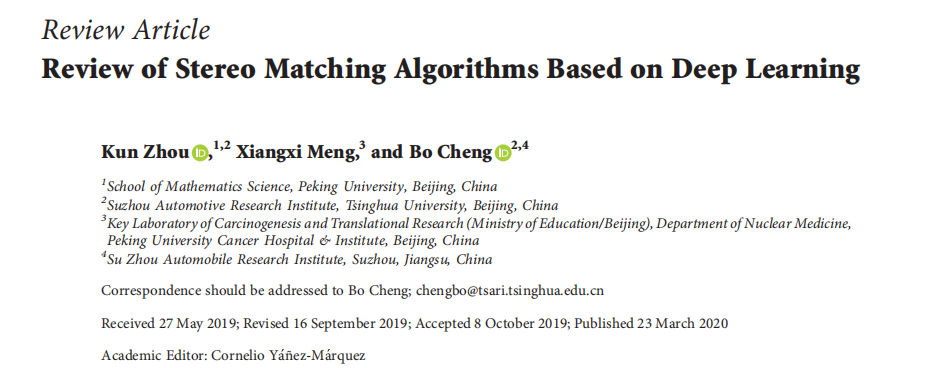
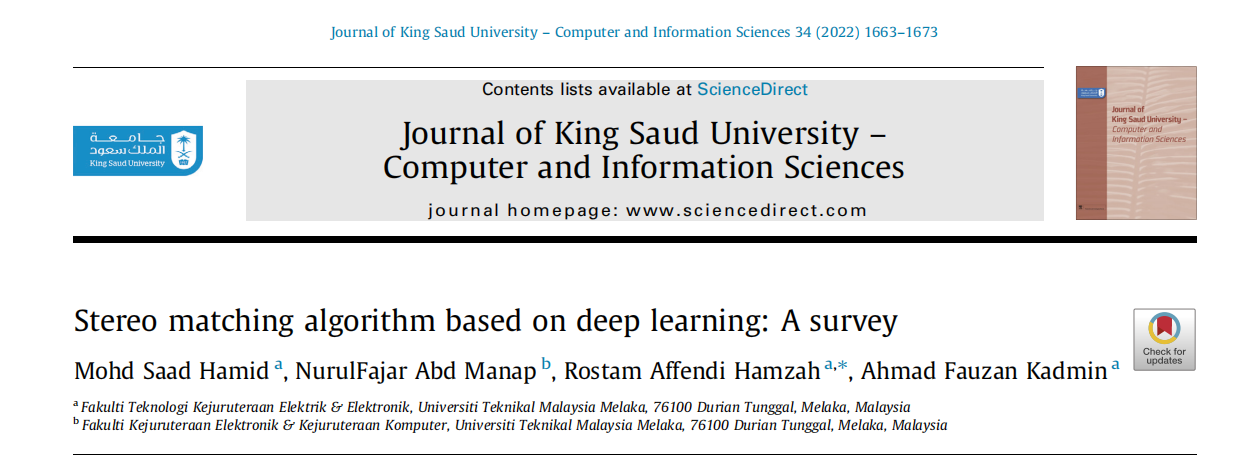
本周首先阅读了论文:

**Review of Stereo Matching Algorithms Based on Deep Learning**



以及

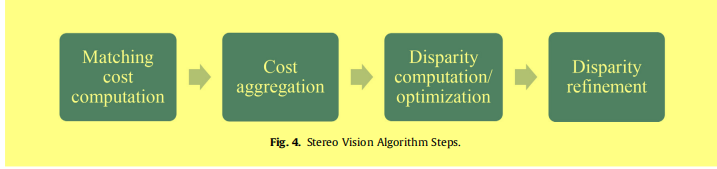
**Stereo matching algorithm based on deep learning: A survey**



对于第二篇文章主要介绍的是非端到端以及端到端中的主要算法的实现（相对于上篇的实现以及相关叙述会更加详细）所以，这篇报告主要会基于第一篇的介绍顺序进行介绍，同时包含第二篇文章的详细介绍相关网络的部分（图片以及文字）。

首先，文章的主线结构是将相关处理方法分成四个主要类型：传统方法、非端到端方法、端到端方法、非监督学习算法，来介绍stereo matching的发展进程。

对于stereo matching的算法的阶段划分，和之前阅读过的一个关于传统匹配方法的阶段划分是一致的（虽然这种区分并不明显或者会缺少某些阶段（如有些基于全局匹配的方法和局部匹配的阶段划分也是不同的））：



文章关于介绍传统双目匹配算法的相关优缺点总结如下：

1、传统算法通常采用手工设计的特征和后处理函数，这些特征和函数需要根据经验进行选择和调整。例如，SGM算法使用了基于梯度的特征，并采用了后处理函数来优化结果。

2、传统算法通常需要大量的计算资源和时间，因为它们需要处理大量的像素和计算量。例如，SGM算法需要计算每个像素的代价函数，并使用动态规划算法来优化结果。

3、传统算法通常无法处理复杂的场景和光照变化（缺乏对全局上下文信息的建模能力）。例如，SGM算法在处理遮挡和光照变化时表现不佳。

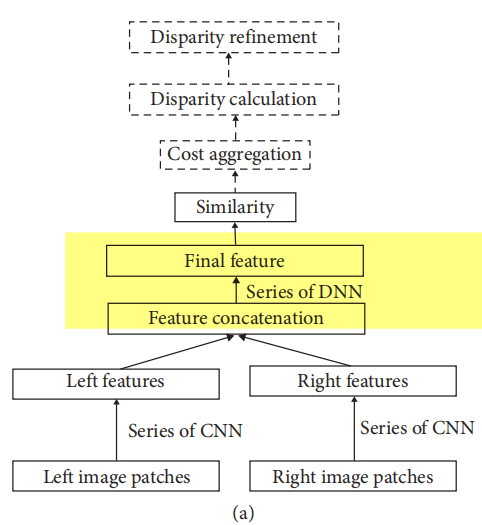
4、传统算法通常无法处理大规模的图像数据，因为它们需要处理大量的像素和计算量。例如，SGM算法在处理高分辨率图像时表现不佳。

传统的方法并不这一篇文章的重点，因此所占篇幅也是比较小的。

之后就有了引进神经网络改进传统方法的非端到端的算法，非端到端的思想主要仍然是基于传统算法的阶段划分，并尝试采用神经网络用于替代上述算法阶段的其中一个部分（文章举例了两个主要的网络，主要用于采用CNN\DNN代替**特征提取**部分）：

1、 MC-CNN：它采用深度孪生网络（由多个CNN和DNN层组成）来衡量两个9x9图像块之间的相似性。通过交叉基于成本聚合和半全局匹配对相似性（成本）进行了改进，并进行了左右一致性检查以消除遮挡区域的错误。

2、Siamese：通过共享权重的方式来比较两个输入样本的相似性。Siamese网络由两个相同的子网络组成，这两个子网络共享相同的权重参数。每个子网络将输入样本映射到一个低维特征空间中，然后通过比较这些特征向量来计算样本之间的相似性。对于每对输入样本，通过计算它们在特征空间中的距离（如欧氏距离或余弦相似度），并与预定义的相似性阈值进行比较。如果样本对属于相似类别，则希望它们的距离较小；如果样本对属于不相似类别，则希望它们的距离较大。对比损失会根据这些距离的大小来更新网络的权重参数，以最小化相似样本对的距离并最大化不相似样本对的距离。



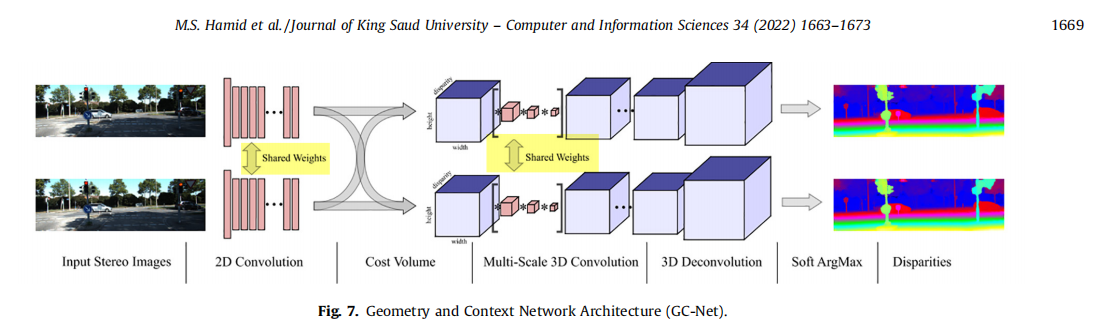
（Siamese网络架构，同时对于黄色标注部分占用较多时间的问题（多个DNN），文章中提到了一种改进方法，就是采用在欧几里得空间中的点积代替该部分）

对于端到端网络结构，则采用了非传统（与上文提到的四个阶段划分没有明确关系）结构，端到端视差估计网络的主要实现思想是将立体匹配流程的所有步骤无缝地融合到一个网络中进行联合优化，直接从立体图像中生成密集的视差图。通过训练一个神经网络，使其能够从立体图像对中直接预测出 disparity map，而无需手工设计的特征或中间步骤。

划分常见实现思路为2D编码-解码结构（encoder-decoder）和包含3D卷积的级联细化和正则化模块。2D编码-解码结构由一系列堆叠的2D卷积神经网络（CNN）组成，其中使用了一些跳跃连接来捕捉详细信息，并提高性能。3D结构则利用了视差维度，通过使用3D CNN来进行视差估计。

主要介绍以下几个代表性的网络：

1. DispNet方法提出了一维相关层来进行成本计算（第一个端到端的视差估计网络）。将图像对作为输入并直接预测视差。由于网络接收整个图像作为输入，相比非端到端方法中采用的Siamese网络，它具有更高的效率。在对象遮挡、重复模式或无纹理区域等困难区域难以找到正确的对应点。
2. CRL两阶段架构：第一阶段进行初始预测，第二阶段通过生成跨多个尺度的残差信号进行进一步细化/校正。两个阶段的主要结构与DispNetC类似，两个阶段的输出相加得到最终的视差。这种更复杂的结构导致了更强大的表示能力，而两阶段架构有助于捕捉细化信息。但是速度较上者更慢。



（截取自另外一篇论文）

3、GC-Net首次使用3D卷积网络来聚合成本体积的上下文信息。他们在计算成本体积时没有折叠特征维度，而是形成了一个四维成本体积，其中包含了沿着视差维度连接的图像对的特征，并通过3D卷积网络给出视差预测。

无监督立体匹配方法（不需要ground truth训练集）的几种实现方式。这些方法主要依靠最小化光度变换误差来驱动网络的无监督学习。其中，通过空间变换和视图合成，提出了几种无监督学习方法：

DeepStereo图像合成网络，通过从附近图像中选择像素来生成新的视图。Deep3D网络解决了新视图合成的问题，通过最小化像素级重构损失，从输入的左图像（即源图像）生成右视图。

基于Deep3D，将单目深度估计问题重新定义为两个子问题，即视图合成过程和标准立体匹配。网络的主要结构是Deep3D和DispNet的组合，Deep3D提供了另一个视图，而DispNet根据初始图像和新视图预测视差。

Garg提出了第一个用于单视图深度估计的无监督网络，使用图像重构损失。该网络通过使用预测的深度生成目标图像的逆变换来重构源图像。尽管由于单个图像的整体尺度模糊性，这种单目深度不仅在绝对意义上不准确，而且在恢复细节方面也不准确，但该网络的性能与单视图深度估计的最先进的监督方法相当。

Ren等人采用双线性采样网络生成图像，从而得到完全可微分的训练损失。Yu等人扩展了图像重构损失，并结合空间平滑损失进行无监督光流学习。然而，直到Godard等人的工作，他们都没有利用预测之间的几何一致性。Godard等人证明了仅通过图像重构解决问题会导致质量较差的深度图像。

由于该部分现阶段实现准确性以及速度都不是非常理想，所以仍然又较大的发展空间。

最后，对于三种基于神经网络的实现思路，大致的缺点和不足总结如下表：

|  |  |  |  |
| --- | --- | --- | --- |
| Non-end-to-end | MC-CNN, content CNN, SGM-Net | 简单以及能够实现相对于传统方法更好的效果 | 计算负担高；  接受域有限，缺乏上下文信息以及仍然需要用后处理 |
| End-to-end | PSMNet, GC-Net | 设计简单，部分网络能够实现速度和准确度上较好的折中 | 成本负担大，内存占用大；  时间成本高；需要ground truth数据 |
| Unsupervised | LR-consistency check | 不需要ground truth数据 | Poor performance |