前言

人体动作捕捉技术（简称人体动捕技术）是影视游戏行业中常用的技术，其可以实现精确的人体姿态，运动捕捉，但是用于此的设备昂贵，很难在日常生活中广泛应用。视频人体动作捕捉技术指的是输入视频片段，捕捉其中场景中的人体运动信息，基于这种技术，可以从互联网中海量的视频中提取其中的人体运动姿态数据，具有很广阔的应用场景。本文打算介绍视频人体动作捕捉相关的一些工作并且笔者的一些个人看法 。 **如有谬误，请联系指出，转载请联系作者，并且注明出处。谢谢。**

∇ 联系方式： **e-mail**: FesianXu@gmail.com **QQ**: 973926198 github: https://github.com/FesianXu

人体动作捕捉技术

人体动作捕捉技术，简称人体动捕技术（Motion Capture, Mocap）是通过某些传感器，捕捉场景中人体运动的姿态或者运动数据，将这些运动姿态数据作为一种驱动数据去驱动虚拟形象模型或者进行行为分析。这里的传感器既可以是惯性传感器IMU，红外光信标（也可能是会发出红外光的摄像机），也可以是RGB摄像头或者是RGBD摄像头等。根据人体是否有佩戴传感器和佩戴的传感器是否会主动发送定位信号，我们可以把人体动捕技术大致分为：

1. 被动形式的人体动捕技术
2. 主动形式的人体动捕技术

被动形式的人体动捕技术

被动形式人体动捕：如Fig1.1和Fig1.2所示，此时人体佩戴的是会反射特定红外激光的光标，而在场景周围部署多个红外激光摄像头，这类型的激光摄像头会主动向人体发射特定的红外激光，该激光打到光标上会反射，摄像头通过接受这些反射光计算每个光标的空间位置。

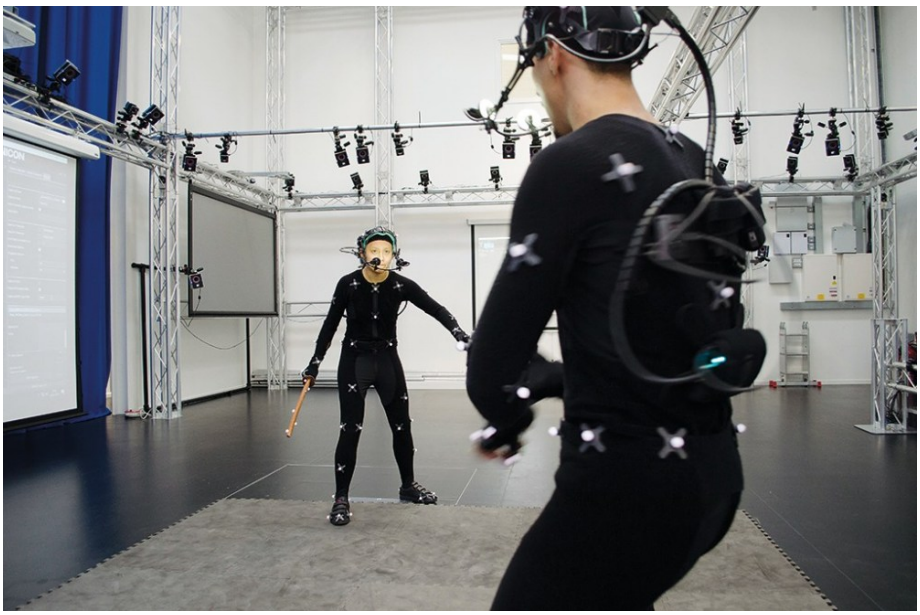


Fig 1.1 人体佩戴着能够反射特定红外激光的光标，周围部署多个特定的红外激光摄像机。

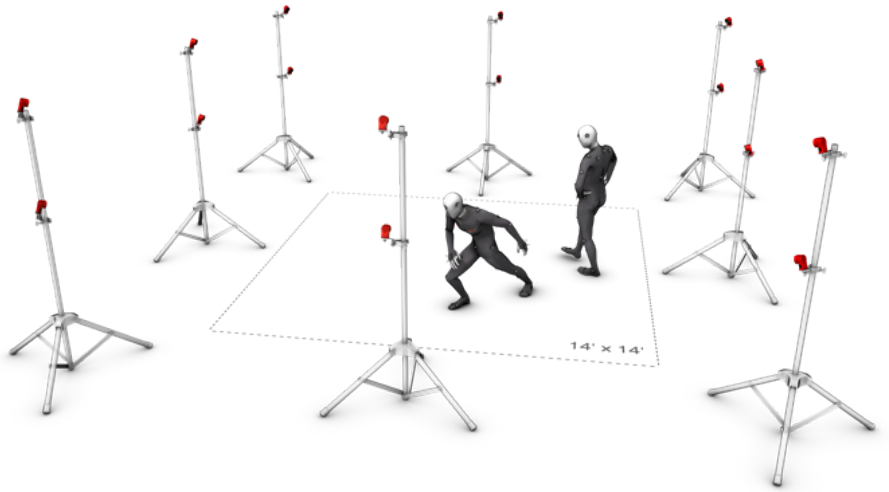


Fig 1.2 基于红外光标的方案需要在场景周围部署多个特定的红外激光摄像头，成本高昂。

当然，现在基于RGB图片的人体姿态估计技术已经趋于成熟，目前已经有着很多很棒的相关工作[4,5]，我们以2D人体姿态估计为例子，通过部署多个经过相机标定的摄像头，我们在每个摄像头上都进行2D人体姿态估计得到每个人体的关节点，通过多视角几何，我们能够计算出关节点的空间位置。通过这种技术，人体不需要佩戴传感器或者光标，能够实现比较轻松的动作采集。这种技术我们称之为 **视频人体动捕技术** ，其中采用了多个摄像头的我们称之为 **多目视频人体动捕技术**，如果只有单个摄像头，我们则称之为 **单目视频人体动捕技术**，而这也是本文的主要需要介绍的内容。

主动形式的人体动捕技术

主动形式的人体动捕技术需要人体佩戴特定传感器，这些传感器或可以自己发射特定的激光信息到周围部署的摄像头，实现多目定位，或可以通过牛顿力学原理计算初始位置记起的每个时刻的空间位置状态（我们称之为惯性导航），如Fig 1.3所示，该类型的方案通常比 被动形式的方案要精准，但是要求人体佩戴更为

昂贵的专用设备，因此场景还是受限于大规模的影视游戏制作上。（虽然基于IMU的设备通常比红外激光设备要便宜，但是精度不如基于光学定位的，而且容易受到环境磁场的干扰）

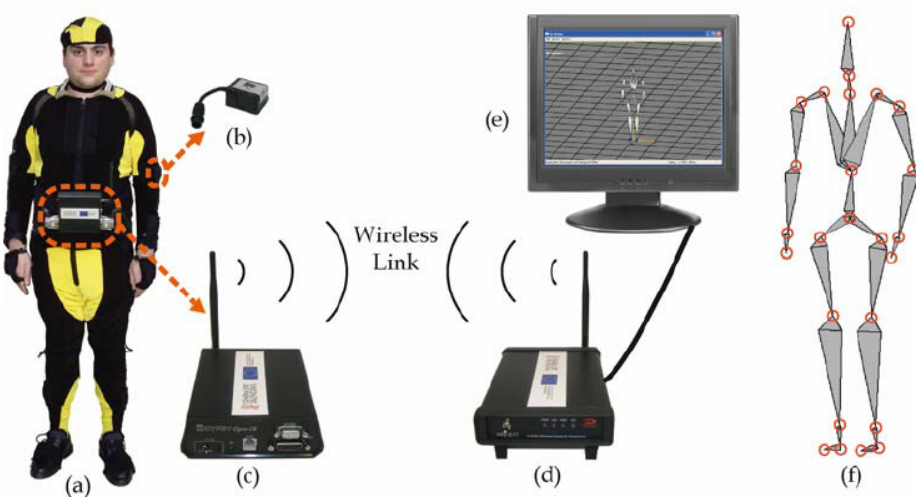


Fig 1.3 人体佩戴主动式传感器，比如IMU，进行姿态相关的数据采集（比如旋转角，地磁，加速度等），然后计算出相对于初始位置的末态空间位置。

说回单目视频

单目RGB视频占据着目前互联网的大部分流量，是名副其实的流量之王。以人为中心的视频占据着这些视频的很大一部分，如果我们能较好的从其中提取出人体的动作和姿态信息，将能够很大程度上帮助我们进行更好的行为分析和虚拟动作生成，甚至可以在一定程度上取代之前谈到的依赖于专业设备的人体动捕技术，在影视游戏制作上助力。当然，基于单目视频毕竟存在某些信息的损失，比如自我遮挡，投影歧义性等，为了解决这些问题，或多或少要引入某些先验知识[21,22]。

单目视频人体动捕技术

从单目视频里面提取人体的动作信息，我们首先要知道什么称之为动作信息（motion）。在影视游戏领域，动作信息通常指的是 **每个关节点相对于其父节点的旋转（rotation），以及整个骨架的朝向旋转（orientation）和偏移（translation）** 。因为这个“动作信息”的定义贯穿着这整篇文章，因此笔者需要进行更为细致的介绍。

动作信息

在影视游戏领域，我们经常会要求动画师设计一个虚拟形象模型，如Fig 2.1所示。为了设计某些特定动作，我们通常会基于某个初始状态（一般是人体呈现T字形，称之为Tpose）进行修正得到最终需要的动作。为了实现这个修正，我们需要描述模型的每个关键部位，通常用关节点表示，如Fig 2.1所示，其中pelvis

节点我们一般视为整个骨架的根节点，而其他的连接节点如果看成多叉树的话，就是子节点。这种树状结构，我们称之为 **关节点树**。

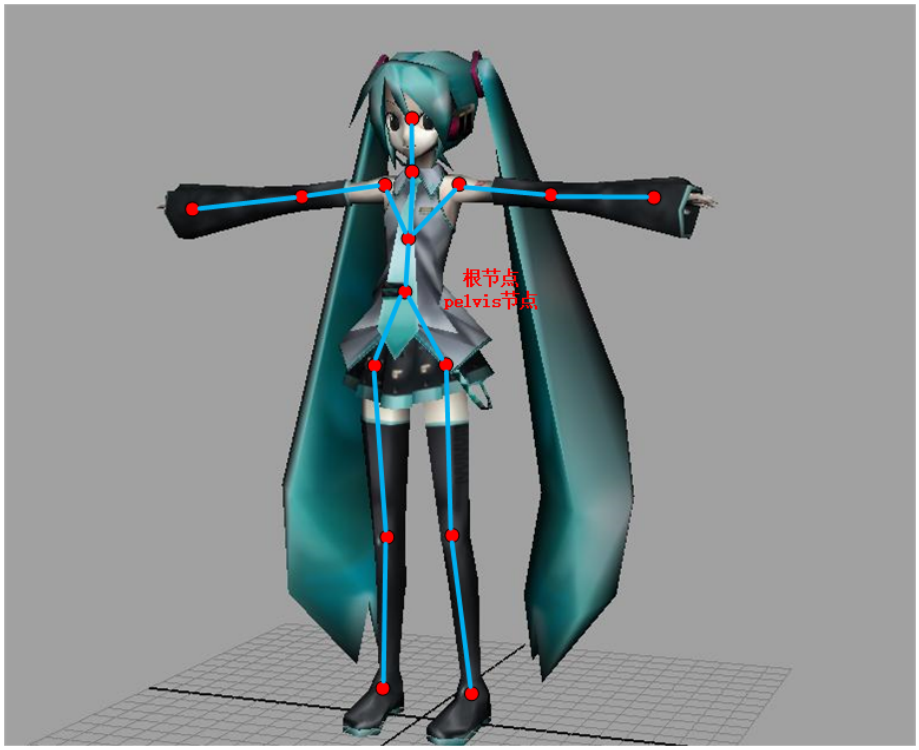
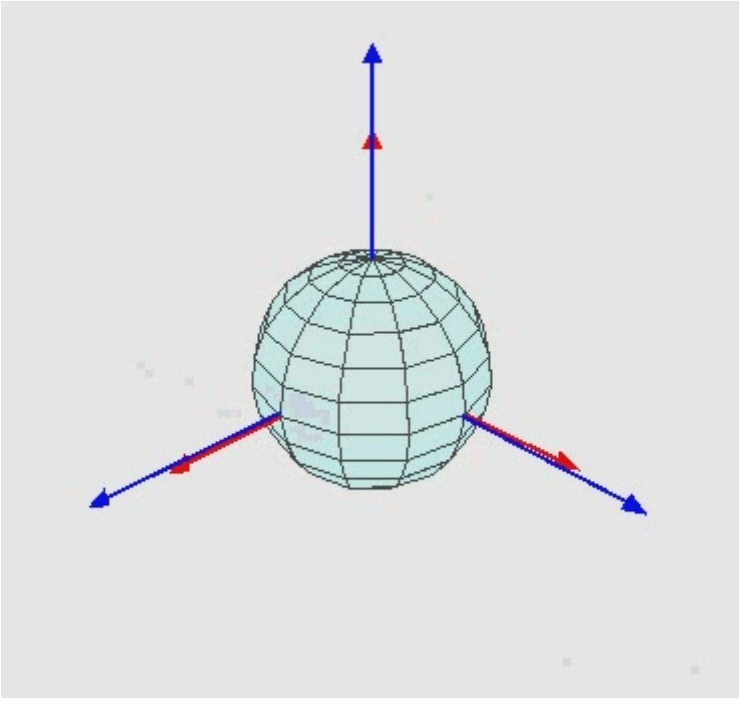


Fig 2.1 虚拟形象 初音未来 的Tpose模型，如果给虚拟模型绑定关节点，关节点的移动和旋转会带着蒙皮跟随变化。

除了根节点之外，其他所有节点都只有旋转信息（一般用欧拉角描述[6]），如Fig 2.2所示，通过三元组的欧拉角（即是分别围绕着XYZ轴旋转特定角度）旋转，可以实现大多数的空间旋转操作，从而修改模型的动作。注意到我们对于某个关节点进行旋转操作，那么其关节树下的其他子关节点也会跟着一起运动，这个和我们的常识是一致的。说回到根关节点，其旋转信息表示的是整个骨架的朝向，并且根节点还有偏移信息，可以表示骨架在世界坐标系下的空间位置。



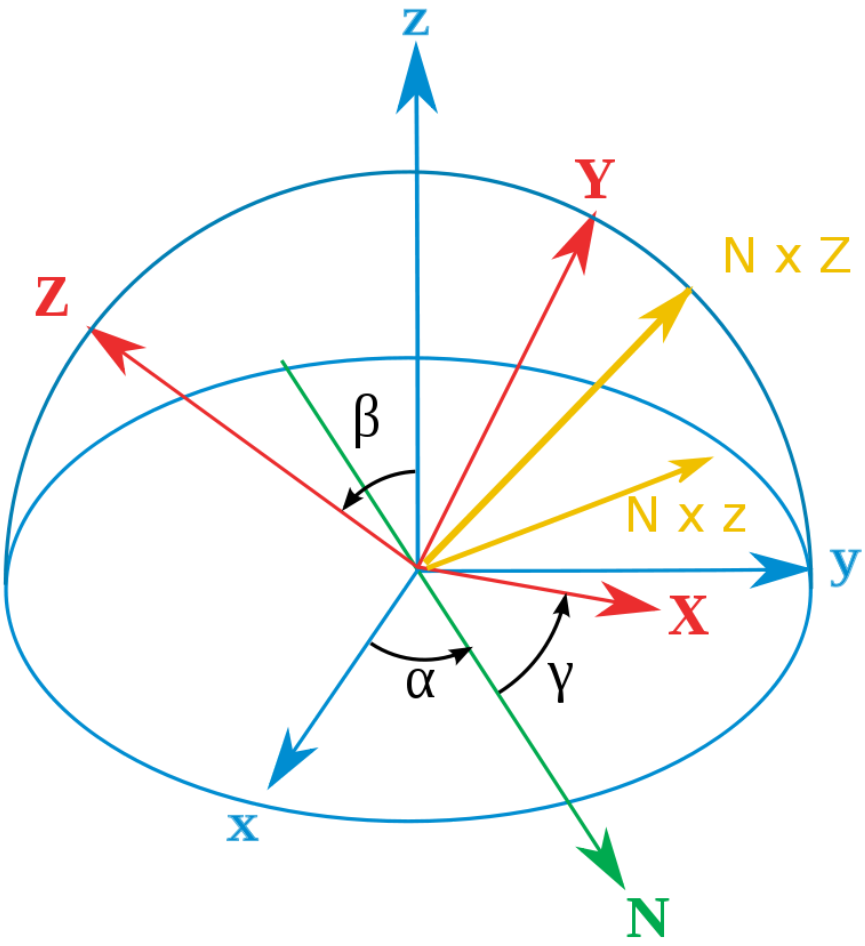


Fig 2.2 欧拉角，分别绕着**XYZ**轴旋转特定角度，可以实现大多数空间旋转（有时候会存在万向节的现象）。

因此，通过一系列的旋转信息和偏移信息，我们可以描述一个骨架的一系列动作。

基本技术路线

从技术大方向来说，可以有以下两种思路：

1. 先从单目图片中提取出人体的关节点空间坐标数据，通常用欧式坐标表示。然后通过反向动力学（Inverse Kinematics，IK）计算得到每个关节点的旋转信息。
2. 直接通过模型从单目图片中估计出人体每个关节点的旋转信息。

一般来说，第一种方案是一种两阶段的方法，需要借助人体姿态估计的方法，比如[4,5]等去估计出2D人体姿态后，通过一些方法[7]可以从2D姿态数据中估计出3D姿态数据。当然也可以直接估计出3D姿态数据[8]，而不用2D姿态作为中间输入。然而单纯的欧式空间坐标还需要转换成旋转欧拉角才能满足后续的使用要求，为了实现这种转换，需要采用反向动力学的方式，去约束计算得到每个关节的旋转信息。这个转换的过程中存在着若干问题：

1. 反向动力学约束求解旋转信息计算速度慢，而且可能存在多解的情况，不是特别鲁棒。
2. 即便采用了反向动力学[9]去计算每个关节的旋转，但是某些关节（特别是双臂，双腿等可以自我旋转的）的“自旋转”是无法描述的，比如Fig 2.3所示，我们的中间输入是3D人体姿态，无法表述这种自旋转，因此即便采用了IK，计算得到的旋转角也会缺少一个自由度。当然，这个也并不是完全不能解决，如果姿态估计的结果能够精确到手指，给出部分手指的关节点数据，那么通过IK的约束还是可以恢复出手臂的自旋转的自由度的。

Fig 2.3 某些动作只存在肢体部分的自我旋转，也就是只有一个自由度的旋转，这种旋转即便采用IK也不能计算到，这部分的信息是完全的损失了。

鉴于这些问题，我们的第二种方案尝试直接从单目图片中估计人体的每个关节点的旋转信息，这种方案是一种一阶段的方案，只需要输入RGB图片，就可以输出每个关节点的旋转数据，而且这种方案有一定的机制可以恢复出双臂，双腿的自旋转自由度，因此该方案是本文的主要讨论内容，细节留到后文继续讲解。（注解：准确说，这里的“恢复出”不准确，应该是通过添加先验，把一些明显人体不能做出来的动作进行了排除）

基于运动估计的动捕技术

在本节中，我们主要讲解技术路线中第二种方案，也就是直接从RGB图片中估计人体每个关节点的旋转数据，我把这种方法称之为 **基于运动估计的人体动捕** 。 为了实现这个技术，必须要引入数字人体模型。数字人体模型是将人体形象参数化，使得可以通过若干个参数去表征人体的形状，动作等特征，数字人体模型有很多，比如SMPL模型[10] (该模型只能表示人体的基本属性，比如形状，运动，没有手势，表情等细节的参数化)，SMPL-X模型[11] (该模型是SMPL模型的扩展，可以表述手势，表情等细节)，Total Capture模型[12] （同样也是可以表征人体的基本属性，并且有手势，表情等细节）。

我们在之前的博客中介绍过最为流行的SMPL模型[10]，在本文稍微在介绍一下，更多细节请移步博文[10]，如有该基础的读者可以省略该部分内容。

SMPL模型

SMPL模型用以参数化人体模型的基本属性，比如动作姿态，形状等，该模型在[13]提出，其全称是**Skinned Multi-Person Linear (SMPL) Model**，其意思很简单，Skinned表示这个模型不仅仅是骨架点了，其是有蒙皮的，其蒙皮通过3D mesh表示，3D mesh如Fig 3.1所示，指的是在立体空间里面用三个点表示一个面，可以视为是对真实几何的采样，其中采样的点越多，3D mesh就越密，建模的精确度就越高（这里的由三个点组成的面称之为三角面片），具体描述见[14]。Multi-person表示的是这个模型是可以表示不同的人的，是通用的并且可以泛化到各个不同的人体的。Linear就很容易理解了，其表示人体的不同姿态或者不同升高，胖瘦（我们都称之为形状shape）是一个线性的过程，是可以控制和解释的（线性系统是可以解释和易于控制的）。那么我们继续探索SMPL模型是怎么定义的。

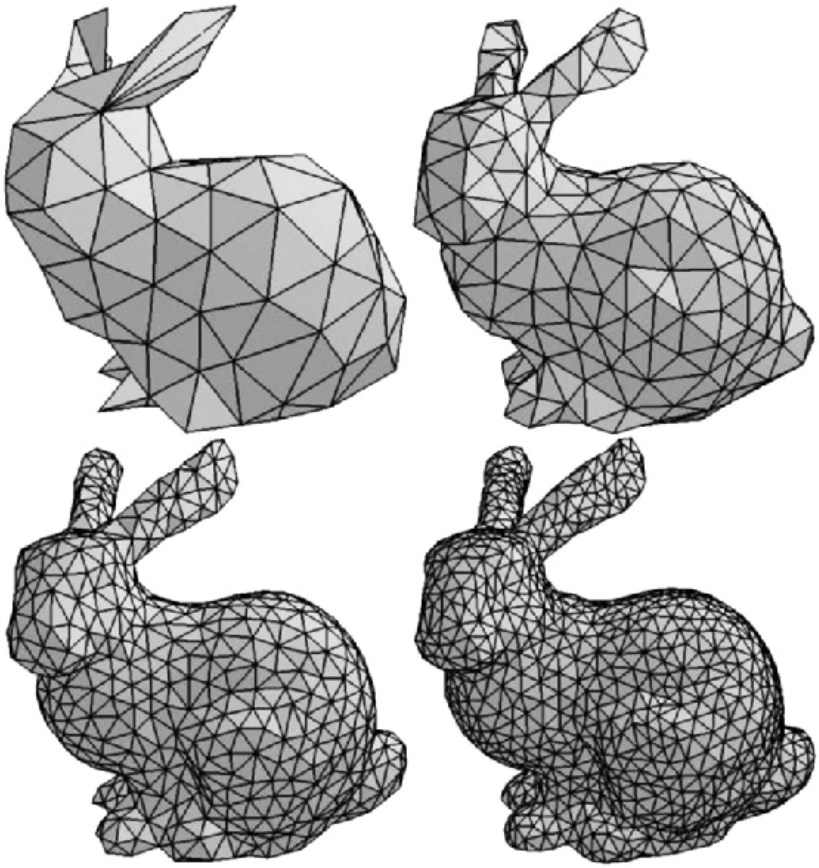


Fig 3.1 不同分辨率的兔子模型的**3D mesh**。

在SMPL模型中，我们的目标是对于人体的形状比如胖瘦高矮，和人体动作的姿态进行定义，为了定义一个人体的动作，我们需要对人体的每个可以活动的关节点进行参数化，当我们改变某个关节点的参数的时候，那么人体的姿态就会跟着改变，类似于BJD球关节娃娃[15]的姿态活动。为了定义人体的形状，SMPL同样定义了参数$\beta \in \mathbb{R}^{10}$，这个参数可以指定人体的形状指标，我们后面继续描述其细节。

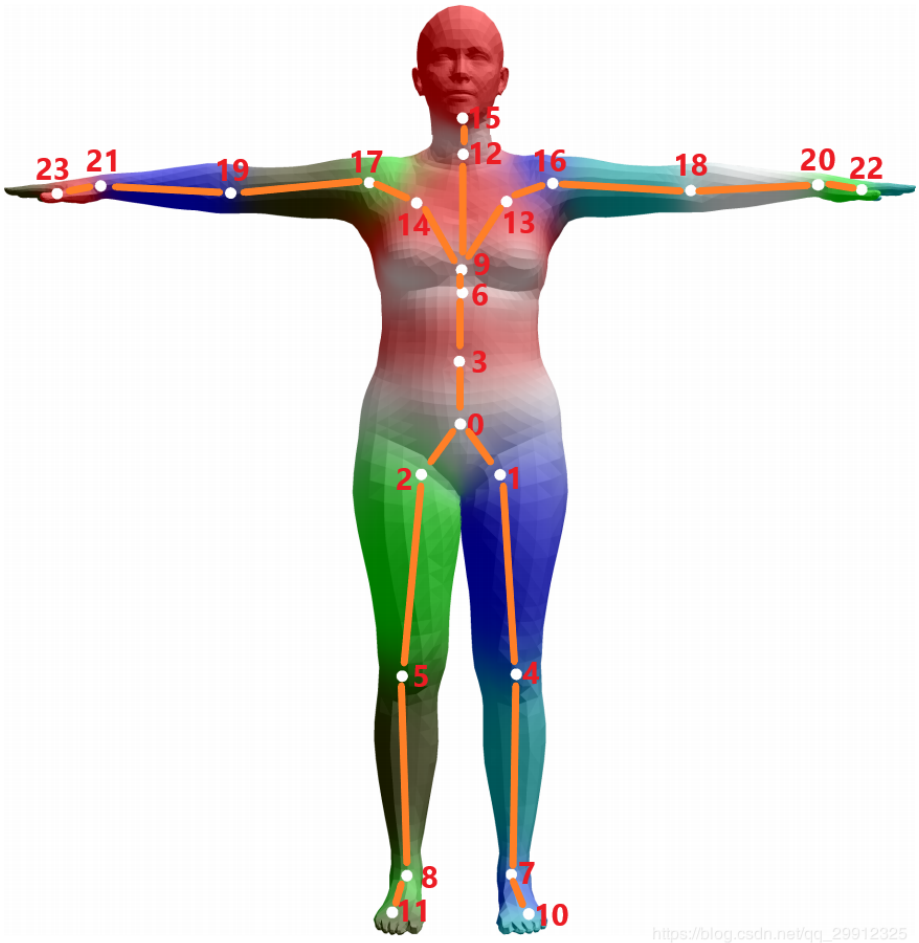


Fig 3.2 SMPL模型定义的24个关节点及其位置。

总体来说，SMPL模型是一个数字人体参数化模型，其通过两种类型的参数对人体进行描述，如Fig 6所示，分别有：

1. 形状参数（shape parameters）：一组形状参数有着10个维度的数值去描述一个人的形状，每一个维度的值都可以解释为人体形状的某个指标，比如高矮，胖瘦等。
2. 姿态参数（pose parameters）：一组姿态参数有着24×3维度的数字，去描述某个时刻人体的动作姿态，其中的24表示的是24个定义好的人体关节点，其中的3并不是如同识别问题里面定义的(x, y, z)空间位置坐标（location），而是指的是该节点针对于其父节点的旋转角度的轴角式表达(axis-angle representation)（对于这24个节点，作者定义了一组关节点树），当然，轴角式在计算中经常会换算成易于计算的欧拉角表达。

具体的β和θ变化导致的人体mesh的变化的效果图可视化，大家可以参考博文[16]和[17]。

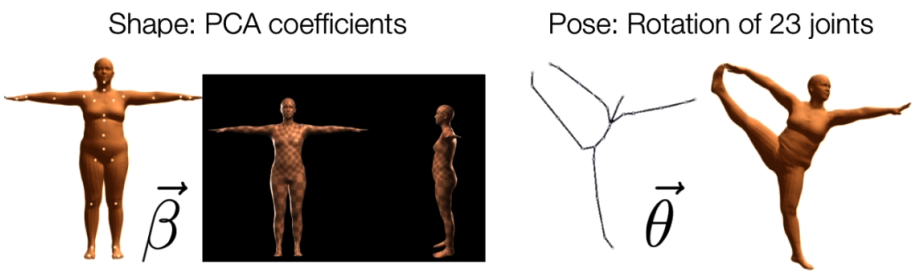


Fig 3.3 形状参数和姿态参数，原图出自[18]。

相信看到现在，诸位读者对于这种通过若干个参数去控制整个模型的姿态，形状的方法有所了解了，我们对于一个模型的形状姿态的mesh控制，一般有两种方法，一是通过手动去拉扯模型mesh的控制点以产生mesh的形变；二是通过Blend Shape，也就是混合成形的方法，通过不同参数的线性组合去“融合”成一个mesh。

基本技术路线

基于数字人体模型，我们只需要从RGB图片中估计出数字人体模型的参数即可，比如若使用SMPL数字人体模型，那么我们需要估计的参数有$69+10+3+3=85$个（这里的旋转使用了轴角式表达，因此只有3个参数，具体见[10]）：

1. 关节点的旋转信息，$\theta \in \mathbb{R}^{23\times 3}$
2. 人体的形态参数，$\beta \in \mathbb{R}^{10}$
3. 相机外参数，$\mathbf{t}_{\text{cam}} \in \mathbb{R}^2$， $s \in \mathbb{R}^1$， $\mathbf{R} \in \mathbb{R}^3$。

需要说明的是，我们一般假设渲染相机是弱透视相机[23,24]，意味着相机外参数有尺度缩放系数s和相对于场景的偏移$\mathbf{t}_{\text{cam}} = [\text{tx}, \text{ty}]$。需要注明的是，关节点中的根关节点，也就是Fig 3.2中的0号关节点的旋转信息是作为相机外参数看待的，表征了人体的朝向(orientation)信息，因此特别地，我们把该节点的旋转信息独立出，作为相机的旋转矩阵外参数，也就是有$\mathbf{R} \in \mathbb{R}^3$。

我们的基本思路就是通过模型去预测回归出SMPL模型参数。[19]中提到的HMR模型是一种经典的方法，如Fig 3.4所示，对于输入的场景，首先用检测算法对其中的人体位置进行确定并且裁剪得到人体的包围框。然后用Resnet-50 [25] 作为图片特征提取器，截取自最后的平均池化层的特征输出，得到图片特征$\phi \in \mathbb{R}^{2048}$。

为了更好地回归出SMPL模型参数，不能直接一次性地用网络回归出这些参数，而是通过迭代(iteration)的方式进行逐步的优化的。具体来说，如图Fig 3.5所示，首先初始化一个SMPL参数$p \in \mathbb{R}^{85}$，和特征输出$\phi \in \mathbb{R}^{2048}$进行拼接后，得到回归输入特征$\Phi \in \mathbb{R}^{2133}$。通过两层的全连接网络作为回归器，回归出SMPL模型参数$p \in \mathbb{R}^{85}$并将其反馈给输入前端，继续拼接并且循环刚才的过程。一般迭代次数设置$T=3$。

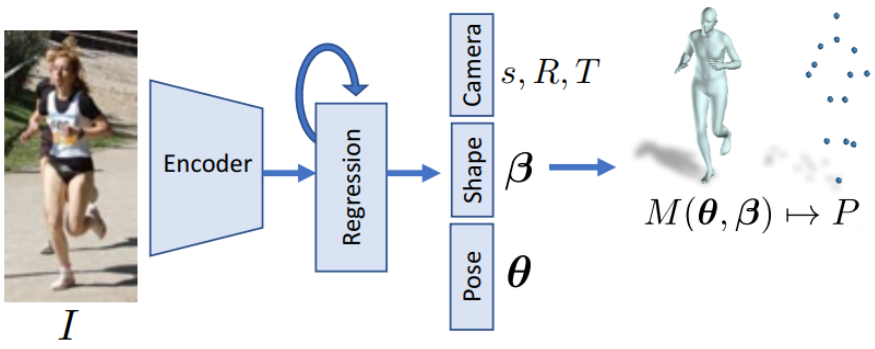


Fig 3.4 通过模型回归出人体的SMPL模型参数[19]。

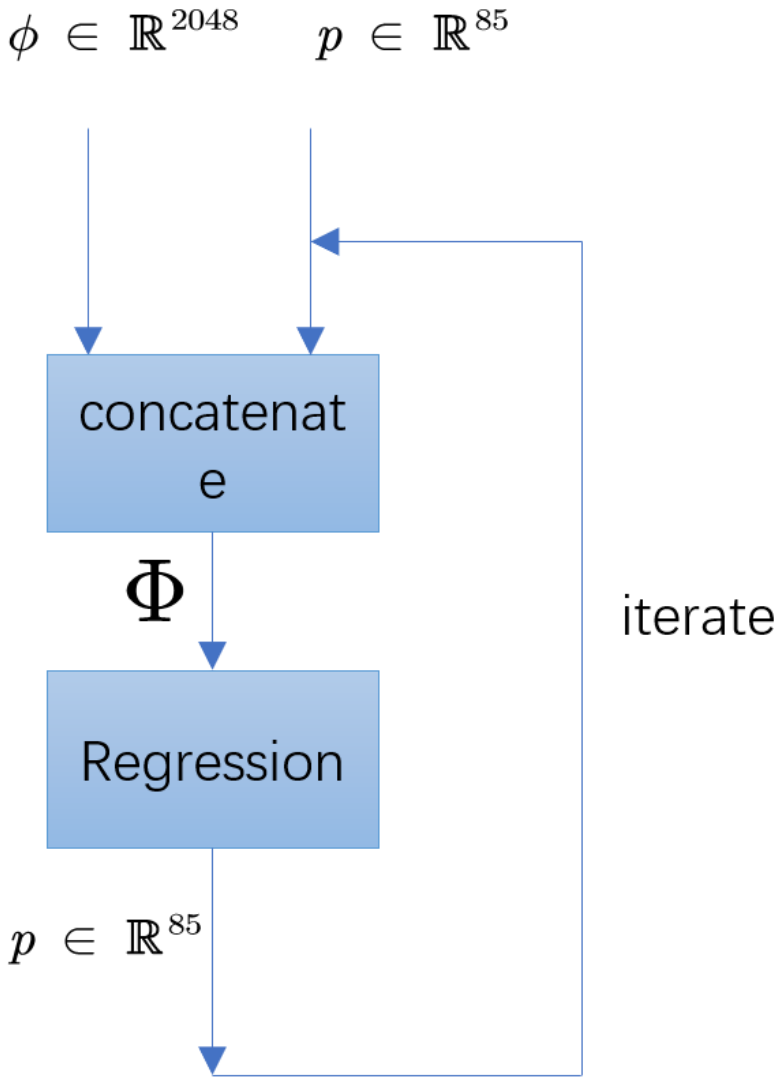


Fig 3.5 HMR中利用迭代去更好地回归出SMPL参数。

迭代过程的代码实例如下：

```
for i in range(n_iter):
    xc = torch.cat([x, pred_pose, pred_shape,
pred_cam], 1)
    xc = self.fc1(xc)
    xc = self.drop1(xc)
    xc = self.fc2(xc)
    xc = self.drop2(xc)
    pred_pose = self.decpose(xc) + pred_pose
    pred_shape = self.decshape(xc) + pred_shape
    pred_cam = self.deccam(xc) + pred_cam
```

通过这个方法，我们可以回归出以上提到的SMPL参数，

训练阶段

考虑到时序信息

Reference

[1]. Kocabas, Muhammed, Nikos Athanasiou, and Michael J. Black. "VIBE: Video inference for human body pose and shape estimation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[2]. https://zhuanlan.zhihu.com/p/115049353

[3]. https://zhuanlan.zhihu.com/p/42012815

[4]. Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7291-7299).

[5]. Fang, H. S., Xie, S., Tai, Y. W., & Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2334-2343).

[6]. https://en.wikipedia.org/wiki/Euler_angles

[7]. Pavllo D, Feichtenhofer C, Grangier D, et al. 3D human pose estimation in video with temporal convolutions and semi-supervised training[J]. arXiv preprint arXiv:1811.11742, 2018.

[8]. Pavlakos, G., Zhou, X., Derpanis, K. G., & Daniilidis, K. (2017). Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7025-7034).

[9]. https://en.wikipedia.org/wiki/Inverse_kinematics

[10]. https://blog.csdn.net/LoseInVain/article/details/107265821

[11]. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, D., & Black, M. J. (2019). Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10975-10985).

[12]. Joo, H., Simon, T., & Sheikh, Y. (2018). Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8320-8329).

[13]. Loper M, Mahmood N, Romero J, et al. SMPL: A skinned multi-person linear model[J]. ACM transactions on graphics (TOG), 2015, 34(6): 1-16.

[14]. https://whatis.techtarget.com/definition/3D-mesh

[15]. https://baike.baidu.com/item/BJD%E5%A8%83%E5%A8%83/760152?fr=aladdin

[16]. https://www.cnblogs.com/xiaoniu-666/p/12207301.html

[17]. https://blog.csdn.net/chenguowen21/article/details/82793994

[18]. https://khanhha.github.io/posts/SMPL-model-introduction/

[19]. Kanazawa, A., Black, M. J., Jacobs, D. W., & Malik, J. (2018). End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7122-7131).

[20]. Kocabas, M., Athanasiou, N., & Black, M. J. (2020). VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5253-5263).

[21]. Zhen, J., Fang, Q., Sun, J., Liu, W., Jiang, W., Bao, H., & Zhou, X. SMAP: Single-Shot Multi-Person Absolute 3D Pose Estimation. ECCV 2020

[22]. Rempe, Davis, Leonidas J. Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. "Contact and Human Dynamics from Monocular Video." Proceedings of the European Conference on Computer Vision (ECCV) 2020

[23]. https://blog.csdn.net/LoseInVain/article/details/102883243

[24]. https://blog.csdn.net/LoseInVain/article/details/102698703

[25]. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity mappings in deep residual networks." In *European conference on computer vision*, pp. 630-645. Springer, Cham, 2016.

[26].