论文

从单幅图像学习场景深度信息固有的歧义性

何雷^①, 董秋雷^{①②③*}, 胡占义^{①②③}^① 中国科学院自动化研究所模式识别国家重点实验室, 北京 100190^② 中国科学院大学, 北京 100049^③ 中国科学院脑科学与智能技术卓越创新中心, 北京 100190

* 通信作者. E-mail: qldong@nlpr.ia.ac.cn

收稿日期: 2016-01-06; 接受日期: 2016-03-22

中国科学院战略先导专项 (批准号: XDB02070002)、国家自然科学基金 (批准号: 61333015, 61421004) 和北京市自然科学基金 (批准号: 7142152) 资助项目

摘要 从单幅图像学习场景深度信息是目前机器学习的一个重要问题, 这类方法的原理依据是人类视觉系统可以从单幅图像感知深度. 人类单眼感知深度是一种长期进化的能力, 其对应的“脑加工机理”仍远未清楚, 所以将这种“原理”直接作为“计算依据”的合理性似乎并不充分. 事实上, 三维场景到二维图像的成像过程, 存在严格的几何映射关系. 当没有任何先验信息时, 场景深度与相机焦距存在固有的歧义性. 所以, 本文认为目前文献中从单幅图像学习场景深度信息这类方法, 由于没有考虑焦距的因素, 在原理上存在固有的缺陷. 本文通过对实际图像进行测试, 证实了这种歧义性的存在. 从单幅图像学习深度信息, 应该考虑相机内参数的影响, 至少是相机有效焦距的影响.

关键词 单幅图像 场景深度复原 成像几何 相机焦距 歧义性

1 引言

在过去的几年中, 从单幅图像学习场景深度信息有很多报道, 这类方法的原理依据是人类视觉系统可以从单幅图像感知深度. Hoiem 等^[1] 假设图像中的物体由许多平面组成, 从单幅图像重建出了具有可信度的室外场景. Hedau 等^[2] 将这种简单的几何假设推广到估计房间的空间布局, 取得了很好的效果. Schwing 和 Urtasun^[3] 也证实了这种几何假设是一种有效的估计房间空间布局的策略. Saxena 等^[4,5] 使用条件随机场学习单幅图像的像素级深度. 而后 Saxena 等^[6] 又对前期工作进行了扩展, 假设图像由许多超像素组成, 搭建了单幅图像三维生成系统 Make3D. Liu 等^[7] 引入连续和离散变量表征超像素的深度和位置关系, 将估计深度的问题转化为高阶 Markov 随机场的优化问题. 为了降低运算的复杂度, Baig 等^[8] 提出了轻量级的深度恢复方法 “im2depth”, 可运行在手机平台上.

单幅图像的深度估计和法向量的估计是紧密相关的, 利用学习到的表面法向量信息, 可以更好地估计深度. Fouhey 等^[9] 提出具有视觉可区分和几何信息可表达的基元, 用于推断单幅图像的表面法向量. 在此基础上, Fouhey 等^[10] 提出了中级约束凸、凹边缘线索, 用于表面法向量的学习. Ladicky

引用格式: 何雷, 董秋雷, 胡占义. 从单幅图像学习场景深度信息固有的歧义性. 中国科学: 信息科学, 2016, 46: 811–818, doi: 10.1360/N112015-00285

等^[11]提出使用结合上下文和基于分割的线索, 构建 Boost 框架, 将学习表面法向量的问题转化成局部编码系数的回归问题. Häne 等^[12]提出了一种将表面法向量的分类器运用到估计单幅图像深度的方法, 并取得了很好的结果.

学习单幅图像的深度, 除了参数的方法外, 还有很多非参数的方法. Karsch 等^[13,14]使用 Liu 等^[15]的稠密特征流方法对相似图像的深度进行迁移, 然后使用 Markov 随机场求解最优深度, Konrad 等^[16]也提出了类似的方法.

除了以上传统的方法外, 卷积神经网络也已成功地应用于单幅图像的深度估计. Eigen 等^[17]首先提出堆叠两个神经网络来学习深度的方法, 结合了全局和局部信息. 在此基础上, Eigen 和 Fergus^[18]又提出使用多尺度的卷积神经网络来学习深度, 并获得了较高的准确率. Wang 等^[19]提出一种新的卷积网络结构, 用来估计单幅图像的表面法向量. Liu 等^[20,21]提出了一种基于全卷积网络和超像素池方法, 在统一的卷积神经网络框架下, 充分利用了深度卷积网络和条件随机场的优势. Li 等^[22]使用深度神经网络的特征输出来估计单幅图像的深度, 并用条件随机场进行调优.

学习单幅图像深度的数据集一般是通过 Kinect 或激光扫描得到的, 对应的二维图像事实上是在同一相机、固定相机参数下采集的, 如 NYU dataset^[23], Make3D dataset^[6]和 KITTI dataset^[24]. 同一相机意味着不同图像对应的像素物理尺寸是相同的, 固定相机参数说明所有图像对应的相机焦距是相同的. 这样, 在这种数据库下学习到的模型, 理论上无法扩展到其他相机下获得的“单幅图像”. 既使文献中有关于“对其他相机单幅图像深度信息学习的推广能力”的报道, 如在 Make3D^[6]中有对从 Internet 下载的单幅图像深度信息推断能力的报道, 本文认为这种“深度学习的推广能力”也缺乏必然性和应有的理论基础, 特别是缺乏“成像几何”的基础, 同理, Kendall 等^[25]学习摄像机外参数的方法也存在这个问题. 下节将指出, 从单幅数字图像学习场景深度信息存在固有的歧义性, 这种歧义性导致从单幅图像学习深度信息时, 如果不考虑焦距等相机内参数, 是不可能准确学习到深度信息的.

2 从单幅图像学习场景深度信息固有的歧义性

首先, 这里的场景深度是指物体与相机光心之间的垂直距离. 下面将分析指出, 相机的焦距与物体深度在图像上存在固有的不可区分性. 同时, 由于数字图像是对“模拟图像离散的结果”, 离散步长与焦距在数字图像上也无法区分, 所以, 在成像参数未知时, 仅仅“相机的有效焦距”才有意义. “有效焦距”是一个无量纲的量, 表示“焦距为多少个像素尺寸”, 所以, 此时的深度也仅仅表示有多少个“像素尺寸”的距离, 而不是通常意义下以米或厘米等物理量纲单位下的距离. 下面将对此进行分析.

2.1 深度与焦距在图像上的不可区分性

理论上来说, 长焦距下相机距离物体远时所成的像, 可以与短焦距下相机在距离物体近时对同一物体所成的像完全相同, 即深度与焦距在图像上的不可区分性. 为了表示方便, 假定相机的成像模型符合针孔模型 (the pinhole model), 且不失一般性, 为了表示清晰, 假定空间物体为线状物体. 如图 1 所示, 空间线段 S 在 (f_1, O_1) 下所成的像为 I_1 , 在 (f_2, O_2) 下所成的像为 I_2 , $I_1 = I_2$. 所以, 当不知道相机焦距时, 由于 $I_1(I_2)$ 对应的深度不唯一, 所以理论上从 $I_1(I_2)$ 无法得到唯一的深度 D .

2.2 相机焦距和数字图像的有效焦距

数字图像是从模拟图像离散得到的, 每个像素具有一定的物理尺寸. 给定一幅数字图像, 一般并不

图 1 (网络版彩图) 深度与焦距在图像上的不可区分性
Figure 1 (Color online) Indistinguishability between the scene depth and the focal length in images

图 2 (网络版彩图) 相机焦距和数字图像的有效焦距
Figure 2 (Color online) Focal length of a camera and valid focal length of a digital image

知道该图像对应的像素物理尺寸, 此时讨论该图像对应的相机物理焦距也已失去意义. 在这种情况下, 人们只能以“有效焦距”为对象进行讨论. “有效焦距”是指“物理焦距与像素物理尺寸的比值”, 本质上是一个无量纲的量. 如图 2 所示, 空间线段 S 在 (f_1, O) 下所成的像为 I_1 , 在 (f_2, O) 下所成的像为 I_2 . 尽管此时 $I_1 \neq I_2$, 但如果对 I_1 的离散步长是对 I_2 离散步长的 f_1/f_2 倍, 则 I_1 对应的离散图像与 I_2 对应的离散图像应该完全相同 (假定不存在离散误差). 图 2 说明, 仅仅知道离散图像, 有效焦距是唯一一个有效的量, 对应相机的具体物理焦距已失去意义. 所以, 在利用机器学习方法进行场景深度信息学习时, 当相机物理焦距和数字图像像素物理尺寸未知时, 学习到的场景深度不可能是物理深度, 而只能是物理深度与像素物理大小的比, 即像素单位下的深度. 当然, 假如数据库为同一相机且在固定焦距下采集的图像, 如 NYU dataset, Make3D dataset 和 KITTI dataset 数据库, 则此时在模型学习时可以直接转换成特定物理量下的深度.

3 歧义性测试实验

3.1 实验

从第 2 节的讨论知, 从单幅图像学习场景深度信息存在深度与焦距之间的固有的歧义性. 下面对这个问题进行了实验测试, 证实了本文结论的正确性.

在实验室场景中, 采集测试数据集 250 组, 每组为同一相机在不同焦距下的两幅图像: 在 50 mm 的焦距和 105 mm 的焦距下, 通过改变相机与场景之间的深度, 使其具有近似相同的视场范围, 每组两幅图像的实际深度相差至少都在 1 m 以上.

目前, 文献中 Liu 等^[20,21] 和 Eigen 等^[17,18] 学习单幅图像深度的方法表现优异, 为了验证深度与焦距之间的这种固有的歧义性, 本文利用这两种方法对引用文献中的 NYU dataset 进行训练, 得到模型后对采集的 250 组室内图像进行测试, 部分实验结果如图 3 所示. 图 3 中从左到右的焦距依次是 105 mm 和 50 mm, 第 1 行是 RGB 图像, 第 2 行是 Liu 等^[20,21] 方法的深度估计结果, 第 3 行是 Eigen 等^[17,18] 方法的深度估计结果, 以下各行以此类推. 此外, 蓝色表示近距离, 红色表示远距离.

本文使用人机交互的方式对每组不同焦距下的图像的深度进行了度量. 图 4 为其中一组图像使用 Liu 等^[20,21] 方法计算的结果, 左图焦距为 105 mm, 右图焦距为 50 mm. 选取固定物体的匹配区域, 读取其深度值. 匹配区域的平均深度值如表 1 所示. 通过计算发现, 学习平均深度相差 0.07506m, 而实际这两幅图像的深度相差 2 m, 显然学习出来的深度值不正确. 通过这种度量方式, 使用 Liu 等^[20,21] 方法和 Eigen 等^[17,18] 方法对 250 组图像进行了测试, 实验结果如表 2 所示, 对应的深度估

图 3 部分不同焦距的图像深度估计结果

Figure 3 Some results of depth estimation from single images with different focal lengths

计的误差率分别为 89.76% 和 87.2%。实验结果说明, 基于单幅图像的场景深度学习方法存在固有的深度和焦距之间的歧义性。本文认为不考虑图像有效焦距下的单幅图像的深度学习方法缺乏严谨的理论支持。

这里需要指出的是, 尽管从单幅图像学习场景深度存在深度与焦距之间的这种固有的歧义性, 为什么文献中仍会有好的结果的报道呢? 这主要是因为文献 [4~7, 14, 17~21] 中引用的数据集, 是同一相机在固定焦距下拍摄的, 此时相机物理焦距和像素离散步长都不变, 不存在焦距与深度之间的这种歧义性。否则, 这种理论上固有的歧义性在单幅图像的场景深度信息学习中是无法避免的。

3.2 讨论

本文中, “相机内参数对单幅图像学习深度有影响” 指的是如不考虑相机内参数 (特别是有效焦距), 则会出现以下歧义情况: 理论上来说, 长焦距下相机距离物体远时所成的像, 可以与短焦距下相

图 4 人机交互判断学习到的深度的正确性

Figure 4 Evaluation on depth estimation accuracy via human-computer interaction

表 1 Liu 等^[20,21]方法估计匹配区域的深度值(米)

Table 1 Liu et al. [20,21] methods estimate the depths of the matching areas (m)

Estimated mean depth in the left area	Estimated mean depth in the right area	Difference of the mean depths
2.12731	2.20237	0.07506

表 2 250 组图像深度估计的统计结果

Table 2 The statistical results of the depth estimation from 250 pairs of images

Methods	Number of testing image pairs	Number of false estimations	Error rate
Liu et al. [20,21]	250	224	89.6%
Eigen et al. [17,18]	250	218	87.2%

机在距离物体近时对同一物体所成的像完全相同,即深度与焦距在图像上的不可区分性.如果不考虑焦距信息,那么图像和深度具有一对多的对应关系,造成学习深度的不准确,即歧义性.如图 1 所示,图像 I_1 和 I_2 是一样的,当不知道两幅图像的焦距信息 f_1 和 f_2 时,单幅图像对应的深度可能是 D_1 ,也可能是 D_2 ,造成歧义性.

考虑到深度与焦距在图像上的不可区分性,以及相机焦距和数字图像的有效焦距,本文建议在模型学习阶段,需要同时将相机的内参数,特别是有效焦距,作为输入量参与学习.由于文献 [4~7,14,17~21] 的数据集 (NYU dataset, Make3D dataset) 都是定焦的,此类数据无法反映本文所述的歧义性.因此,下一步工作计划搭建含不同焦距的 RGB-D 数据,联合使用 RGB-D 数据与对应的焦距信息构建深度学习模型,以期减少从单幅图像学习场景深度信息固有的歧义性.

需要指出的是,即便焦距已知,由三维射影几何理论可知,从单幅图像(在无任何场景先验的情况下)也无法唯一地确定场景深度,如图 5 所示,空间中两个纹理形状相似的同类物体,如 S_1 和 S_2 ,成像都为 I ,当已知焦距信息 f 时,其对应的深度仍然是一对多的关系.这类歧义性是永远无法消除的.但考虑到在实际场景中,纹理形状相似的同类物体往往尺寸比较相近,因此,这种歧义性在实际运用中对深度估计精度的影响较有效焦距的影响要小得多.

图 5 (网络版彩图) 已知焦距学习单幅图像深度的歧义性

Figure 5 (Color online) Ambiguity of learning depth from single images with a fixed camera's focal length

4 结论

本文分析指出, 从单幅图像恢复场景深度信息在理论上存在固有的歧义性, 并通过实验对这种歧义性的存在性进行了验证. 要想从单幅图像可靠地学习场景深度, 在模型学习阶段, 需要同时将相机的内参数, 特别是有效焦距, 作为输入量参与学习. 下一步工作计划搭建含有不同焦距的 RGB-D 图像数据, 进一步探索相机内参数已知下的基于单幅图像的场景深度信息学习方法, 以期减少从单幅图像学习深度的歧义性.

参考文献

- 1 Hoiem D, Efros A A, Hebert M. Automatic photo pop-up. *ACM Trans Graphics (TOG)*, 2005, 24: 577–584
- 2 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*. Berlin: Springer, 2010. 224–237
- 3 Schwing A G, Urtasun R. Efficient exact inference for 3d indoor scene understanding. In: *Proceedings of the 12th European Conference on Computer Vision*. Berlin: Springer, 2012. 299–313
- 4 Saxena A, Chung S H, Ng A Y. Learning depth from single monocular images. In: *Proceedings of Advances in Neural Information Processing Systems*, Vancouver, 2005. 1161–1168
- 5 Saxena A, Chung S H, Ng A Y. 3-d depth reconstruction from a single still image. *Int J Comput Vision*, 2008, 76: 53–69
- 6 Saxena A, Sun M, Ng A Y. Make3d: learning 3d scene structure from a single still image. *IEEE Trans Pattern Anal Mach Intell*, 2009, 31: 824–840
- 7 Liu M, Salzmann M, He X. Discrete-continuous depth estimation from a single image. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 2014. 716–723
- 8 Baig M H, Jagadeesh V, Piramuthu R, et al. Im2depth: scalable exemplar based depth transfer. In: *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, Steamboat Springs, 2014. 145–152
- 9 Fouhey D F, Gupta A, Hebert M. Data-driven 3d primitives for single image understanding. In: *Proceedings of the IEEE International Conference on Computer Vision*, Sydney, 2013. 3392–3399
- 10 Fouhey D F, Gupta A, Hebert M. Unfolding an indoor origami world. In: *Proceedings of 13th European Conference on Computer Vision*. Berlin: Springer, 2014. 687–702
- 11 Ladicky L, Shi J, Pollefeys M. Pulling things out of perspective. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 2014. 89–96
- 12 Häne C, Ladicky L, Pollefeys M. Direction matters: depth estimation with a surface normal classifier. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015. 381–389
- 13 Liu C, Yuen J, Torralba A. Nonparametric scene parsing via label transfer. *IEEE Trans Pattern Anal Mach Intell*, 2011, 33: 2368–2382

- 14 Karsch K, Liu C, Kang S B. Depth transfer: depth extraction from video using non-parametric sampling. *IEEE Trans Pattern Anal Mach Intell*, 2014, 36: 2144–2158
- 15 Liu C, Yuen J, Torralba A. Sift flow: dense correspondence across scenes and its applications. *IEEE Trans Pattern Anal Mach Intell*, 2011, 33: 978–994
- 16 Konrad J, Meng W, Ishwar P, et al. Learning-based, automatic 2D-to-3D image and video conversion. *IEEE Trans Image Process*, 2013, 22: 3485–3496
- 17 Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: *Proceedings of Advances in Neural Information Processing Systems*, Montreal, 2014. 2366–2374
- 18 Eigen D, Fergus R. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 2015. 2650–2658
- 19 Wang X L, Fouhey D F, Gupta A. Designing deep networks for surface normal estimation. In: *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015. 539–547
- 20 Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015. 5162–5170
- 21 Liu F Y, Shen C H, Lin G S, et al. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans Pattern Anal Mach Intell*, in Press. doi: 10.1109/TPAMI.2015.2505283
- 22 Li B, Shen C, Dai Y, et al. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015. 1119–1127
- 23 Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgb-d images. In: *Proceedings of the 12th European Conference on Computer Vision*. Berlin: Springer, 2012. 746–760
- 24 Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: the kitti dataset. *Int J Robot Res*, 2013, 32: 1231–1237
- 25 Kendall A, Grimes M, Cipolla R. PoseNet: a convolutional networks for real-time 6-DOF camera relocalization. In: *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 2015. 2938–2946

The inherent ambiguity in scene depth learning from single images

Lei HE¹, Qiulei DONG^{1,2,3*} & Zhanyi HU^{1,2,3}

1 *National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;*

2 *University of Chinese Academy of Sciences, Beijing 100049, China;*

3 *CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing 100190, China*

*E-mail: qldong@nlpr.ia.ac.cn

Abstract Scene depth inference from a single image is currently an important issue in machine learning, its underlying rationale is the possibility of human depth perception from single images. Human depth perception from single images is the consequence of extended evolution, its underlying brain mechanism is far from well established whereby it seems imprudent to compute the depth based merely on such a fact. In fact, the 3D-to-2D object imaging process must satisfy some strict projective geometric relationship, and without prior knowledge on the camera's intrinsics, some ambiguity exists between the scene depth and the camera's focal length. We think since the camera's intrinsics were not accounted for, the currently reported single-image based depth learning approaches in the literature invariantly contain a crucial deficiency in theory. The ambiguity between the depth and focal-length is also verified by real images. We think in order to increase the accuracy of learning the scene depth from single images, cameras' intrinsics should be taken into account, at least the focal-length should also be used as input in both learning and inference phases.

Keywords single image, scene depth inference, imaging geometry, focal-length, ambiguity

Lei HE obtained the B.S. degree from Beijing University of Aeronautics and Astronautics, China. He is currently a Ph.D. candidate at the National Laboratory of Pattern Recognition, Chinese Academy of Sciences. His research interests include computer vision, machine learning, and pattern recognition.

Qiulei DONG received the B.S. degree in Automation from the Northeastern University in 2003, the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2008. Currently, he is an associate professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. His research interests include motion analysis, 3D computer vision, and pattern classification.

Zhanyi HU received his B.S. in automation from the North China University of Technology, Beijing, China, in 1985, and Ph.D. in computer vision from the University of Liege, Belgium, in 1993. Since 1993, he has been with the National Laboratory of Pattern Recognition at the Institute of Automation, Chinese Academy of Sciences, where he is now a professor. His research interests are Robot Vision, which include Camera Calibration and 3D Reconstruction, Vision Guided Robot Navigation. He was the local chair of ICCV2005, an area chair of ACCV2009, and the PC chair of ACCV2012.