计算机视觉课程报告

张津宁
21821083

摘要

本文主要推荐"Ganerated hands for real-time 3d hand tracking from monocular rgb."[1]这篇论文提出的使用单目RGB摄像头来跟踪手部姿态的实现。本文首先提出手部姿态跟踪的相关问题，随后给出这篇论文[1]的解决方法，并分析其实现细节以及创新点，最后根据这篇论文给出的实验结果数据评价该方法的性能。

1. 相关问题

手部姿态估计与跟踪问题，目前大量应用于AR/VR等用户交互模块中，它对于检测精度、速度、成本的要求也是非常高的。对于手部姿态估计问题，目前以及有很多不同的解决方案。例如采用图像结合深度信息的RGB-D设备来做姿态估计，或者使用多视点相机来获得额外的深度信息，以便于估计手部的 3D姿态。而在这篇文章中提出了新的思路：仅依赖一个单视点、无深度信息的相机作为输入源来完成 3D姿态估计。这种单目RGB的方法虽然会损失一些精度，但是也降低了检测设备的限制，目前来说是非常好的单目RGB条件下的检测方法。

2. 解决方法

这篇文章的特点在于以下两方面：1.一部分的训练数据由GAN网络生成；2.仅需要一个RGB摄像头采集的视频流作为输入。由于仅依赖单目RGB数据，又要估计3D姿态，那么必然会有从 2D到 3D转化过程中产生的不确定性，这篇文章采用了一个修改过的ResNet，结合一个平行投影层来同时估计 2D与3D的姿态，最终综合 2D与 3D的估计结果，使用最小化一个能量函数的优化方法与手的骨骼模型做拟合来得到 3D姿态。另一方面，为了扩充这一神经网络的训练数据集，作者通过一篇 2017 年的Cycle-Consistent GAN方法[2]来生成一部分训练集，以此提升整个检测模型的精度。

2.1. 训练数据生成

文中提到了训练集数据生成的必要性。由于网络训练时需要知道当前输入图像对应的正确 3D姿态数据来用于计算loss，做真实训练数据的采集，尤其是标定手部关节的位置姿态信息是比较费时的。因此这里考虑使用GAN网络生成一部分训练集：将预先绘制好的缺乏真实感的手部图像通过GAN网络转化为“真实”的图像。这一操作一定程度上消去了生成数据与真实数据之间的差异性，防止最后训练出来的网络对于真实数据的性能不好。另一方面，由于生成数据是由预先设定好的手部模型绘制来的，它的姿态标定信息就非常容易获得。

其中负责将生成图像转化到“真实”图像的GAN网络称为几何一致的CycleGAN网络，它参考了[2]中的CycleGAN结构，使得在训练时不需要合成数据与真实数据一一对应，实际上对于合成的手部图像，我们也找不到对应的真实图像。网络结构如下页图 2 所示。

该网络的输入为生成与真实的图像，网络结构的左右对称。左侧对生成图像的输入做“synth2real”后，交给判别器，接着又通过“real2synth”转化一次，与原先的输入做一次L1loss。这就是所谓的CycleGAN，在保留GAN结构的同时引入一次逆向转化的loss，来确保生成图像与“真实”转化结果之间保持一种更加靠近“双射”的映射关系。另外在最左侧的结构中还加入了一个几何一致性的交叉熵（通过两个图像轮廓来计算），来确保CycleGAN网络在做映射时保持手部的轮廓不变。对于轮廓的提取，这里使用一个预先训练好的 2 分类UNet网络来完成。

另外，为了进一步扩充生成的数据集，借助上面的轮廓信息我们可以替换多种背景。也可以在手部上层叠加一个额外的mask来制造出遮挡效果。如图 1 所示。文中提到这样的后续处理方式可以降低CycleGAN网络的复杂度，让它只处理纯色背景的图像数据。

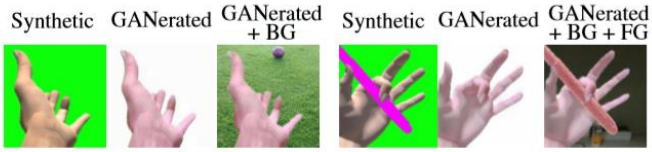

图 1 Synthetic images

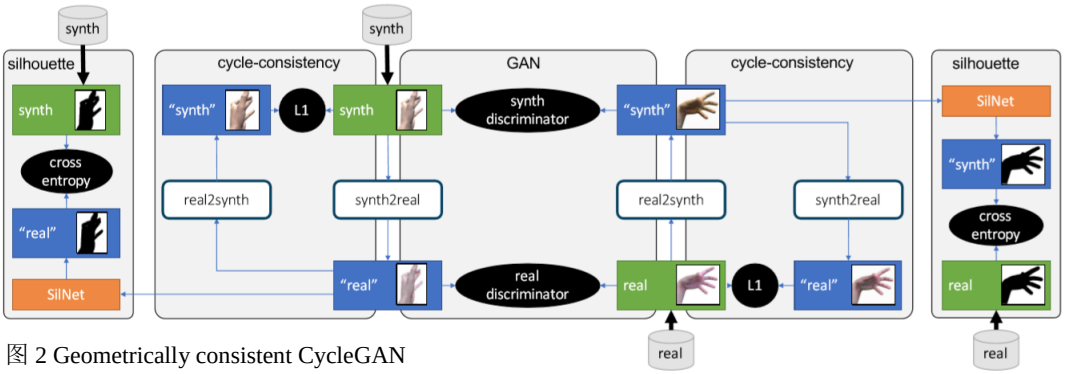

图 2 Geometrically consistent CycleGAN

2.2. 手部结点姿态回归

对于手部姿态的检测，这里使用一个CNN网络来预测手部的 21 个结点的位置（每根手指 4 个结点，加上手腕一个结点）。网络结构如图 3 所示。输入为一张图像数据，通过 10 个残差块组成的一个残差网络ResNet来得到一组 3D位置的中间结果（相对于根结点的位置）。随后通过一个可微的平行投影层ProjLayer来将 3D中间结果转化为 2D的heatmap。接着再使用这些 2D heatmap作为卷积层的输入来生成最终的 2D heatmap和 3D 位置信息。这里的loss计算需要同时对最终 2D、3D结果以及中间 3D结果来做。这里我不是很明白为什么需要一个 3D中间结果以及投影层生成的一组 2D heatmap中间结果，文中只是提到这样可以更好地结合 2D和 3D的预测。

这个网络的训练数据集同时包括了真实数据以及生成数据（占 60%）。另外在实际使用时，需要读取上一帧的 2D预测结果来框取当前帧输入的手部图像。

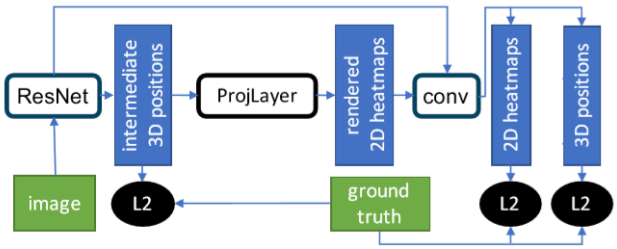

图 3 RegNet architecture

2.3. 骨骼拟合

对于上面RegNet生成的 2D heatmap与 3D相对位置预测结果，这里将它们拟合到一个手部骨骼模型上。由于人的手部骨骼模型大部分结点的旋转自由度有一定限制，采用拟合的方式可以更好地遵循这些限制，避免一些现实中不可能的姿态。首先定义一组参数$\Theta = (t, R, \theta)$。其中t为根结点空间位置坐标，R为根结点的旋转欧拉角，θ为剩余 20 个结点的旋转角度。另外，为了对每个用户的手部做针对性检测提高准确率，需要预先得到用户手部各个骨骼的长度信息，这里通过一段 30 帧的特定姿势的图像输入来获取。

有了骨骼模型后，采用一个能量函数来表达拟合的误差：

$$E(\Theta) = E_{2D}(\Theta) + E_{3D}(\Theta) + E_{limits}(\Theta) + E_{temp}(\Theta)$$

对于 2D项，计算heatmap上的极值点与对应结点投影到heatmap上的平面距离：

$$E_{2D}(\Theta) = \sum_j \omega_j ||\Pi(M_j(\Theta)) - u_j||_2^2$$

其中$M_j(\Theta)$表示第j个结点的空间坐标，Π表示平行投影，u_j表示极值点。文中提到了 2D项对于 3D预测结果是必要的，因为下面的 3D项只计算了 3D的结点相对根结点的坐标，结合 2D项后才可以拟合绝对空间坐标。

对于 3D项，计算某一角度集合下手部骨骼模型的用户相关的结点位置与RegNet输出的结点位置的距离，可以用来解决仅有 2D结点位置时产生的歧义问题。

$$E_{3D}(\Theta) = \sum_j ||M_j(\Theta) - M_{root}(\Theta) - z_j||_2^2$$

其中z_j为用户相关的结点坐标：

$$z_j = z_{p(j)} + \frac{||M_j(\Theta) - M_{p(j)}(\Theta)||_2}{||x_j - x_{p(j)}||_2}(x_j - x_{p(j)})$$

其中$p(j)$为j的父结点。z实际上就是根据RegNet输出的x作为骨骼的方向向量，再结合用户相关的骨骼长度数据来得到用户相关的节点位置。

对于limits项，由于手部的各个关节的运动自由度受到一定的限制，这里对超过旋转范围的情况作出惩罚：

$$E_{limits}(\Theta) = ||max(0, \theta - \theta^{max}, \theta^{min} - \theta)||$$

对于temp项，为了保证对于视频流的输出结果中每一帧的连贯性，这里尽量减少每帧间梯度的差异：

$$E_{temp}(\Theta) = ||\nabla\Theta^{prev} - \nabla\Theta||_2^2$$

我们对能量函数使用梯度下降法求使得它最小的Θ参数值，得到最终的手部姿态估计结果。

3. 实验与结果

为了评价检测的准确率，引入PCK（Percentage of Correct Keypoint）分数作为评价标准：定义一个正确的关键点，若检测结果落在关键点的临域（球或者圆）中时认为检测结果正确。在图 4 的左侧显示了采用不同训练数据集后（以及加入投影层后）得到的准确率结果：明显发现在仅有生成数据（不进行CycleGAN转化）的情况下准确率比最好的情况低了一半，这说明仅靠计算机绘制生成的训练样本与实际的真实样本分布存在很大差异。对比红色曲线可以发现，这种分布差异有很大一部分是体现在图像的背景上的，解释了我们替换背景来丰富训练集的必要性。在引入CycleGAN转化后，发现准确率明显提升，在加入投影层后，准确率进一步提升，说明采用生成中间 3D结果的RegNet是有利的。作者给出的解释是投影层加强了 2D与 3D结果的一致性。

在图 4 的中间给出了该方法与其他手部姿态检测方法在Stereo数据集上的 3D PCK准确率比较。其中Ours w指在数据集中拿 80%训练，拿 20%测试的结果，是所列出方法中最好的。Ours wo表示不采用Stereo数据集作为训练数据，直接作为测试数据后得到的准确率。可以发现仍然保持 80%左右的准确率，说明该方法的通用性。

在图4的右侧给出了在Dexter+Object与EgoDexter数据集上的 2D PCK准确率。由于这些数据集不提供手部根结点坐标，该方法不能输出结点的 3D绝对位置。可以看到在这两个个数据集上该方法仍然领先。

但是该方法也存在一定限制和不足。首先对于背景和手部的颜色相似的情况结果较差，另外当多只手同时出现并且距离较近时，该方法也会变得不稳定。最后，受限于输入数据有限，该方法仍然无法超过目前的RGB-D方法。

4. 总结

这篇文章的亮点在于使用CycleGAN生成训练数据用于CNN检测网络的训练，避免了真实数据采集困难的问题。这篇文章并没有给出仅采用真实数据训练后的准确率实验结果，所以我们很难判断这里使用的CycleGAN转化生成的伪“真实”图像与真实图像分布之间到底有多少差异。但是从目前的实验结果来看，CycleGAN确实对准确率有很大提升。另一方面，为了弥补RGB-only数据中深度信息的确实，作者创新地使用了平行投影层，并让网络同时输出 2D与 3D预测结果，最后综合考虑所有预测结果来确定手部姿势。目前看来这一弥补措施与RGB-D方法还有差距，仍有提升空间。

参考文献

[1] Mueller, Franziska, et al. "Ganerated hands for real-time 3d hand tracking from monocular rgb." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[2] Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE international conference on computer vision. 2017.

[3] De Boer, Pieter-Tjerk, et al. "A tutorial on the cross-entropy method." Annals of operations research 134.1 (2005): 19-67.

[4] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems. 2014.

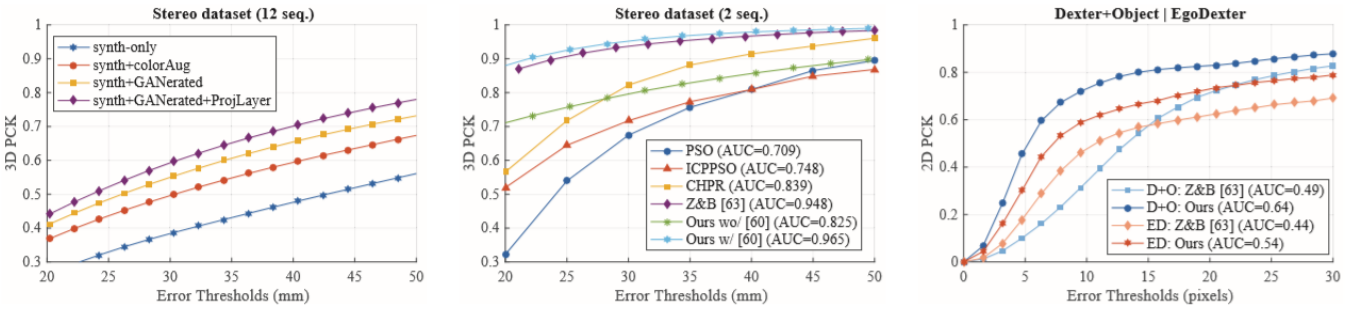

图 4 Results