深度立体回归的几何和上下文端到端学习

摘要

我们提出了一种新颖的深度学习架构，用于从整流的立体图像对中回归差异。我们利用问题几何的知识来形成一个使用深度要素表示的成本量。 我们学会了使用三维卷积合并上下文信息超过这个数量。 差异值从中回归使用建议的可区分软argmin操作的成本量，这允许我们端到端地训练我们的方法子像素精度无需任何额外的后处理或正规化。 我们在场景上评估我们的方法Flow和KITTI数据集以及KITTI我们设置了一个新的最先进的基准，同时显着快于竞争方法。

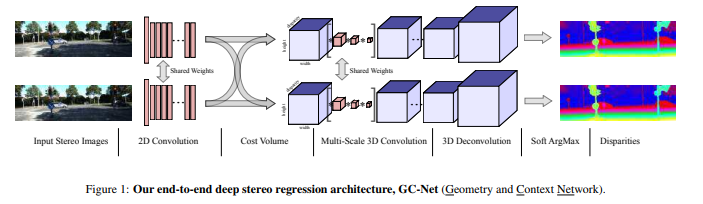
1介绍

准确估计三维几何立体图像是许多计算机视觉的核心问题应用，包括自动驾驶汽车和无人机[2]。在本文中，我们特别感兴趣的是计算整流的立体图像对之间的每个像素的差异。为实现这一目标，立体算法的核心任务是计算两个像素之间的对应关系图片。在真实场景中实现稳健性非常具有挑战性。目前最先进的立体声算法通常在无纹理区域，反射表面，薄结构和重复图案方面存在困难。很多立体声算法旨在通过池化或缓解这些故障基于梯度的正则化[15,23]。但是，经常这样需要在平滑表面和检测详细结构之间进行折衷。相比之下，深度学习模型已经取得了成功直接从原始学习强大的表示对象分类[28]，检测[17]和语义分割[31,3]中的数据。这些例子证明了这一点深度卷积网络对于理解语义非常有效。他们擅长于分类任务使用大型训练数据集进行监督。我们观察到了立体算法的这些挑战性问题的数量将受益于全球语义背景的知识，而不是仅仅依赖于局部几何。例如，给定车辆挡风板的反射表面，立体声算法很可能是错误的，如果它完全依赖于反射表面的局部外观以计算几何形状。相反，理解它是有利的这个表面的语义上下文（它属于一个载体）来推断局部几何。在本文中，我们将展示如何学习端到端的立体回归模型，具有理解更广泛的上下文信息的能力。迄今为止，利用深度学习表示的立体算法主要集中于使用它们生成一元术语[48,32]。应用成本匹配在估计像素差异时，深度一元表示的表现很差[32,48]。仍然使用传统的正则化和后处理步骤，例如半全局块匹配和左右一致性检查[23]。这些正规化步骤受到严格限制，因为它们是手工设计的浅层功能仍然容易受到上述问题的影响。本文提出这个问题，我们可以使用我们对立体几何的理解，通过深度学习来制定整个立体视觉问题吗？的主要贡献本文是一种端到端的深度学习方法来估算来自单个整流图像对的每像素差异。我们的架构如图1所示。它明确地说明了原因关于几何形状通过形成成本量，同时也推理使用深度卷积网络的语义公式。我们通过两个关键想法实现这一目标

•我们学习直接从数据中合并上下文，采用三维卷积来学习过滤成本体积超过高度×宽度×视差尺寸，

•我们使用完全可微分的软argmin函数，并允许我们回归子像素差异来自差异成本量的值。第3节介绍了这个模型。在第4节中，我们评估我们在合成场景流数据集[36]和集合上的模型KITTI 2012和KITTI的最新基准测试，2015年数据集[14,35]。 最后，在4.3节我们提出

证明我们的模型具有学习语义的能力推理和背景信息。



2.相关工作

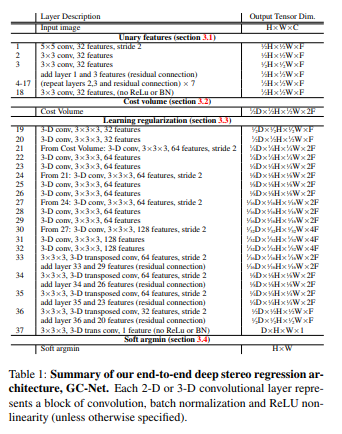
从立体图像对计算深度的问题已经研究了很长一段时间[5]。一项调查Scharstein和Szeliski [39]提供了立体声分类算法执行某些子集：匹配成本计算，成本支持聚合，视差计算和优化，或视差细化。这项调查还描述了第一个Middlebury数据集和相关的数据集评估指标。 KITTI数据集[14,35]更大数据集来自从具有激光雷达真实情况的移动车辆收集的数据。这些数据集首先得到改善手工设计技术，适用于立体声的所有组件我们提到一些值得注意的例子。匹配成本是对像素不相似性的度量潜在对应的图像位置[25]，其中绝对差异，平方差异和截断差异就是例子。基于渐变[16]或二进制模式的局部描述符，例如CENSUS [45]或可以使用Brief [7,22]。在假设下，对图像内容的认识可以更多地结合具有相似外观的相邻像素他们更有可能来自同一个表面和差距。这些技术的调查由提供Tombari等人。 [43]。本地匹配成本也可以在全球框架内进行优化，通常最小化能量函数组合本地数据项和成对平滑期。可以完成全局优化使用图形切割[27]或信念传播[26]，可以延伸到倾斜的表面[6]。一种流行而有效的全局优化的近似是Semi-Global匹配（SGM）的Hirschmuller [¨24]，其中动态规划优化了能量函数的路径形式在很多方向。

除了提供比较立体算法的基础，来自这些数据集的地面实况深度数据提供了使用机器学习以各种方式改进立体声算法的机会。张和Seitz [52]交替优化了视差和马尔可夫随机场正则化参数。 Scharstein和Pal [38]学习条件随机场（CRF）参数和李和Huttenlocher [29]训练非参数CRF模型使用结构化支持向量机。学习可以也被用来估计传统的信心立体算法，如随机森林方法Haeusler等人。 [19]。这种信心措施可以改善Park和Yoon [37]所示的SGM结果。可以训练深度卷积神经网络匹配图像补丁[46]。深度网络训练匹配Zbontar和Le-展示了9×9图像块，随后是非学习成本聚合和正则化。Cun [47,49]产生了当时最先进的结果。罗等。提出了一个显着更快的网络，用于计算本地匹配成本，作为使用Siamese网络的差异的多标签分类[33]。 Chen等人的多尺度嵌入模型。 [9]也提供了良好的本地匹配分数。另外值得注意的是DeepStereo的工作Flynn等人[12]，它学习了成本量的总和用一个单独的条件颜色模型来预测小说多视图立体声设置中的视点。梅耶等人。创建了一个大的合成数据集来训练用于视差估计的网络（以及光学流程）[34]，改进了最先进的技术。作为一个变种在网络中，提出了一维相关性视差线是乘法近似的立体声成本量。相比之下，我们的工作并没有崩溃计算成本量时的特征维度使用三维卷积来结合上下文。虽然这项工作的重点是双目立体声，但它值得注意的是，深度卷积网络的表征能力也能够实现深度估计单眼图像[10]。深度学习相结合由Liu等人提出的连续CRF。 [30]。而不是用标记的地面真相监督训练，未标记的立体声对可用于训练单眼模型[13]。

在我们的工作中，我们不应用后处理或正规化。 我们通过形成明确地推理几何完全可区分的成本量和合并背景来自具有3-D卷积架构的数据。 我们不学习概率分布，成本函数或分类结果。 相反，我们的网络能够直接从立体图像回归子像素的视差估计对。

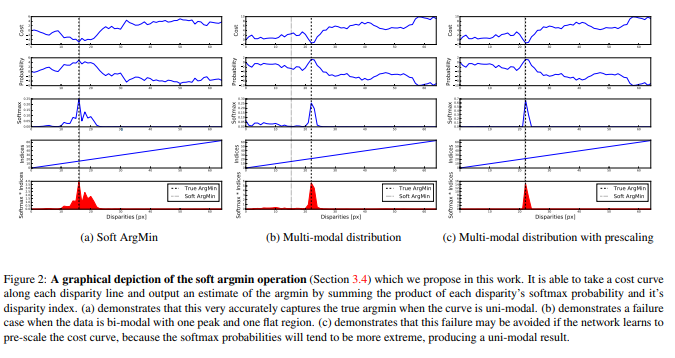
3.学习端到端的差异回归

而不是设计立体算法的任何步骤我们想从中学习端到端的映射使用深度学习的图像对到视差图。我们希望直接从中学到更优的功能数据。此外，这种方法有望减少很多工程设计的复杂性。然而，我们的目的不是天真地将机器学习架构构建为立体声模型的黑盒子。相反，我们提倡使用数十年的多视图几何研究[20]的见解来指导建筑设计。因此，我们通过开发代表传统立体声中每个主要组件的可分层来形成我们的模型管道[39]。这使我们能够在利用我们的几何知识的同时学习整个模型的端到端立体声问题。我们的架构，GC-Net（几何和上下文网络）如图1所示，更详细表1中的逐层定义。在其余部分中本节我们将详细讨论每个组件。后来，在4.1节中，我们提出了证明我们的定量结果设计决策。3.1。一元特征首先，我们学习用于计算的深度表示立体声匹配成本。而不是计算立体声使用原始像素强度匹配成本，这是常见的使用要素表示。动机是比较描述符对光度计外观中的模糊性更加鲁棒，并且可以包含局部上下文。在我们的模型中，我们通过a学习深度表示2-D卷积运算的数量。每个卷积层后面跟着一个批量标准化层和整流的线性非线性。减少计算量需求，我们最初使用5×5卷积滤波器两个步骤对输入进行子采样。在这一层之后，我们追加了8个残差块[21]，每个块都由两个串联的3×3卷积滤波器。我们的最终模型架构如表1所示。我们形成了一元特征通过这些传递左右立体图像层。我们在左右之间共享参数塔更有效地学习相应的功能。



3.2。成本量

我们使用深度一元特征来计算立体声通过形成成本量来匹配成本。虽然天真方法可以简单地连接左右特征映射，形成成本量允许我们约束模型以保持我们对立体视觉几何的了解的方式。对于每个立体图像，我们形成维度高度×宽度×（最大视差+ 1）×特征尺寸的成本量。我们通过连接实现这一目标每个一元特征与其相应的一元从跨越每个视差水平的相反立体图像，并将它们打包到4D体积中。至关重要的是，我们通过此保留了特征维度操作，不像以前使用点积的工作抽取特征维度的样式操作[32]。这使我们能够学习合并可以对特征一元进行操作的上下文（第3.3节）。我们发现，形成具有级联特征的成本量可以提高性能，而不是减去特征或使用距离度量。我们的直觉是通过保持这个特征是一元的，网络有机会学习绝对表示（因为它不是距离度量）和携带这通过成本量。这给了架构学习语义的能力。相反，使用距离度量将网络限制为仅学习相对功能之间的表示，并不能承载绝对功能表示到成本量。



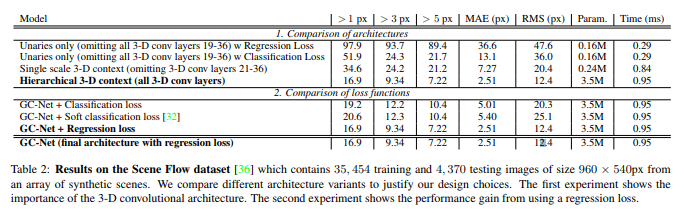
3.3。语义学习

鉴于这种差异成本量，我们现在想学习正则化函数，该函数能够考虑该卷中的上下文并改进我们的差异估计。一元之间的匹配成本永远不可能完美，即使使用深度特征表示。对于

例如，在均匀像素强度的区域（例如，天空）中，对于任何基于特征的成本曲线将是平坦的

在固定的本地环境中。我们发现像这样的地区可以导致跨视差维度的多模态匹配成本曲线。因此，我们希望学会规范和提高这个量。我们建议使用三维卷积运算来过滤和细化这种表示。三维卷积能够从中学习特征表示高度，宽度和差异尺寸。因为我们计算每个一元特征的成本曲线，我们可以从这种表示中学习卷积滤波器。在4.1节我们展示这些三维滤波器对学习环境的重要性并显着改善立体声性能。三维卷积的困难在于额外的维度是计算时间的负担培训和推理。深度编码器 - 解码器任务专为密集预测任务而设计的它们的计算负担是对子采样特征图进行编码，然后在解码器中进行上采样[3]。我们将这个想法扩展到三个维度。通过分抽样输入两步，我们也降低了三维成本量大小为8倍。我们形成了三维正则化网络有四个级别的子采样。作为一元一体已经被二次子采样，其特征是通过总因子32进行子采样。这允许我们明确地利用具有宽视野的上下文。我们申请每个编码器级别串联两个3×3×3个卷积。

为了使用原始输入分辨率进行密集预测，我们采用3-D转置卷积进行上采样解码器中的音量。描述了完整的体系结构见表1。子采样有助于在减少计算的同时增加每个特征的感知域。然而，它还通过降低空间精度和细粒度细节失去决议。因此，我们在上采样之前添加每个更高分辨率的特征图。这些残留层具有保留较高频率信息的优点，而上采样特征提供具有较大视野的细致特征图。最后，我们应用单个3-D转置卷积（反卷积），具有步幅2和单个特征输出。该层是进行密集预测所必需的原始输入维度，因为功能一元是子采样率为2。这导致了决赛，正规化成本量，大小为H×W×D。



3.4。 可区分的ArgMin

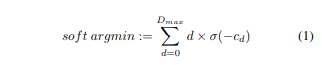
通常，立体算法产生最终成本量从匹配成本的一元。 从这卷，我们可以通过执行argmin操作来估计差异成本量差异维度。 但是，此操作有两个问题：

•它是离散的，无法产生子像素差异估计，

•它不可区分，因此无法区分

使用反向传播训练。为了克服这些限制，我们定义了一个软argmin1这是完全可区分的，能够倒退平滑的差异估计。 首先，我们转换预测成本，cd（对于每个差异，d）从成本量到a

通过取每个值的负数来计算概率。我们将差异中的概率量标准化使用softmax操作的尺寸σ（·）。 然后我们接受每个差异的总和d，由其归一化加权可能性。 图2中显示了一个图形说明并在（1）中以数学方式定义：



这项操作完全可以区分，并允许我们训练并且回归差异估计。 我们注意到[4]首先引入了类似的功能，并将其称为软注意机制。 在这里，我们将展示如何将其应用于立体回归问题。

但是，与argmin操作相比，它的输出受所有价值观的影响。 这使它容易受到影响多模态分布，因为输出不会采取最有可能的。 相反，它将估计加权平均值所有模式。 为了克服这个限制，我们依靠网络的正则化产生差异概率分布主要是单峰的。 网络还可以预先调整匹配成本来控制标准化的峰值（有时称为温度）post-softmax概率（图2）。 我们明确省略来自最终卷积层的批量归一化一元塔允许网络从数据中学习。



在下一节中，我们将展示制定我们的作为回归问题的模型允许我们以亚像素精度回归并且优于分类方法。此外，制定回归模型可以利用基于光度重投影误差的无监督学习损失[13]。