浙江工业大学

本科毕业设计论文

开题报告

题目：双目视觉立体匹配算法设计与实现

作者姓名 王 灏

指导教师 宣琦研究员

专业班级 通信工程 1301

学 院 信息工程学院

提交日期 2017 年 3 月 1 日

双目视觉立体匹配算法设计与实现

1　课题研究背景

近年来深度学习技术开始被广泛应用到各个领域中，特别是在图像处理领域，深度学习的特征提取特性表现尤为优秀。随着深度学习技术的深入发展，研究人员开始尝试在一些早已被充分研究的问题中应用神经网络，试图提高图像处理的效率和准确率。随着无人驾驶、机器人导航技术的发展，如何从二维的图像中提取出三维的深度信息成为了一个重要的命题。虽然自20世纪60年代以来，不断有人提出不同的方法来提取二维图像中的三维深度信息。准确率较高的传统的方法如激光测距仪，结构光等，受限于提取设备的繁琐复杂，无法得到广泛应用。而设备简单的传统双目视觉方法，受限于立体匹配算法（如SAD块匹配和SIFT算法）的不成熟，又无法提供准确的视差数据。在无人驾驶，机器人导航等技术迅猛发展的今天，我们亟需设计一个能从成本低廉的设备中提取出准确率较高的立体视觉算法。

2　概述

立体视觉是一种从二维平面图像中恢复出三维深度信息的技术。双目立体视觉系统是通过模拟人的双眼，仅需要两台安装在同一水平线上的数字摄像机，所得图像对经过立体矫正就可以投入使用，具有实现设备简单，成本低廉，并且可以在非接触条件下测量距离的优点。在机器人导航系统中双目立体视觉技术可以用于环境检测、障碍识别。在工业自动化控制系统中双目立体视觉技术可用于零部件的识别安装、产品质量检测。在安防监控系统中双目立体视觉技术常用于人流检测，危害报警。在无人驾驶技术中双目视觉技术还可用于道路环境的检测。

经典的双目立体视觉方法实现的一般步骤为：相机内参外参的离线标定，双目相机图像矫正，立体匹配以及光学三角形计算深度信息。通常采用不同水平位置的相机拍摄的两个图像，通过立体匹配找出对应点，计算视差d，视差指的是同一对象在左图像和右图像中的水平位置的差异——同一对象在左图像中的位置为（x，y），在右图像中的

位置为（x-d，y）。已知视差，我们可以使用以下关系计算它的深度z：

$$z = \frac{fB}{d}$$

其中f是相机的焦距，B是相机中心之间的距离。上述步骤中的立体匹配算法是双目视觉技术的核心问题，也是我们目前研究的重点。立体匹配的精度和速度对于立体视觉系统有着很大的影响。为提高双目视觉方法的性能，基于卷积神经网络进行立体匹配的双目视觉算法由下述几个部分组成。

2.1 训练模型

在KITTI 2012数据集[1] 中，从每个真实视差已知的图像位置处，提取出一对消极的和一对消极的9×9的训练图像对，构成二元分类数据集。并将等量的积极和消极图像对，作为输入数据，训练基于Tensorflow的卷积神经网络模型。

2.2 离线标定

通过常见的棋盘标定法，利用Matlab的stereo camera模块对双目摄像头进行离线标定。

2.3 数据获取

通过双目摄像头获取图像数据对。

2.4 双目矫正

利用Matlab标定所得的畸变参数和位置矩阵，在Opencv图像处理库中对双目摄像头所获取的左右图像对进行立体矫正，消除由于摄像头位置或摄像头本身性能导致的光学畸变。

2.5 立体匹配

将已矫正的左右图像对中的每个像素点，在可能的视差范围内，从左右图像对中依次提取出9×9的图像块，输入训练好的卷积神经网络模型，选取相似性得分最高的点，作为匹配点，并通过半全局匹配，左右一致性检查，中值滤波和双边滤波等后处理进一步优化视差图。

2.6 3D 恢复

根据所得的视差图，在Opencv图像处理库中进行3D恢复，计算3D坐标。

3 国内外现状

计算机立体视觉理论开始于20世纪60年代，美国麻省理工学院的Robert最先提出把二维图像分析推广到三维景物分析，标志着计算机立体视觉技术的诞生，并逐步发展成一门新的学科[2]。特别是20世纪70年代末，Marr等创立的视觉计算理论对立体视觉的发展产生了巨大影响，现已形成了从图像获取到最终的景物可视表面重建的比较完整的体系[2]。目前常见的立体视觉算法是双目立体视觉算法。立体视觉算法的核心是立体匹配算法。

Kong和Tao在2004年提出采用平方距离的总和（SAD）来计算初始匹配代价。他们训练了一个模型来预测三个类别的概率分布：初始视差正确的，由于前景目标过大导致初始视差不正确的，以及由于其他原因导致初始视差不正确的。预测概率被用来调整初始匹配代价[3]，并在2006年提出通过组合由计算归一化的不同的窗口大小和中心的互相关获得的预测[4]。

Zhang和Seitz在2007年提出使用一种替代优化算法来估算马尔科夫随机场超参数的最优值[5]。Scharstein和Pal构建了一个新的30个立体对的数据集，并使用它来得到条件随机场的参数[6]。Li和Huttenlocher在2008年提出了一个带非参数代价函数的条件随机场模型，并使用结构化支持向量机来得到模型参数[7]。

Haeusler等人在2013年使用了一种随机森林分类器来组合若干置信度度量的方式[8]。同样的，Spyropoulos等人在2014年训练了一个随机森林分类器来预测匹配代价的置信度，并且使用预测结果作为马尔科夫随机场中的软约束来减少立体视觉法的误差[9]。

Zbontar和LeCun在2015年提出使用卷积神经网络进行立体匹配。他们使用相似和不相似的图像对构建二元分类数据集来进行有监督的训练。卷积神经网络的输出用于初始化立体匹配代价。并配合一系列的后处理步骤：基于交叉的代价聚合，半全局匹配，左右一致性检查，亚像素增强，一个中值滤波器和一个双边滤波器来提高视差图的准确性。[10]

目前在主流的立体视觉评测库KITTI和Middlebury中，Zbontar和LeCun提出的MC-

CNN-acrt架构准确率高达2.43%（排名第五），所以使用卷积神经网络进行立体匹配算法的设计非常适合。

卷积神经网络是深度学习模型的一种，目前众多的研究院校和企业（Google，Facebook，阿里巴巴，百度，腾讯）都致力于深度学习模型的研究、应用、推广和优化。谷歌著名的AlphaGo围棋程序正是基于两个神经网络的算法设计的。深度学习模型在图像匹配中表现优异，其基本思想是通过有监督或者无监督的方式学习层次化的特征表达，从而达到从底层到高层的描述。主流的深度学习模型包括自动编码器、受限玻尔兹曼机、深度置信网络以及卷积神经网络等。下面详细介绍目前较为火热并强大的卷积神经网络，这也是在立体匹配算法中准备使用的技术。

1962年Hubel和Wiesel[11]通过对猫视觉皮层细胞的研究，提出了感受野(Receptive Field)的概念，即猫的视觉系统是分级的，这种分级可以看成是逐层迭代、抽象的过程。后来研究者便将这种逐步抽象的分层模型命名为深度学习模型。1984年日本学者Fukushima基于感受野概念提出的神经认知机(Neocognitron) [12]可以看作是卷积神经网络的第一个实现网络。

1988年LeCun等人将BP神经网络算法引入CNN，LeCun等人结合BP算法实现的LeNet-5模型在数字识别领域的表现强大，在银行支票的手写体字符识别中，识别正确率达到商用级别[13]。这是第一个真正多层结构的学习算法，它利用空间相对关系减少参数数目以提高训练性能。2006年，Hinton提出了深度置信网络（DBN），一种深层网络模型。使用一种贪心无监督训练方法来解决问题并取得良好结果。DBN（Deep Belief Networks）的训练方法降低了学习隐藏层参数的难度。并且该算法的训练时间和网络的大小和深度近乎线性关系[14]。Siamese网络通过在全连接层前添加两个共享权重的子网来得到两张图片的特征，并计算特征间的距离来比较图像的相似度，非常适合于双目视觉算法中的立体匹配[15]。

4 研究目标及主要内容

本篇文章的研究目标为设计一个双目视觉立体匹配算法。其主要内容包含如下：

1 通过文献检索查找国内外分类领域的最新研究，阅读比较各个算法的优劣，并针对本篇文章的研究目标设计可行的网络模型和整体方案。

2 从KITTI 2012已知视差的立体数据集中构建二元分类数据集（含数量相等的匹配

的数据和不匹配的数据）。

3 基于Tensorflow框架，设计并训练卷积神经网络模型。

4 双摄像头采集标定图片，并在Matlab中进行双目标定。

5 根据离线标定所得的相机参数进行利用Opencv对采集的图片进行双目矫正。

6 将矫正后的图片上传到服务器中，利用训练好的卷积神经网络模型进行立体匹配和后处理。

7 根据所得的视差图进行3D恢复，提取深度信息。

5 研究方法及手段

5.1 研究方法

1 文献：利用各个文献检索数据库，互联网检索以及博客文章的阅读，查询相关领域的最新技术以及发展趋势。

2 调研：通过与该领域公司内的技术人员交流沟通，了解目前基本解决方案以及最新的发展方向，并咨询目前该相关领域的难点和突破。

3 实验：通过对初步方案的不断实验完善，对于模型的摸索，调整，通过实验数据优化模型并寻找改进的方案。

5.2 研究手段

通过利用国内外文献资料搜索、互联网平台、校园数据库等途径，进行相关课题资料的收集、整理与分析，通过软件调试等手段来完善本课题。

5.3 所需工具

Python开发环境: 是一种面向对象的解释型计算机程序设计语言，Python 是由Guido van Rossum 在八十年代末和九十年代初，在荷兰国家数学和计算机科学研究所设计出来的。像Perl语言一样，Python 源代码同样遵循 GPL(GNU General Public License)协议。Python是纯粹的自由软件，语法简洁清晰，特色之一是强制用空白符(white space)作为语句缩进。Python具有丰富和强大的库。它常被昵称为胶水语言，能够把用其他语言制作的各种模块，尤其是C/C++，很轻松地联结在一起。

Matlab: 是matrix&laboratory两个词的组合，意为矩阵工厂（矩阵实验室）。是由美国Mathworks公司发布的主要面对科学计算、可视化以及交互式程序设计的高科技计算环境。它将数值分析、矩阵计算、科学数据可视化以及非线性动态系统的建模和仿真等诸多强大功能集成在一个易于使用的视窗环境中，为科学研究、工程设计以及必须进行有效数值计算的众多科学领域提供了一种全面的解决方案，并在很大程度上摆脱了传统非交互式程序设计语言（如C、Fortran）的编辑模式，代表了当今国际科学计算软件的先进水平。

Opencv图像处理库： OpenCV是一个基于BSD许可（开源）发行的跨平台计算机视觉库，可以运行在Linux、Windows和Mac OS操作系统上。它轻量级而且高效——由一系列 C 函数和少量 C++ 类构成，同时提供了Python、Ruby、MATLAB等语言的接口，实现了图像处理和计算机视觉方面的很多通用算法。OpenCV于1999年由Intel建立，目前由Willow Garage提供支持。

Tensorflow: 是谷歌基于DistBelief进行研发的第二代人工智能学习系统，其命名来源于本身的运行原理。Tensor（张量）意味着N维数组，Flow（流）意味着基于数据流图的计算，TensorFlow为张量从流图的一端流动到另一端计算过程。TensorFlow是将复杂的数据结构传输至人工智能神经网中进行分析和处理过程的系统。TensorFlow 表达了高层次的机器学习计算，大幅简化了第一代系统，并且具备更好的灵活性和可延展性。TensorFlow一大亮点是支持异构设备分布式计算，它能够在各个平台上自动运行模型，从手机、单个CPU / GPU到成百上千GPU卡组成的分布式系统。

6 进度安排

2016.12.20--2017.3.1 ：收集相关资料文献，学习Tensorflow框架，从KITTI2012数据集中构建二元分类数据集，初步训练网络模型。完成外文翻译、文献综述；熟悉课题，有初步设计方案；

2017.3.1—2017.3.15 ： 完成开题报告并有初步算法实现结果。

2017.3.16--2017.4.30 ： 完成立体视觉算法的初始版本，有较高的准确率。

2017.5.1--2017.5.31 ：算法的优化。撰写毕业论文初稿。

2017.6.1--2017.6.20 ：论文修改，毕业答辩，提交相关文档资料。

参考文献

[1] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The KITTI dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[2] Marr D C.A Computational Investigation into the Human Representation and Processing of Visual Information [M]. San Francisco：W．H．Freeman and company，1982．

[3] Kong D, Tao H. A method for learning matching errors for stereo computation.[C]. british machine vision conference, 2004.

[4] Kong D, Tao H. Stereo Matching via Learning Multiple Experts Behaviors[C]. british machine vision conference, 2006.

[5] Zhang L, Seitz S M. Estimating Optimal Parameters for MRF Stereo from a Single Image Pair[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2): 331-342.

[6] Scharstein D, Pal C. Learning conditional random fields for stereo[C]. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2007.

[7] Li Y, Huttenlocher D P. Learning for stereo vision using the structured support vector machine[C]. computer vision and pattern recognition, 2008: 1-8.

[8] Haeusler R, Nair R, Kondermann D, et al. Ensemble Learning for Confidence Measures in Stereo Vision[C]. computer vision and pattern recognition, 2013: 305-312.

[9] Spyropoulos A, Komodakis N, Mordohai P, et al. Learning to Detect Ground Control Points for Improving the Accuracy of Stereo Matching[C]. computer vision and pattern recognition, 2014: 1621-1628.

[10] Zbontar J, Lecun Y. Computing the stereo matching cost with a convolutional neural network[C]. computer vision and pattern recognition, 2015: 1592-1599.

[11] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture

in the cat's visual cortex[J]. The Journal of physiology, 1962, 160(1): 106-154.

[12] Fukushima K, Miyake S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position[J]. Pattern recognition, 1982, 15(6): 455-469.

[13] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[14] Hinton G E, Osindero S, Teh Y, et al. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7): 1527-1554.

[15] Bromley J, Guyon I, Lecun Y, et al. Signature verification using a Siamese time delay neural network[C]. neural information processing systems, 1993: 737-744.