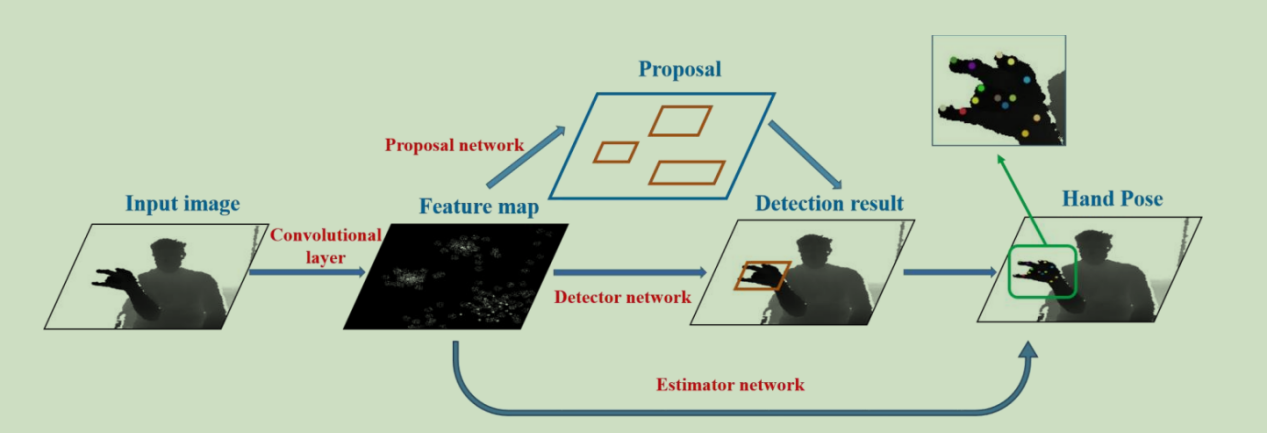
**Deep Learning for Integrated Hand Detection and Pose Estimation**

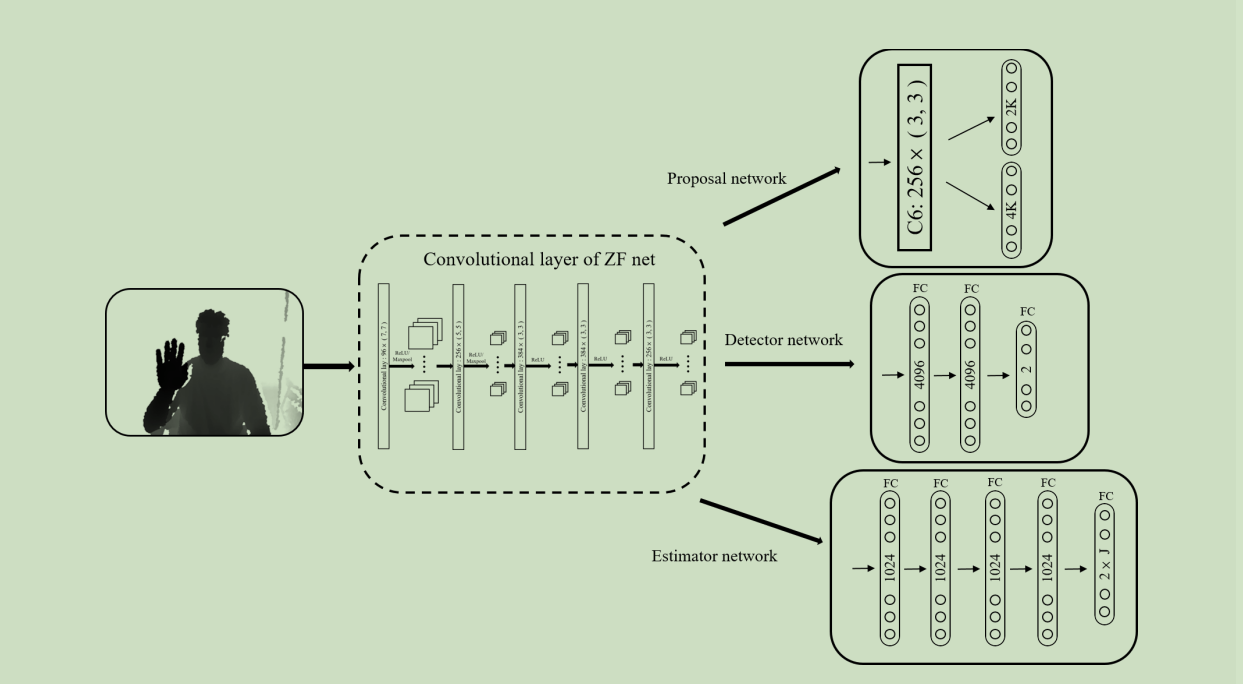


系统框架总览

该系统包含三个**卷积神经网络**(proposal, detector and estimator network)框架和一个**优化阶段**(optimization stage)。

**系统流程：**

首先,对原始深度图像做卷积运算得到一个特征图像（feature map）;然后,proposal network利用特征图像生成一系列手部的候选对象，detector network对生成的候选对象进行运算得到最有可能是手得对象，并在由原始图像得到的特征图像上标记出；接着, estimator network利用框选出的手部对象，回归出每个关节的坐标；最后，对这些回归出的位置进行优化。下图为神经网络框架:



**实验结果：**

准确率: 90%

误差: ±8.92 pixels.

速率: 74毫秒每帧(13.5FPS)

**Depth-based 3D Hand Pose Tracking**

**本文所做的主要贡献**

第一个贡献是一种新的检测跟踪策略，通过考虑使用RNN的预测历史结果，将基于CNN的单帧检测方法扩展到多帧跟踪方法。第二个贡献是使用RNN模拟3D模型对输入数据的拟合。

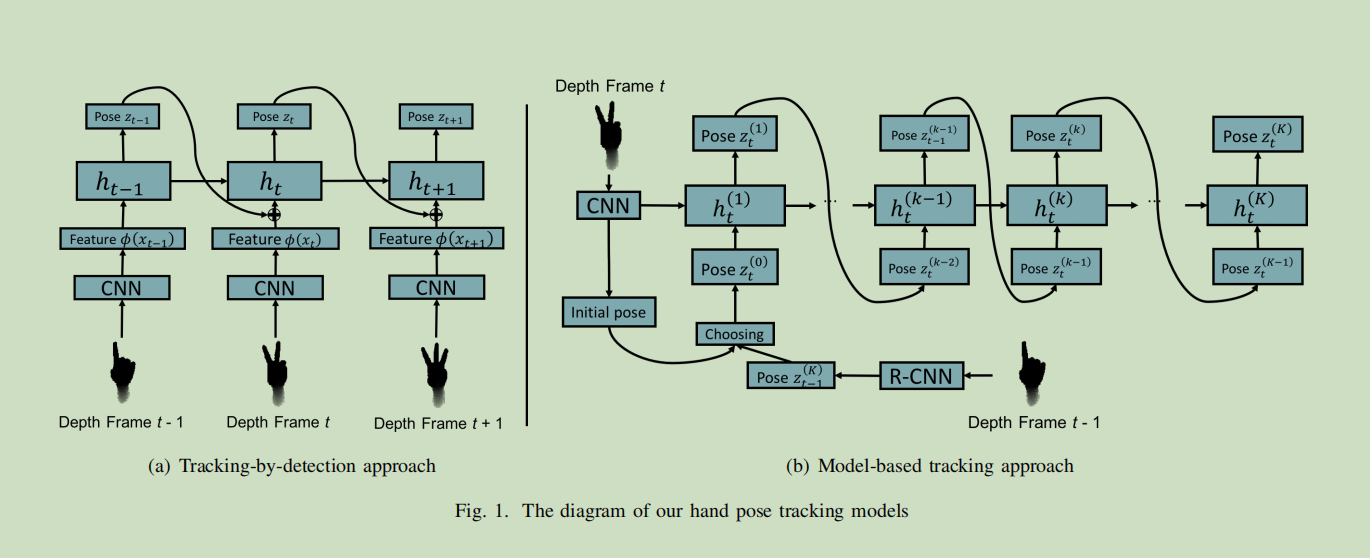
**本文提到的其他方法**

**基于模型生成的手部跟踪**(the model-base generative tracking)和**基于判别式手势检测**(the discriminative hand pose detection)。各自的缺点如下:

基于模型生成的手部跟踪: 当检测器在发生阻塞时导致模型初始化失败，跟踪框架无法从初始错误中恢复。

基于判别式手势检测（自己添加的）：鲁棒性不强，难以处理自遮挡的情况。

而本文则是结合上述的两种方法创建一种更强健的监测策略。



手姿势跟踪模型框架图

**框架说明：**

第一个部分是RNN输出根据CNN生成的图像的视觉特征和RNN的记忆状态（记忆的先前的帧特征）计算出的手姿态预测(如上图(a)所示)。换句话说，RNN起到记忆来自先前帧的手姿势预测的作用。所以，基于CNN的手部姿势估计的结果现在受到先前帧而不是当前输入帧的影响。最后，预测可以选择性地强调输入中的关注区域，然后在下一帧中将其馈送到卷积网络中。上图（ａ）所示，生成每一帧的特征都需要结合前一帧的特征。

第二个部分是一种新的基于生成模型的跟踪，其中RNN中的每个时间步骤可被视为传统的基于模型的跟踪中的有效近似模型拟合步骤。从训练数据中学习相似度函数并将其集成到RNN的学习中。CNN用于从当前深度帧中提取特征，并在必要时产生良好的初始手势（如图1（b）所示）。上图（ｂ）中，深度图像中的当前帧和上一帧同时作为网络的输入，然后网络利用两副图的特征回归出当前帧中的位置。

**实验结果：**

准确率: 89-99%

误差: 20~25mm

**Robust 3D Hand Pose Estimation in Single Depth Images:**

**from Single-View CNN to Multi-View CNNs**

**本文所做的主要贡献**

将单幅深度图像投影到三个正交的平面，然后并行训练三个CNN以将每个视图的投影图像映射到其相应的热图，然后将其融合在一起以估计3D手部关节位置。

**本文提到的其他方法**

最常见的手部姿势估计技术可以分为**模型驱动方法**和**数据驱动方法**两大类。

**数据驱动方法(Data-driven)。**用于手部位置估计的**数据驱动型(Data-driven)**方法主要有以下几种:等距自组织映射(isometric self-organizing map)、随机森林(random forests)、卷积神经网络(CNN)。这些方法的主要目的是将图像特征映射到手姿势参数。数据驱动型方法的优点:不需要复杂的模型校准(不需要对手的大小和运动约束进行明确的说明)，并且受不良初始化的影响较小。基于CNN的手姿态估计方法的主要困难在于精确的3D手姿势回归。从输入图像到三维位置的直接映射是高度非线性的，具有较高的学习复杂度和较低的网络泛化能力。一种替代方式是将输入图像映射到一组热图，这些热图表示图像中关节位置的概率分布，并通过模型拟合从深度图像恢复关节的三维位置。在该方法中，热图仅提供手关节的2D信息，并且未充分利用深度信息。所以本文使用多视图投影的方法尽可能多的利用深度信息。

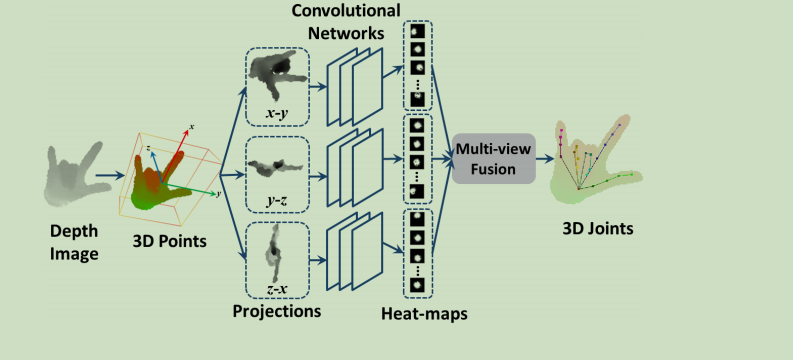
**模型驱动方法(Model-driven)。模型驱动**的方法通常用输入图像观测值来拟合可变形的三维手模型，从而找到最佳的手姿态参数，这种方法已经证明是非常有效的，尤其是在深度相机中。同时此类方法存在一定缺陷；例如，使用此类方法需要明确定义手的解剖尺寸和手部运动约束，以匹配输入图像。同时，由于手部姿态参数的高维性，使得迭代模型拟合过程收敛到最佳姿态对初始化很敏感。

**本文的方法描述**

首先，将输入的单幅深度图像投影到三个正交的平面上(x-y,y-z,x-z);然后,将每个投影图像送入是各自对应的CNN(一共三个)中，生成一组手关节热图。最后，通过将三个视图的热图与预先学习的手姿态先验(priors)融合，最终获得3D关节位置并同时减轻模糊估计。

**本文的优点**

1. 使用多视图 CNN可同时生成正面、侧面和顶部视图的热图，从中可以估计更为准确的手关节的 3D位置。
2. 受模糊估计的影响更小。在模糊估计的情况下，单视图CNN不能很好地区分热图中的多个热点，其中只有一个可以对应于真实的关节。 利用提出的多视图CNN，使用来自其他两个视图的热图可以帮助此消除模糊性。
3. 使用三维视图投影能产生一个潜在的运动学约束，因此无需手动定义手部的大小参数。



系统总览图

**实验结果：**

准确率: 95~98%

误差: 14~18±15%mm

速率: ≥70FPS

**DeepHand: Robust Hand Pose Estimation by Completing a Matrix Imputed with Deep Features**

**本文所做的主要贡献**

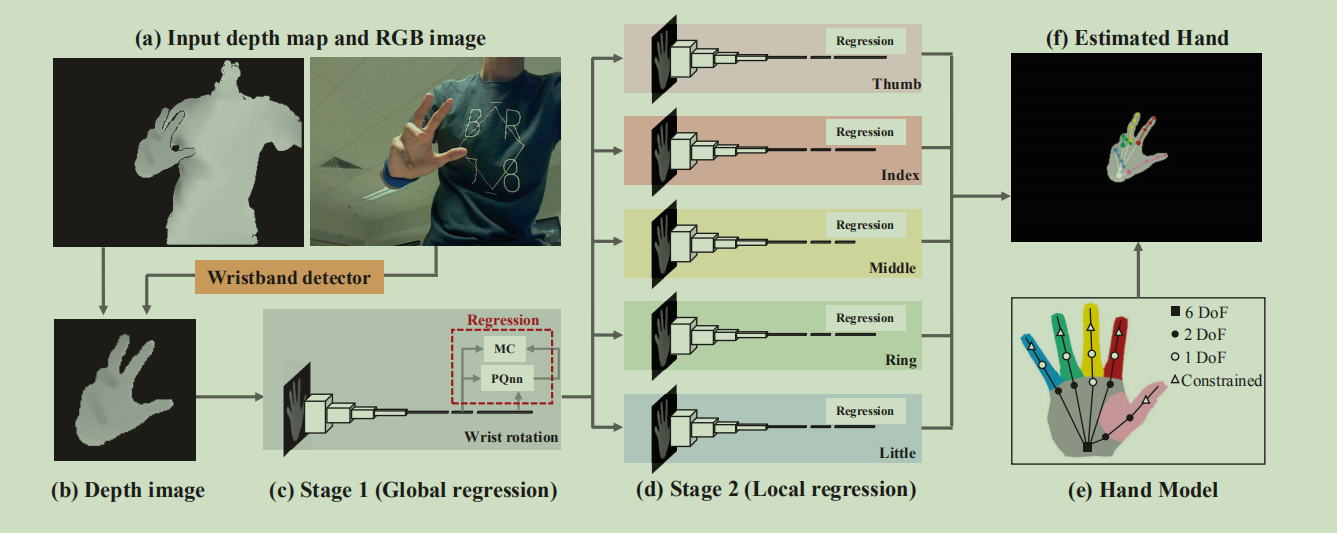
1. 使用deep ConvNets,使用全局方向或手指关节的低维度和判别式表示作为激活特征来初始化姿势矩阵，其有助于使用量化的生成结果从大量预先计算的激活特征中有效地检索最近的临近值(neighbors)。
2. 提出了一种使用初始化姿势矩阵估计关节角度参数的有效矩阵完成(completion)方法(An efficient matrix completion method for estimating joint angle parameters using the initialized pose matrix.)。
3. 在标准计算机上保持实时帧速率的同时，以一种原则性的方式将全局姿态方向和手指关节结合在一起的一种用于手势估计的层次结构管道结构。

**手部模型描述**

本文使用具有21自由度（DOF）的运动学手模型作为手势估计的标准(见下图e)，表示为H（θ，φ）。θ表示18个关节角度参数的集合，φ是手的3个全局平移参数（x，y，z）的集合。

**本文的方法描述**

首先，我们使用深度图和基于RGB的腕带检测器(下图ａ,b部分）提取包含手部的区域。 然后，将获得的深度图像被馈送到ConvNet，输出激活特征。 此激活特征使用矩阵完成方法与填充数据库中的其他特征同步，并估计全局姿势参数(下图c)。 基于该全局姿势初始化，我们以相同的递归方式(下图d)估计其余的局部关节参数。 最终的手姿势显示在多媒体屏幕上(下图f)。

****

系统整体框图

**实验结果：**

准确率: 97%

误差: 16.35mm

速率: ≈32FPS

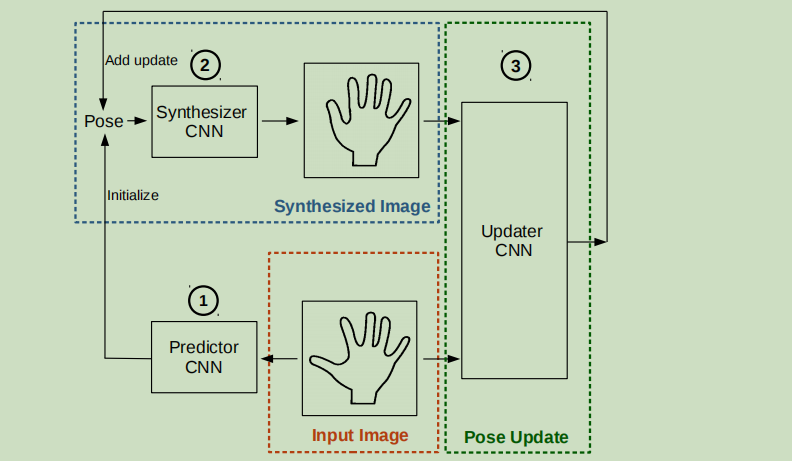
**Training a Feedback Loop for Hand Pose Estimation**

**本文所做的主要贡献**

1. 从原始的深度图像中学习，以生成逼真的手部深度图像，用来进行反复迭代，用于细化初始姿势估计和纠正前面的产生的错误。
2. 使用了三个结构较为简单的CNN，大大提到了预测时的速度。
3. 不使用手工制作的相似度函数和优化算法。 我们学习的是预先更新，以改善当前手的估计，根据输入深度图像和训练数据进行姿势为此估计生成的图像

**本文的方法的整体描述**

首先,我们将单幅深度图像输入到Predictor CNN(下图①)中用于预测深度图像的三维姿势的初始估计。 然后，将该初始姿势的估计输入到Synthesizer CNN(下图②)中用于合成一副含有三维姿势信息的图像。接着,将含有三维姿势信息的图像和原始深度图像送入到Updater CNN中用于输出改进后的姿势估计，再将该输出反馈到Predictor CNN再次进行姿态合成。最后，重复迭代上面的步骤数次即可得到较为准确的姿势预测。



系统整体框图

**实验结果：**

准确率: 97%

误差: 16.5~27mm

速率: ≥400FPS

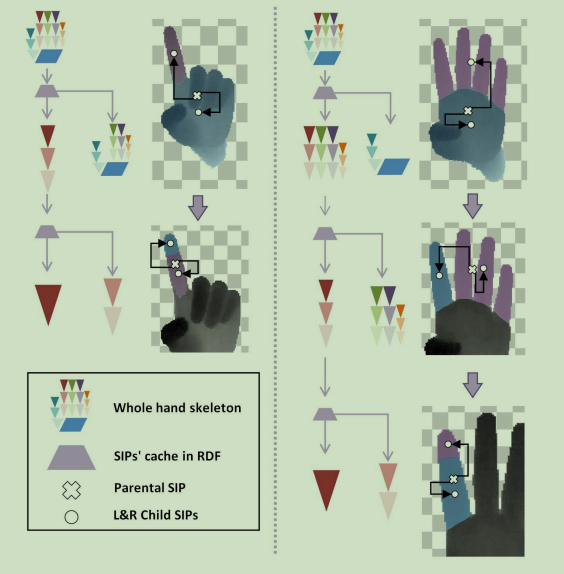
**3D Hand Pose Estimation Using Randomized Decision Forest with Segmentation Index Points**

**本文所做的主要贡献**

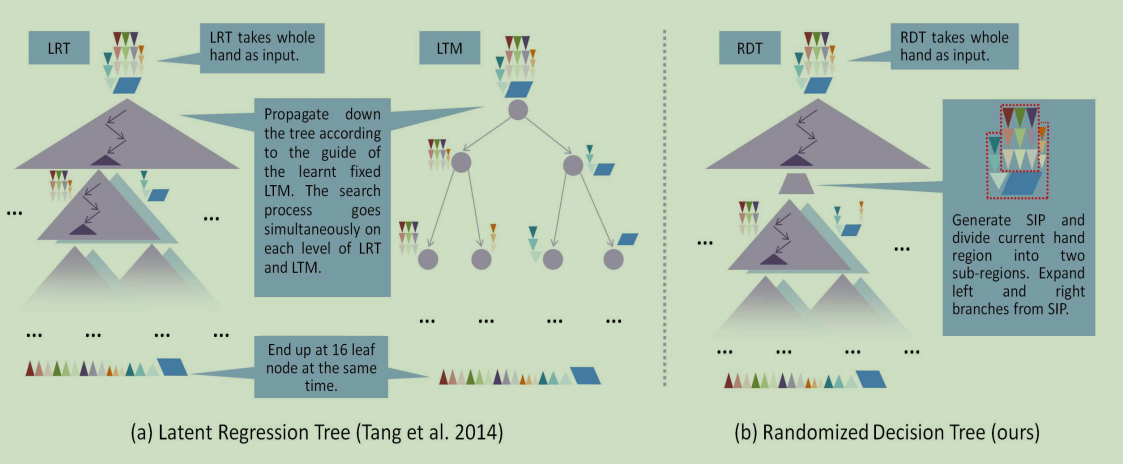
1. 提出了SIP的概念(segmentation index points,分段索引点)。SIP表示骨架关节子集的质心，其位于从SIP直接或递归扩展的分支的叶子处。 我们将SIP集成到框架中，有两个主要改进：首先，我们设计了一种新的森林生长策略，其中SIP用于自适应地引导骨骼关节分区和随机特征选择。 其次，我们加快了训练过程，因为在非叶节点中估计最多两个SIP而不是骨骼关节。
2. SIP的使用使得决策树能够将最终的骨架定位推迟到后期（即终止叶子），从而避免为更精确的决策积累更多的信息。
3. SIP基于三维骨架关节的聚类结果以无人监督的方式训练，因此对不同的姿势变化是自适应的。
4. 新的森林培育程序由一个随机的独特二元特征指导，这取决于SIP，它不依赖于特定的手模型。

**本文的方法的整体描述**

此算法可以看作是一个分而治之的搜索过程。 我们递归地聚类骨架关节并将探测中的部分划分为两个更精细的子区域，直到到达叶节点，该叶节点表示骨架关节的位置。 该图显示了定位食指尖的两个示例。 不同的手势导致不同的手子区域分割，因此不同的SIP和不同的树结构。下图显示了定位食指尖的两个示例．以左边的食指查找过程为例讲解，首先，找到手点云集群的质心作为父SIP;然后,以父SIP可以将手部分为两个子SIP(即食指的部分的质心和手掌部分的质心),接着将子SIP中可继续分割的SIP作为新的父SIP,继续进行上述的操作,直到到达叶节点，该叶节点即表示对应关节点的位置．反复迭代多次即可找到所有的关节点．

****

定位食指的示例

**** 与其他随机森林算法的不同之处

**实验结果：**

准确率: 96%~98%

误差: 未提到

速率: 55.5~62.5FPS(single CPU)

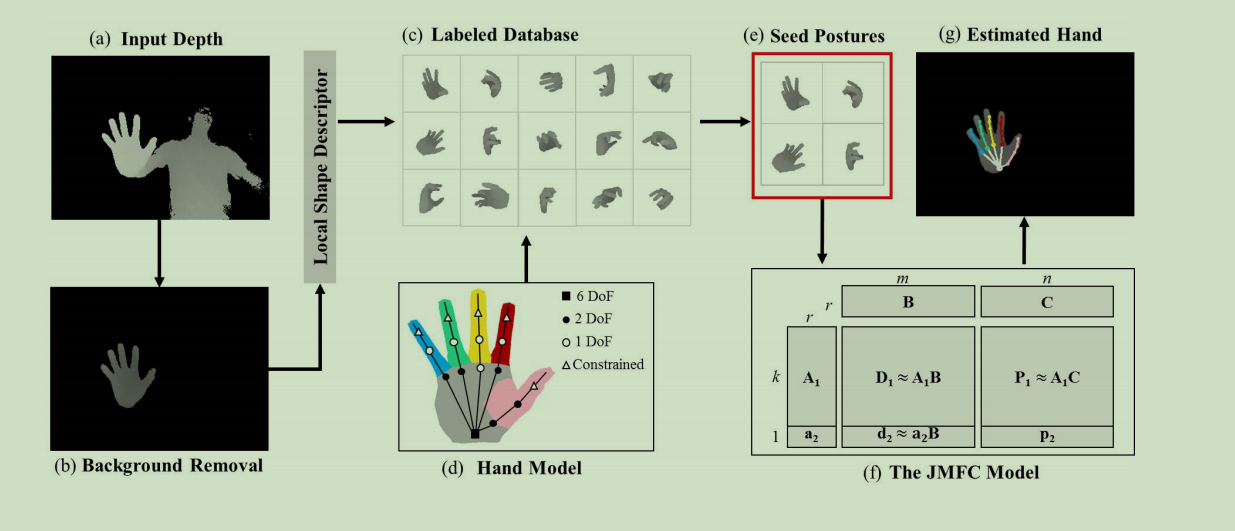
**A Collaborative Filtering Approach to Real-Time Hand Pose Estimation**

**本文所做的主要贡献**

1. 提出联合矩阵分解和完成（joint matrix factorization and completion,JMFC）算法，以基于每帧估计来自最近邻居的未知姿势参数。
2. 使用模仿真实三维手势的合成手模型构建手姿势库。
3. 通过使用姿势聚类，FAST特征点检测器和简要描述符的组合，从姿势库中检索有效的最近邻居。
4. 能实现实时手姿态估计，且能使用未经训练的参数。

**算法概述**

首先,去除深度图中的背景噪声,然后提取出手部区域（（a） -> （b））。然后, 我们使用局部形状描述符从各种手配置的标记数据库中检索最近的neighbors（（c） -> （d））。接着,将提取的neighbors作为JMFC模型的种子姿势，使用矩阵分解和完成过程估算未知的联合参数（手的整体姿势）（（e） -> （g））

****

系统整体框架

**实验结果：**

准确率: 85%~90%,98%~100%

误差: 18~30ｍｍ

速率: ≈29FPS

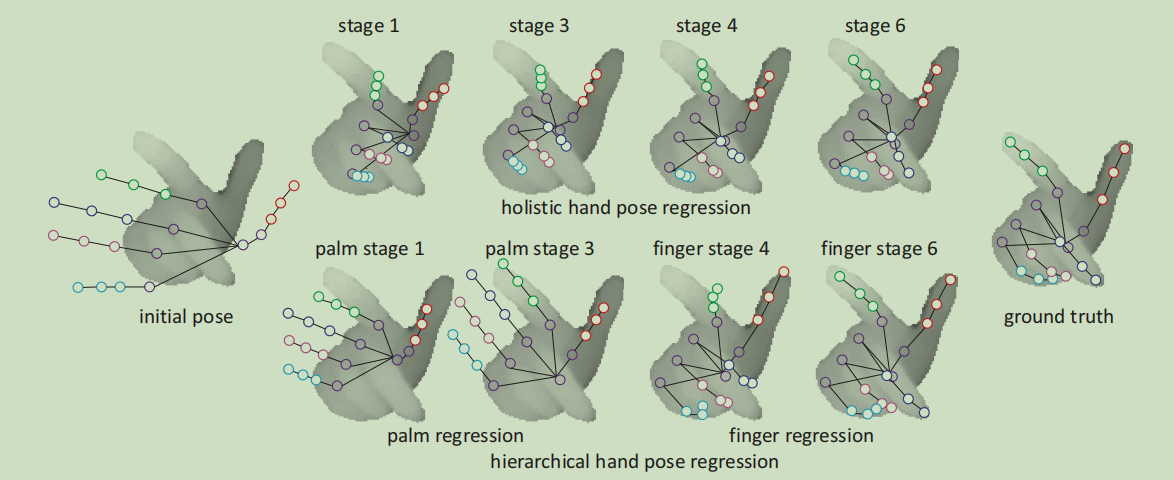
**Cascaded Hand Pose Regression**

**本文所做的主要贡献**

1. 提出了一种级联回归方法，该方法在大视点和复杂的手势下更加稳健。 它由级联姿势回归框架直接驱动，其中通过一系列弱回归量逐步估计对象姿势，并且每个弱回归量使用依赖于前一阶段的估计姿势的特征。这种姿势索引特征提供了更好的几何不变性并简化了学习任务.
2. 扩展了3D关节对象的框架。 提出了三维姿势索引功能。
3. 我们的第二个贡献是一种原则性的分层方法，适用于铰接对象的结构。由于手是铰接结构，不同的物体部分通常表现出不同的变化量和自由度。因此，将所有部件一起回归是不必要的，并且导致收敛慢和精度降低。我们的分层方法按照其清晰度复杂度的顺序依次对不同部分的姿态进行回归。首先估计更容易根部（例如手掌）的姿势。然后对更难的子部分（例如手指）的估计进行条件化。根部的姿势因此更容易。分层方法不仅收敛速度快，而且更准确。

**回归过程**

首先,从深度图像中得到一个粗略的初始手姿势；然后，手姿势通过六个阶段迭代更新直至接近真实的三维手部姿态．第一行是整体回归，先对手掌进行整体回归，接着是对手指进行整体回归；第二行是分层回归，先对手掌进行回归，然后对手指进行回归，第而行时对第一行回归的细化．

****

级联分层回归结构框图

**实验结果：**

准确率: 95%~100%

误差: 7~10ｍｍ

速率: >300 FPS(CPU, single thread)