用 3D 扫描设备收集的图像以及对应的参考深度图(ground-truth depths)(http://make3d. stanford. edu),这些图像的深度范围是 1.2 ~81.89 m。为了降低计算复杂度,本文基于 LLOM 的深度识别系统仅包括 80 个 LLOM 网络。其中:被标注为路面(road)的图像块对应 $N_k(k=0,\cdots,16)$ 网络;被标注为草地(grass)的图像块对应 $N_k(k=17,\cdots,41)$ 的网络;被标注为树木(tree)的图像块对应 $N_k(k=42,\cdots,58)$ 的网络;被标注为建筑物(building)的图像块对应 $N_k(k=59,\cdots,80)$ 的网络;而被标注为天空(sky)的图像直接得出深度值为 81.89。

图 3 为深度估计值与其参考深度之间的匹配程度,即估计精度。由于被标注为天空(sky)的图像块深度直接赋值为 81.89,因此其深度估计的精度由语义标注算法来决定。从图 3 中可以看出,深度估计精度不是特别理想,主要原因在于本文将每一类别的语义图像划分在 20 个深度级别以内。这种划分注定了估计结果只是一个近似值,而图 3 的实验数据却是估计值与参考深度值的精确匹配,因此,导致实验结果欠佳。

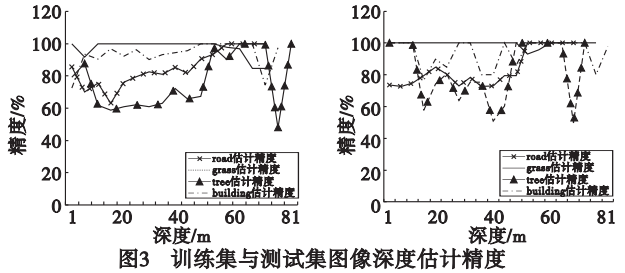

图3　训练集与测试集图像深度估计精度

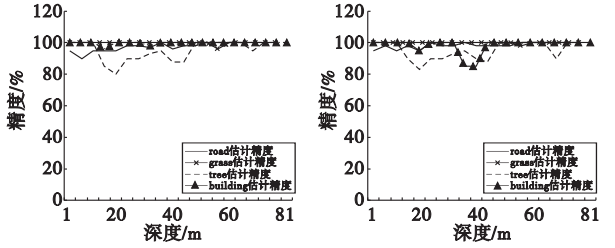

图4　容许0.5 m误差的估计精度

为了更客观地评价本文算法,采用近似匹配策略衡量算法性能,即容许估计值与参考深度值有一定的误差存在。在这种匹配标准的前提下,本文算法估计的深度与参考深度的匹配精度明显提高许多。图 4 为容许 0.5 m 误差的结果。

4　结束语

基于 LLOM 的图像深度估计算法作为一个尝试,为图像深度图估计提供了一种可行的发展空间。实验结果显示,该算法对路面、草地以及建筑物类等这些深度渐进变化的图像块来说,能获得较满意的估计结果。对于深度突变性较大的语义类别图像,如树木等,则需要进一步研究更有效的特征提取方法,这将是本文下一步研究的目标。

参考文献:

[1] SAXENA A, SUN M, ANDREW Y N. Make3D: learning 3D scene structure from a single still image[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2009,30(5):824-840.

[2] ZHENG Hai-cheng, SHEN Wei, DAI Qiong-hai,*et al*. Learning nonlinear manifolds based on mixtures of localized linear manifolds under a self-organizing framework[J]. Neurocomputing, 2009, 72(13-15):3318-3330.

[3] 郑慧诚,沈伟. 一种局部化的线性流形自组织映射[J]. 自动化学报,2008,34(10):1298-1304.

[4] EBRAHIMPOUR R, KABIR E, ESTEKY H, *et al*. View independent face recognition with mixture of experts[J]. Neurocomputing, 2008,71(4-6):1103-1107.

[5] XING Hong-jie,HU Bao-gang. An adaptive fuzzy C-means clustering-based mixtures of experts model for unlabeled data classification[J]. Neurocomputing,2008,71(4-6):1008-1021.

[6] GE Liang,ZHU Qing-sheng,FU Si-si. Application of laws' masks to stereo matching[J]. Acta Optica Sinica,2009,29(9):2509-2511.

[7] CARNEIRO G, CHAN A B, MORENO P J,*et al*. Supervised learning of semantic classes for image annotation and retrieval[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(3):394-410.

[8] LI Jia, WANG J. Automatic linguistic indexing of pictures by a statistical modeling approach[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2003,25(9):1075-1088.

[9] 李志欣,施智平. 融合语义主题的图像自动标注[J]. 软件学报,2011,22(4):801-812.

[10] LIU B, GOULD S, KOLLER D. Single image depth estimation from predicted semantic labels[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2010:1253-1260.

(上接第 4349 页)

[9] 李景義,王树宗. UPF 算法及其在目标跟踪问题中的应用[J]. 系统仿真学报,2007,19(3):675-677.

[10] GORDON N J, SALMOND D J, SMITH A F M. Novel approach to nonlinear/non-Gaussian Bayesian state estimation[J]. IEEE Proceedings on Radar and Signal Processing,1993,140(2):107-113.

[11] 王华剑,景占荣,羊彦. 基于改进扩展卡尔曼粒子滤波的目标跟踪算法[J]. 计算机应用研究,2011,28(5):1634-1636,1643.

[12] SARKKA S,VEHTARI A,LAMPINEN J. Rao-Blackwellized particle filter for multiple target tracking[J]. Information Fusion, 2007, 8(1):2-15.

[13] 王相海,方玲玲,丛志环. 基于 MSPF 的实时监控多目标跟踪算法研究[J]. 自动化学报,2012,38(1):139-144.

[14] HORN B K P. Robot vision[M]. [S. l.]: MIT Press,1986:66-69,299-333.

踪实验。其特点分别为:运动目标个数较少、多个目标存在交互运动、较远视角下的多个运动目标。实验是基于 MATLAB R2009b 平台上,硬件环境是 Intel Core i7,CPU 2.67 GHz,4 GB 内存。实验所采用的视频是使用固定在三脚架上的摄像机进行拍摄的若干个视频序列,帧率为 25 fps,分辨率为 320 × 240。

实验 1　本段视频包含了单个及两个目标的运动,即骑自行车的目标从左边驶入场景,随后另一个目标从左步行出现;骑自行车的目标从场景右边驶出,一辆白色的汽车从右边驶入。此段视频包含的运动较为简单,目标个数少,没有出现目标间的运动交汇。视频格式为 avi,共有 1 170 帧。跟踪实验结果如图 3 所示,目标用黑色方框来表示。其中(a)为混合高斯模型检测出的背景结果;(b) ~ (d)分别表示第 34、108、228 帧(这三帧基本表示了此段视频中所包含目标运动的情况)的实验跟踪结果。

实验 2　和实验 1 是同一场景下拍摄的视频,此段视频包含了多个运动目标,即三辆汽车的交互运动,以及行人的运动。视频格式为 avi,共有 578 帧。跟踪实验结果如图 4 所示,目标用黑色方框来表示。其中(a)为混合高斯模型检测出的背景结果;(b) ~ (d)分别表示第 205、222、242 帧(这三帧基本表示了此段视频中所包含目标运动的情况:两辆汽车从左向右行驶;当黑色汽车驶出场景时,另一辆白色汽车从右驶入场景;两辆汽车平行交互行驶;场景下方的行人运动)的实验跟踪结果。

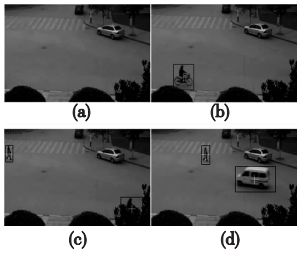

图3　视频1中目标跟踪结果

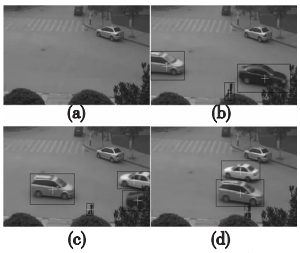

图4　视频2中目标跟踪结果

实验 3　与实验 1、2 的场景不同,远视角场景包含了更多的目标,并且具有目标轮廓较小、位置相对分散和疏密不均的特点。视频格式为 avi,共有 765 帧。跟踪实验结果如图 5 所示(包括三辆汽车和多个行人的运动),目标用黑色方框来表示。其中(a)为混合高斯模型检测出的背景结果;(b) ~ (f)分别表示第 67、335、502、696、762 帧(一辆黄色汽车从左边驶入,从右边驶出;随后一辆黑色汽车从右驶入,从左驶出;接着黑色汽车从左向右行驶;多个行人不规则的运动)的实验跟踪结果。

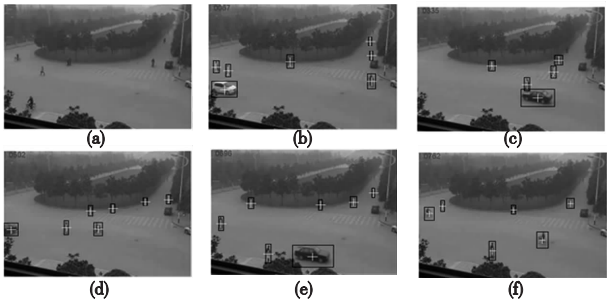

图5　视频3中目标跟踪结果

表 1 记录了处理不同视频段时所需要的运算时间。从实验结果来看,对于处理单一运动目标和多个交互运动目标场景下的目标跟踪问题,本文提出的算法具有一定的实时性。

表 1　实验所需的时间

视频段	视频帧数	每帧处理时间/ms
1	1 170	124
2	578	111
3	765	118

5　结束语

本文提出一种基于 GMM-RBMCDA 的实时监控多目标跟踪方法。该方法利用混合高斯模型作为背景建模方法,并使用形态学方法分割提取出前景目标;在多目标跟踪过程中使用了 Rao-Blackwellized 蒙特卡洛数据关联(RBMCDA)算法。此方法可以处理杂波条件下的跟踪问题,能够有效减少可能的目标交叉及杂波干扰带来的影响,且不需要事先给定目标个数,通过设置目标的存在和消失参数,实现了多个运动目标的自动跟踪。实验结果表明,对于实际场景中所跟踪目标个数无法实现给定的情况,该方法能够较为准确地计算出目标的个数,并且能够同时跟踪多个目标的运动状态,并具有较好的实时性。本文算法存在的一个局限是当场景视角广、目标较小或目标个数增多时,并不能全部给出满意的跟踪效果。另外,当目标之间的距离很近或存在交叉遮挡(或重叠)时,容易误认为是同一个目标。因此,在今后的研究中,还需要借助更多的目标特征,并综合利用模式识别的方法和理论增强前期目标检测预处理的准确性,为后续目标状态估计和跟踪提供更加可靠的保障。

参考文献:

[1] WREN C, AZARBAYEJANI A, DARRELL T, *et al*. Pfinder: real-time tracking of the human body[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,1997,19(7):780-785.

[2] STAUFFER C,GRIMSON W E L. Adaptive background mixture models for real-time tracking[C]//Proc of the IEEE International Computer Society Conference on Computer Vision and Pattern Recognition. [S. l.]:IEEE,1999:246-252.

[3] STAUFFER C,GRIMSON W E L. Learning patterns of activity using real-time tracking[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2000,22(8):747-757.

[4] 何信华,赵龙. 基于改进高斯混合模型的实时运动目标检测与跟踪[J]. 计算机应用研究,2010,27(12):4768-4771.

[5] PETERFREUND N. Robust tracking with spatio-velocity snakes Kalman filtering approach[C]//Proc of the 6th International Conference on Computer Vision. Washington DC:IEEE Computer Society,1998:433-439.

[6] ISARD M, BLAKE A. Contour tracking by stochastic propagation of conditional density[C]//Proc of the Europe Conference on Computer Vision. Berlin:Springer-Verlag,1996:343-356.

[7] ISARD M, MACCORMICK J. BraMBLe: a Bayesian multiple-blob tracker[C]//Proc of the IEEE International Conference on Computer Vision. [S. l.]:IEEE,2001:34-41.

[8] JEPSON A D, FLEET D J, EI-MARAGHI T F. Robust online appearance models for visual tracking[J]. IEEE Trans on Pattern Analysis and Machine Intelligence,2003,25(10):1296-1311.

(下转第 4359 页)