6

浙江工业大学

本科毕业设计论文

文献综述

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

摘要：本文回顾了立体视觉研究的发展和现状，简述了立体视觉技术在工业军事等领域的广泛应用，并介绍了双目立体视觉技术的研究背景和基本原理。随后，本文回顾了卷积神经网络（Convolution Neural Network，CNN）的发展历史，并简单描述了一个卷积神经网络的基本结构。此外本文还回顾了立体匹配算法的发展现状，以及卷积神经网络在立体匹配中的应用。最后，本文介绍了当前主流的立体视觉数据集 KITTI 和 Middlebury 及其对应的评测库。

关键词：卷积神经网络、立体视觉、立体匹配

1　引言

1.1 立体视觉研究的发展历史

众所周知，自然界中的物体都是三维立体的，人类通过双眼可以获取物体的三维立体信息。但一般的摄影系统只能把三维的物体以二维的形式保存和记录下来，丢失了大量的深度信息。计算机立体视觉的开创性工作是从 20 世纪 60 年代中期开始的，美国麻省理工学院的 Robert 把 2 维图像分析推广到 3 维景物分析，标志着计算机立体视觉技术的诞生，并在随后的 20 年中迅速发展成一门新的学科[1]。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[1-2]。

1.2 双目立体视觉的研究背景和基本原理

立体视觉技术在军事、工业的各个方面都有着非常广泛的应用背景，具有很大的商业价值和使用价值。在军事领域中，众所周知，现代的战争是信息化战争，对信息的搜索和利用率往往决定了一场战役的胜负。利用侦察卫星，单兵侦察等手段获取二维图像数据，并通过立体视觉算法进行三维重构，模拟出三维的战场环境，可以为作战指挥提

供信息支持。在机器人导航系统中立体视觉技术可以用于环境检测、障碍识别。在工业检测系统中立体视觉技术常用于产品质量检测。在无人驾驶技术中立体视觉技术还可用于道路环境的检测。

目前对立体视觉的研究主要有三类方法：第一类方法是直接利用测距器（如激光雷达，结构光等）获得深度信息，重建三维模型的方法。这类方法通常需要特殊仪器，成本较高，而且由于设备复杂，应用环境受限；第二类方法是根据光学成像原理及统计假设，仅利用一幅图像所提供的信息推断三维模型的方法。这种方法由于受到单一图像所能提供信息的局限性，难以提供准确的深度信息；第三类方法是利用不同视点上的，是不同时间或不同位置拍摄的两幅或多幅二维图像重构三维模型的方法。第三类方法是目前较为常见的立体视觉方法。双目立体视觉就是第三类方法的常见实现方式。

双目立体视觉一般由两摄像机从不同角度同时获取目标场景的两幅图像，或由单摄像机在不同时刻从不同角度获取目标场景的两幅数字图像，并基于视差原理恢复出场景的三维几何信息，重建场景中物体的三维轮廓及位置。

双目立体视觉技术实现的一般步骤为：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）。已知视差可以根据以下公式

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

计算它的深度 z，其中 f 是相机的焦距，B 是相机中心之间的距离。上述步骤中的双目立体匹配算法是双目视觉技术的核心问题。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[3-4]。

目前立体匹配面临许多挑战性的问题：遮挡匹配问题、弱纹理或重复匹配问题、深度不连续匹配问题、光照变化引起匹配问题等。近几年来随着深度学习（Deep Learning）技术的发展，其在图片、语音方面强大的特征提取表现使得利用 CNN 进行立体视觉匹配变成可能。

2　卷积神经网络

2.1 卷积神经网络的发展历史

1962 年 Hubel 和 Wiesel 在研究猫脑皮层中用于局部敏感和方向选择的神经元时发现其独特的网络结构可以有效地降低反馈神经网络的复杂性，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[5]。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 基于上述提出的感受野概念，构建了神经认知机(Neocognitron)，神经认知机是卷积神经网络的第一个实现网络，其将视觉模式分为多个子模式（特征），然后进入分层连接的特征平面处理[6]。随后，更多的科研工作者对该网络进行了改进。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[7]。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[8]。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[8]。

2.2 卷积神经网络的网络结构

如图 1 所示，卷积神经网络是一个多层的神经网络，每层由多个二维平面组成，每个平面由多个独立神经元组成。

C 层为特征提取层（卷积层），每个神经元的输入与前一层的局部感受野相连，并提取该局部的特征，一旦该局部特征被提取后，它与其它特征间的位置关系也随之确定下来；S 层是特征映射层（池化层），网络的每个计算层由多个特征映射组成，每个特征映射为一个平面，平面上所有神经元的权值相等。特征映射结构采用影响函数核小的 sigmoid 函数作为卷积网络的激活函数，使得特征映射具有位移不变性。

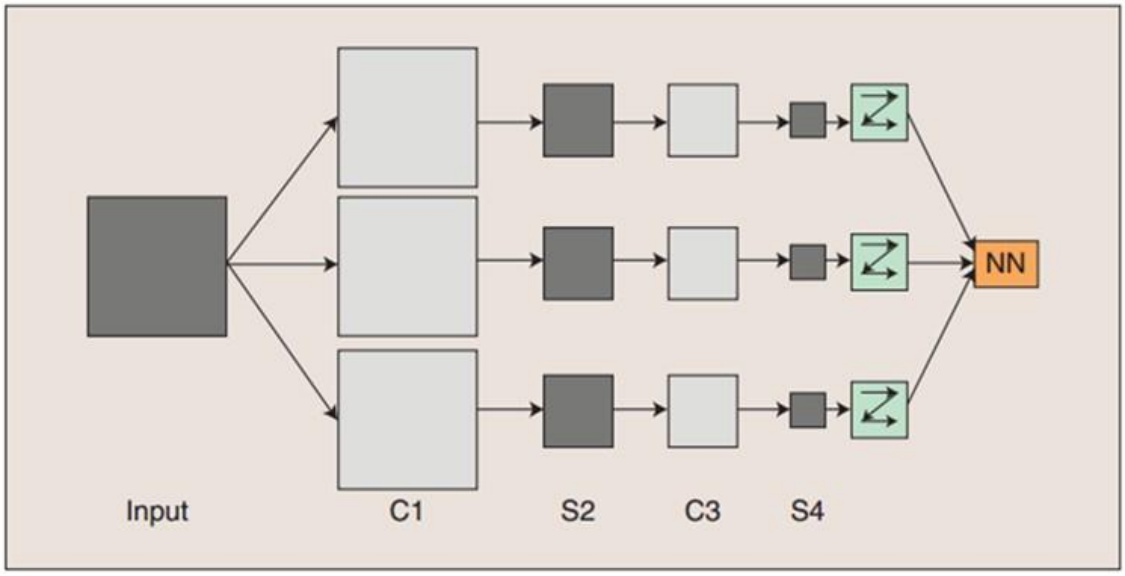

图 1　卷积神经网络的概念示范

此外，由于一个映射面上的神经元共享权值，因而减少了网络自由参数的个数，降低了网络参数选择的复杂度，所以卷积神经网络是一种权值共享网络，网络参数的减少使得能构建更加深（层数）以及更加宽（特征树）的网络结构。卷积神经网络中的每一个特征提取层（C-层）都紧跟着一个用来求局部平均与二次提取的计算层（S-层），这种特有的两次特征提取结构使网络在识别时对输入样本有较高的畸变容忍能力。

3 立体匹配研究及发展现状

3.1 立体匹配算法

3.1.1 区域匹配法

区域匹配算法是以参考图像中某一待匹配像素点为中心，选择一个矩形窗口作为其约束区域，然后，从目标图像中，寻找与其匹配的像素点，同样也以寻找的像素点为中心，选择同样大小一个矩形窗口，用此区域内像素点的属性信息约束该像素点的相似性计算。在左右图像中，两个像素之间的相似性必须满足一定相似性条件，才认为它们是相似的，区域匹配算法的目的是获取稠密视差图。区域匹配算法的缺点在于受图像的仿射畸变和辐射畸变的影响较大，而且像素点约束窗口的大小与形状选择比较困难，选择过大，在深度不连续处，视差图中会出现过度平滑现象；选择过小，对像素点的约束比较少，图像信息没有得到充分利用，容易产生误匹配。

随着研究的深入，开始出现各种改进的算法，如自适应窗体算法，自适应权重算法

等。目前较先进的 PMSC（PatchMatch-based Superpixel Cut）算法采用双层匹配代价计算，首先匹配小正方形的窗口然后在更大的不规则窗口上进行代价聚合，最后进行全局优化。该算法截至 2016 年 12 月在 Middlebury 立体视觉测评中排名第一。[9]

3.1.2 特征匹配算法

特征的匹配算法，主要是基于几何特征信息（边缘、线、轮廓、兴趣点、角点和几何基元等），针对几何特征点进行视差估计，所以先要提取图像的特征点，尽而利用这些特征点的视差值信息来重建三维空间场景。目前较流行的 SIFT(Scale Invariant Feature Transform)就是一种提取局部特征的算法，在尺度空间中寻找极值点，提取位置、尺度、旋转不变量，生成关键点特征描述符，然后根据这些不变量特征进行匹配[10]。SIFT 特征是基于物体上的一些局部外观的兴趣点而与影像的大小和旋转无关[10]。对于光线、噪声、些微视角改变的容忍度也相当高。基于这些特性，它们是高度显著而且相对容易撷取，在母数庞大的特征数据库中，很容易辨识物体而且鲜有误认[10]。使用 SIFT 特征描述对于部分物体遮蔽的侦测率也相当高，甚至只需要 3 个以上的 SIFT 物体特征就足以计算出位置与方位[10]。现今的电脑硬件速度下和小型的特征数据库条件下，辨识速度可接近即时运算[10]。但是基于特征的匹配算法，得到的匹配点对数量较少，无法满足生成视差图的要求，所以使得它的应用领域受到了限制[10]。

3.1.3 CNN 立体匹配

2015 年，Zbontar 和 LeCun 在 CVPR 上提出使用卷积神经网络来计算立体匹配代价，并在 2016 年进一步完善。CNN 立体匹配是一种局部匹配算法，主要是针对立体匹配算法的第一阶段：匹配代价的计算。CNN 立体匹配使用相似和不相似的图像对构建二元分类数据集对卷积神经网络进行有监督的训练，然后通过卷积神经网络进行小图像块的相似性度量。卷积神经网络的输出用于初始化立体匹配代价，最后经过一系列的后处理步骤来进行代价聚合，视差计算以及视差精化[3-4]。

目前在立体视觉评测库 MiddleBury 和 KITTI 中，基于卷积神经网络的立体视觉匹配算法仍然排在前列。

3. 2 立体视觉数据集

立体视觉真实标签数据对于卷积神经网络的构建非常重要，优质的标签数据可以获得更好的训练模型。目前主流的立体视觉数据集是 KITTI 和 Middlebury 数据集。

3.2.1 KITTI 数据集

KITTI 立体数据集[11-12]是从安装在汽车车顶上两个相隔约 54 厘米的摄像机拍摄的经整流的灰度图像对的集合。图像是在白天晴朗和多云的天气下，在 Karlsruhe 市中心和郊区开车记录的。图像分辨率为 1240×376。旋转激光扫描仪安装在左侧相机后面，记录地面的真实深度，标记了约 30%的图像像素。

测试集的地面真实视差并未公开，而是提供了一个在线排行榜供研究人员在测试集上评估他们的算法。每隔三天允许提交一次。错误率是计算那些真实视差和预测视差相差超过三个像素的像素的比例。这就意味着，例如，错误的容许范围为 3 厘米则转换成物理距离为距离相机 2 米，而 80 厘米的容许范围，就是距离相机 10 米。

目前存在两个 KITTI 立体视觉数据集：KITTI 2012 和较新的 KITTI 2015。为了完成立体视觉的计算任务，他们几乎是相同的，但较新的数据集的改进了光流任务的一些方面。2012 年的数据集包含 194 个训练图和 195 个测试图像，而 2015 年的数据集包含 200 个训练图和 200 个测试图。较新的数据集引入了一个微妙但重要的区别：运动车辆是密集标记的，并且汽车玻璃也被包括在评估中。

3.2.2 Middlebury 数据集

Middlebury 立体视觉数据集[13-16]的图像对来自光照可控的室内场景。数据通过结构光来测量真实视差，视差的密度和精度都比在 KITTI 数据集好。该数据集共发布了五个独立的数据集，分别是在 2001 年，2003 年，2005 年，2006 年和 2014 年。

在 2005 年，2006 年，和 2014 年数据集的每一个场景都是根据多种光照条件和快门曝光方式拍摄的，有一组典型的图像对是同一个场景在四种光照条件和七种快门曝光方式下拍摄的共计 28 张图片。一个类似 KITTI 的在线排行榜，显示了所有提交算法的排名表。参与者只有一次提交结果的机会。测试数据集包含 15 幅来自 2005 年和 2014 年数据集的图像。数据集提供全分辨率，半分辨率和四分之一分辨率的图片。误差率的计算是根据全分辨率的，如果算法输出的是半分辨率或四分之一分辨率的视差图，它们在计算错误率之前上采样。

4 总结

双目立体视觉经过几十年的研究已经取得了显著的成果，出现了各种专门的硬件设计和视频速率实时的立体视觉系统，目前立体视觉技术被广泛应用于军事侦察，无人驾

驶，机器人导航以及工业自动化系统中。但是，从普遍的意义来讲，由于难以彻底解决立体匹配的对应点问题，具体的立体视觉系统一般都是有针对性的、不普遍适用的，还无法与人类的双目视觉系统相媲美。

但近年来，随着深度学习技术在计算机视觉处理领域的快速发展，使用深度学习算法建立一个普遍适用的立体匹配网络也成为一个重要的趋势。

参考文献

[1] 周星，高志军.立体视觉技术的应用与发展[A].工程图学学报,2010 年第 4 期，No.4.

[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] Jure Žbontar and Yann LeCun. Computing the stereo matching cost with a convolutional neural network. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[4] Žbontar J, Lecun Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of Machine Learning Research, 2016, 17(65): 1-32.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 闫雅晶. 《纽约时报》与中美关系（1972-1979）[C]., 2013.

[10] Lowe D G. Distinctive Image Features from Scale-Invariant Keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[11] 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.

[12] Fatma Guney and Andreas Geiger. Displets: Resolving stereo ambiguities using object knowledge. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

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

[14] Daniel Scharstein and Richard Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision, 47(1-3): 7–42, 2002.

[15] Daniel Scharstein and Richard Szeliski. High-accuracy stereo depth maps using structured light. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June.2003.

[16] Daniel Scharstein, Heiko Hirschmu¨ller, York Kitajima, Greg Krathwohl, Nera Neˇsi´c, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. German Conference on Pattern Recognition (GCPR), September 2014.