指导教师：杨涛　　提交时间：3.13

CVPR2015 Paper Translation

No: 01
姓名：王淑君
学号：2013300211
班号：HC001310

学习一种基于深度图像的有效手势变换模型

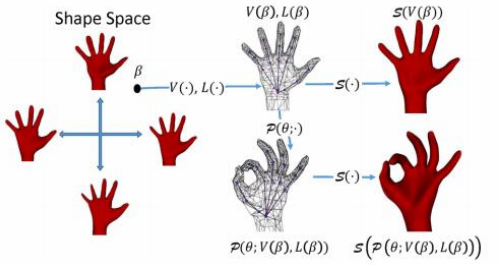
Sameh Khamis12 Jonathan Taylor2 Jamie Shotton
Cem Keskin2 Shahram Izadi2 Andrew Fitzgibbon2
1University of Maryland 2Microsoft Research

摘要

我们描述了如何学习一种复杂且有效率的有关人类手表面的变形的模型。本模型是由一组来自不同受试者表演出的不同的手势噪声深度图像建立的。我们通过学习到的参数模型和位姿模型发现用控制网的循环细分观察到的表面已经变形。本模型同时说明了目标特定形状和目标未知形状的变化。特别地，手势被原始姿态的平均筛目的线性组合参数化为少量的相对向量。这些网孔通过线性混合蒙皮算法衔接起来以形成细分表面的控制网。我们定义一个能量促使每个深度像素能够被我们的模型解释使用一个平滑的细分表面把所有的参数联合起来作一个粗糙的初始化。我们模型的结果通过合成数据和真实数据来演示这些数据表明手势能够由很少的基本组件构成。我们跟其他的方法（包括PCA）相比在我们模型的表征能力方面表现出了实质性的改善，与此同时保持了线性形状理论的效率。

1. 引言

人体三维形变模型已经成为了计算机图形和视觉领域一个伟大的成功故事。从布兰兹和维特尔【7】的面部模型到正在进行的全身的组合形状和姿势模型【5，13，15，12】，这些模型正在见证应用程序的商业发展包括虚拟购物（如：Metail），表演捕捉



图一

我们的可变表面模型考虑姿态（通过动画准备运动模型）和形状（用运动模型在形状空间学会最佳组合）。一组形状参数$\beta \in R^K$在形状空间（左）指定（上中）一个神经网络$V(\beta) \in R^{3\times M}$以及骨骼参数$L(\beta) \in R^{3\times B}$. 一系列关节角度$\theta$变形网格获得一个特定的网格$P(\theta; V(\beta), L(\beta)) \in R^{3\times M}$（左下）使用线性混合蒙皮算法 P(.).一个细分曲面模型 S(.)将网格映射到平滑的 3D 表面（右侧）。同时优化关节角的参数对整个管道三维形状给最好的参数即端到端模型稀疏和嘈杂的真实数据。

技术（如：faceshift）以及视频游戏（如：Kinect 体育竞技）。

然而，据我们所知，迄今为止没有一种三维形变手部模型被建立。手在某种理想的意义上是这样一种模型：

978-1-4673-6964-0/15/$31.00 2015©IEEE

它通常是裸体的，在自然 3D 用户接口方面拥有巨大的潜能。巴里安等人证实非常强健的手跟踪能够给出用户专用手模型，但是需要手动操纵和多摄像头捕获设备。泰勒等人【28】证实从单一的深度相机获取用户专用手模型，但是需长时间的校准序列获得手的所有自由度。

我们假设没有手的三维形变模型，因为目前的技术模型构建需要大量的高质量的数据集。即使手自由度的数目跟身体的数量相差不大，手趋向于展示而不是自我封闭，所以这样的扫描有很多的漏洞。此外，手相对来说较小，所以有关手的图像通常只含有少量的前景像素同时相机噪声会产生很大的影响。最终，手姿势空间可能大一些，关节的数目大约相同，因此，更多的特征抓取会被需要去精确的学习位姿空间。

在这篇文章中，我们克服了这些挑战并且根据从 50 个人身上用单一的 Kinect V2 传感器获取的一些短序列建立了一个三维手形变模型。我们方法的关键是双重。

首先，我们仅仅了解姿势和形状的这些方面：它们都不能被一个标准的操纵模型所解释。这虽然减少了数据需求，却产生了一个优势：我们系统的输出是根据线性混合蒙皮算法产生的一个标准的细分表面模型。这确保我们的方法具有极高的效率。与此相反，像 SCAPE 和 TenBo 模型在测试时包含了额外的线性解决方案，该解决方案在图形处理器上已经可实施，代表重要的其他计算代价。

第二，我们联结了匹配的所有部分扫描模型，而不是对每一个物体尝试分别建立完整的扫描同时进行主成分分析法（PCA）。正如我们在仿真和真实数据实验中展示的那样，这产生了一个更好的模型即使是对闭合的仿真数据，一个用真实扫描数据得到的更好的模型包含了丢失的噪声数据。

我们的主要贡献是一项用来学习有效率的由稀疏和噪音深度数据得到的骨骼驱动形变模型新的技术。这个学习模型包括一组参数化的基础网格骨架参数以及骨长度和皮肤等权重。虽然以前的工作已经学会了其中的一些参数，本文是第一个把所有参数连接起来的，第一个通过捕捉到的数据进行直接阐述的。

1.1. 相关工作

学习一种低维参数化形状的范围扫描或其他 3D 数据的例子已经证明对人体和人脸创建通用形变模型时是有效的【7，4，5，13，10，29，17，27，30，16，31，19】。已经建立了这样的形变模型，例如令人印象深刻的应用，为拟合单眼深度序列或更精确的身体或脸部跟踪已经被证明【7，11，14，21】。

然后，尽管人脸和人体的模型实现了长期成功，我们并没有发现任何已经存在的手部统计模型。这表明这是一个具有挑战的项目，已经存在全身【4，5，13，10】的非直接传输技

术。

艾伦等人【3】也做了与我们相似的工作，表明用线性混合蒙皮算法实现的细分曲面模型来代替标准模型。然而，重要的是，他们的不同之处在于基础表面上的位移映射。为了避免自相交有一定的大小限制以及其形状基础被迫配合输入扫描。此外，他们的优化步骤是顺序的(坐标下降法)而不是并行的，可能会导致不好的局部最优结果。卡什曼和费兹吉本【9】证明形状可变模型使用细分曲面根据极少的数据（3D 轮廓图像）训练模型。然而他们的模型并没区分形状和姿势，同时也没有训练参数形基础。

尤其是对于手来说，李仁济等人从在控制照明情况下单手的前景图像提取出了可见皱纹，局部关节，并且用用户的特定皮肤匹配了一个三维模型。该模型适合一张单一的图像，产生非常简单的有限自由度的手模型。阿尔布雷克特等人，去另一个极端创造非常详细、物理真实的手模型。一个更自动化的技术由泰勒【28】等人提出，给定噪声输入深度序列能够生成个性化手模型，用户的手旋转 180 度同时清晰地表现出手指。持续优化通过一个平滑细分曲面用尽可能严格的正规化产生高质量的用户具体操纵手模型共同解决了联系和模型参数，即使不是一个形状。虽然这个过程是自动化的，手要覆盖整个范围的手关节，以及必要的更长的序列，导致更复杂的捕捉需求和更昂贵的优化。

虽然没有明确本文研究对象，我们假设形状将被证明一个重要的健壮手姿，在之前它已经被全身跟踪证明过了。成年人手的解剖结构的研究已经显示出相当大的变化【8】，这显然是明显的性别和年龄差异。最近高质量工作，离线、表演手的捕捉，利用多个摄像机平台重申我们直觉的估计有关特定用户手型和手姿的重要性。巴郎等人【6】利用多相机平台和泊松表面重建个性化的手网，然后覆盖手部皮肤。他们用复杂的双手和手部接触展示出高质量的结果，然而，这个系统侧重于姿势而不是形状构建，这需要耗时的手工方式来进行。

2. 模型

现在我们描述手可变表面模型。我们将三角网格作为基本单元并大面积使用。本文讨论的所有网格表示人类的右手（但是并没有限定其他方面的特定手部模型)，网格包含 M 个顶点数，使用一个固定的三角网格。因此我们把这个网格表示为 3*M 的矩阵。第 M 列代表第 M 个顶点的位置。

如下详细解释，我们同时使用由 B 块骨头组成的手部骨架结构，使用一个 3*B 的矩阵来代表这些骨骼的位置。再一次，第 B 列代表第 B 块骨头的位置。这些骨头具有固定的层次结构，第一块骨头是位于根部的骨头，从第二块骨头开始$b \in \{2,\ldots,B\}$，它们父节点的索引为π_b。

在图一中，我们的形状模型使用一系列形状参数$\beta \in \mathrm{R}^K$对一个正对手掌

的姿势的一个手掌网格$V(\beta)$形状和骨架$L(\beta)$进行线性参数化。我们的姿势模型不考虑关节，通过对一个搬出姿势的手网格$P(\theta;V(\beta),L(\beta))\in\mathbb{R}^M$形成一个中立手网格和骨骼的映射。最后，我们的表皮模型使用循环细分把姿势映射到一个平滑的表面$S(P(\theta;V(\beta),L(\beta))\subset\mathbb{R}^3$。在本节的其余部分中,我们详细描述这些函数的具体形式。

2.1. 形状模型

我们的形状模型遵循我们的直觉，人手的单一姿势形状（或骨架）的变化比较紧凑，可以被描述为一个低维的线性子空间。因此我们使用一组网格矩阵基$\mathcal{V}=\{V_k\}_{k=1}^K\subset R^{3\times M}$和骨架位置基矩阵$\mathcal{L}=\{L_k\}_{k=1}^K\subset R^{3\times B}$把这个空间参数化。这两个基的维度数目都是 K 并不是偶然的，我们规定骨骼和皮肤的形状一起变化。特别的，给定一个向量的形状参数$\beta\in R^K$，一个中立的网格

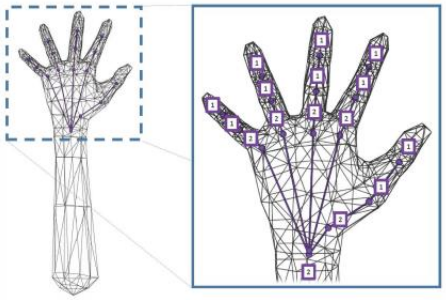
$$V(\beta;\mathcal{V})=\sum_{k=1}^{K}\beta_k V_k \tag{1}$$

一个中立的骨骼

$$L(\beta;\mathcal{L})=\sum_{k=1}^{K}\beta_k L_k \tag{2}$$

被恢复为这些矩阵的线性组合。

此外，我们相信得到第一基础组件 V1 和 L1 代表一个平均网格和骨架用其他的基本组件代表其余的部分。理想情况下，β_i应该（近似的）按比例进行编码，其他的坐标系用一定数量的偏移向量来表示。我们不显示的执行这些想法，使用正规化矩阵(见 3.1.2 和 3.1.3 章节)来代替。



图二

我们的手模型模板包含一个网格和运动骨架。它定义了网格拓扑和骨架结构的学习模型,并在优化时提供一个弱正则化的形状模型。这样的模板不需要精确解剖，模型将调整顶点和骨架,以确保观察手表面时精确的重建。在右边,我们已经标记每个关节的自由度的数量。

我们的线性模型跟 PCA 具有相同的表征能力，尽管不同之处在于它是如何学习的（见下文）。跟参考文献【5】中的模型对比，我们的方法可能在内存和计算方面更有效率。

2.2. 姿势模型

我们模型的一个重要特征是，我们不需要签名提到的形状模型来对由非中立的姿势造成的手表面的变形做出解释。代替的是，我们明确的用参数表示姿势，包括用一个向量θ把一组关节角度（见图二），全局方向和角度变换组合在一起。我们本节定义的姿势模型，详细说明了θ在一个中立姿势中使用对应的骨架$L(\beta)$调用一个网格$V(\beta)$实现连接变形。有很多的公式可以套用，由于线性混合蒙皮算法（LBS）应用的普遍性以及很高的效率，我们

决定使用它。

为了清晰起见，我们暂时放弃依赖形状参数β，并且证明最低有效单元如何让一个固定的网格$V=[v_1 \dots v_n$ 和骨架$L=[l_1 \dots l_b]$变形。这个模型要求每跟骨头被赋予一个固定的旋转矩阵Q_b这代表骨骼的关节的旋转主轴。跟骨头的位置l_b一起，这定义了一个坐标系

$$H_b = \begin{bmatrix} Q_b & l_b \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4} \qquad (3)$$

通过一个精确变换从骨骼空间映射到世界空间。注意，一组$H=\{H_b\}_{b=1}^{B}$隐形定义了一组等价的变换$\{T_b\}_{b=1}^{B}$，T_b映射指的是从骨架坐标系到它的父坐标系π_b（$H_b=T_1 \dots T_{\pi_b} T_b$）。

给定一组姿势参数θ，线性混合蒙皮算法清楚的通过提供一个三维旋转$\tilde{R}_b$表达一个关节，用均匀地旋转矩阵

$$R_b(\theta) = \begin{bmatrix} \tilde{R}_b(\theta) & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 4}, \qquad (4)$$

来表达

每个骨骼 b 的局部坐标系。另外，全球定位和位置由于全局精确变换$R(\theta)\in \mathrm{R}^{4\times 4}$被应用到世界。对每根骨头 b，我们这样获得精确变换$G_b(\theta)$联系骨头 b 的局部坐标系统到世界系统在姿

$$G_1(\theta) = R(\theta) T_1 R_1(\theta) \qquad (5)$$
$$G_b(\theta) = G_{\pi_b}(\theta) T_b R_b(\theta). \qquad (6)$$

势θ下，通过如下迭代：

网格通过一组皮肤权重$\Gamma=\{\alpha_{bm}\ b\in\{1,\dots,B\}\ m\in\{1,\dots M\}\}$，给骨架“植皮”，对于所有的 m 有$\sum_{b=1}^{B}\alpha_{bm}=1$。直觉的，$\alpha_{bm}$表明骨头在移动时附着定点 m 和骨头 b 的强度。

更精确的是，在 V 中第 m 个顶点的姿势θ有关节的位置

$$I_{3\times 4} R(\theta) \sum_{b=1}^{B} \alpha_{bm} G_b(\theta) H_b^{-1} \hat{\mathbf{v}}_m \qquad (7)$$

$I_{3\times 4}$用欧几里得坐标来表示$v_m=[\ v_m^T \quad 1\ ]^T$。注意$H_b^{-1}$首先从中立姿势的定点 vm 的世界坐标系到骨架 b 的局部坐标系的映射。转换方程式$G_b(\theta)$将顶点重新映射到θ的世界坐标系。

为了方便符号标记，现在我们为姿势不变量形状参数[1]（我们迄今为止提到的参数）创建一个三元组$\Upsilon=(\mathcal{V},\mathcal{L},\Gamma)$。重新介绍形状参数$\beta$，通过 LBS 模型$\Phi(\beta;\Upsilon)=(V(\beta),L(\beta),\Gamma)$获得参数，因此我们用$\mathcal{P}(\theta;\Phi(\beta;\Upsilon))\in\mathbb{R}^{3\times M}$通过应用姿势$\theta$来对中立的形状$\beta$手网格表明生成的网格。

2.3. 表面模型

接下来我们对模型用一个控制网格循环细分法来表示实际的表面。给定一个网格 V，循环细分的处理过程时通过对每个三角表面（根据固定的三角）进行迭代细分，用他们周围的数据来平滑顶点的位置。限制性表面$\mathcal{S}(V)\subset\mathbb{R}^3$能够哦那个过执行这个细分过程无数次。

为了避免这个复杂的构建过程我们根据文献【28】用下式代替参数化我们的表面模型。

$$\mathcal{S}(\mathbf{u};V):\Omega\times\mathbb{R}^{3\times M}\mapsto\mathbb{R}^3, \qquad (8)$$

这个映射从位置 u（表面坐标系的二维空间Ω）到一个三维细分表面中的一个

[1] 我们排除保持不变的参数$\{Q_b\}_{b=1}^{B}$

点。在这个定义之下，整个表面能被写成$\mathcal{S}(V) = \{\mathcal{S}(\mathbf{u}; V) : \mathbf{u} \in \Omega\}$。由于空间限制，我们为了函数$\mathcal{S}(\cdot)$的精确细节以及参数化 u 参考文献【28】。然而，它已经满足了我们的需求$\mathcal{S}(\cdot)$及其派生物关于 u 和 V 能够被有效的计算。我们用模型的其他部分组成$\mathcal{S}(\cdot)$生成$\mathcal{S}(\mathbf{u}; \mathcal{P}(\theta; \Phi(\beta; \Upsilon))$，手的三维坐标系 u 的位置用形状$\beta$和姿势$\theta$来表示。

2.4. 完整模型

做个总结，当一组特定参数β被选中时，我们得到了 LBS 手部模型的特定的目标参数$\Phi(\beta; \Upsilon)$。然后我们获得一组用形状β和姿势θ表示成$\mathcal{P}(\theta; \Phi(\beta; \Upsilon))$的网格。最后我们得到坐标在$\mathbf{u} \in \Omega$范围内细分表面用形状$\beta$和姿势$\theta$表示成$\mathcal{S}(\mathbf{u}; \mathcal{P}(\theta; \Phi(\beta; \Upsilon)))$对应的表面。我们下一步的考虑是学习一组$\gamma$变量，这样形状$\beta$和姿势$\theta$就能单独用来描述大部分的人手形状和姿势了。

3. 拟合模型

这项工作的主要贡献是展示如何学习参数γ根据一组使用者手部的噪声深度图像。为此，我们假定我们有很多不同类型的试验者（例如男人，女人，小孩等不同的手型）。对于每个试验对象 s，我们拥有F_s个它们表演的各种各样的手关节深度图像。在每一帧 f 中，一组N_{sf}数据点$\{\mathbf{x}_{sfn}\}_{n=1}^{N_{sf}} \subset \mathbb{R}^3$及提取出的相应的估计值$\{\mathbf{n}_{sfn}\}_{n=1}^{N_{sf}} \subset \mathbb{R}^3$。

3.1. 能量

我们想要使用这些数据来训练参数γ，因此我们的模型能够解释数据并且满足简单的先验条件。我们把能量最小化的问题以γ作为自变量得到方程。

$$E(\Upsilon) = \sum_{s=1}^{S} E^s(\Upsilon) + \lambda_{\text{arap}} E_{\text{arap}}(\Upsilon) + \lambda_{\text{skin}} E_{\text{skin}}(\Upsilon) \quad (9)$$

后面两个加权的系数规范化了基本表示和皮肤权重，如下式表示，每个对

$$E^s(\Upsilon) = \min_{\beta} \sum_{f=1}^{F_s} E^{sf}(\beta; \Upsilon) + \lambda_{\text{shape}} E_{\text{shape}}(\beta) \quad (10)$$

在对象 s 的数据基础上提供了γ的约束。10 式中第二项编码一个形状的先验错误。各项如下

$$E^{sf}(\beta; \Upsilon) = \min_{\theta} \sum_{n=1}^{N_{sf}} E^{sfn}_{\text{data}}(\theta, \beta; \Upsilon) + \lambda_{\text{pose}} E_{\text{pose}}(\theta)$$

用来测试 f 帧中摆好姿势的表面得到的数据质量如何。

3.1.1. 数据项

我们使用的数据项是

$$E^{sfn}_{\text{data}}(\theta, \beta; \Upsilon) = \quad (11)$$
$$\min_{\mathbf{u} \in \Omega} \rho(\|WQ_{sfn}(\mathbf{x}_{sfn} - \mathcal{S}(\mathbf{u}; \mathcal{P}(\theta; \Phi(\beta, \Upsilon))))\|)$$
$$+\lambda_{\text{normal}} \rho^{\perp}(\|1 - (\mathbf{n}_{sfn})^{\top} \mathcal{S}^{\perp}(\mathbf{u}; \mathcal{P}(\theta; \Phi(\beta, \Upsilon)))\|)$$

和$\rho^{\perp}(e)$分别应用于点坐标错误和方形普通错误解决。我们设置缩放矩阵$W = \text{diag}(1, 1, \zeta)$，加上旋转量$Q_{sfn}$旋转$\rho(e)$三维余量，因此$\mathbf{x}_{sfn}$坐标具有了 z 轴，模型深度传感器的相对较高的观察方向的不确定性。

3.1.2. 尽可能精确的正规化

$E_{\text{arap}}(\Upsilon)$尽可能精确的调用了假定$\mathcal{V}$和$\mathcal{L}$的变形。我们使用尽可能精确把正规化能量定义成

$$E_{\text{arap}}(\Upsilon) = D(V_1, V_{\text{template}}) + D^{\dagger}(L_1, V_1, L_{\text{template}}, V_{\text{template}})$$
$$+ \sum_{k=2}^{K} (D(V_k, \varnothing) + D^{\dagger}(L_k, V_k, \varnothing, \varnothing)) \quad (12)$$

V_{template}和L_{template}分别代表粗糙的手型模板（见图二）网孔和骨头的位置

信息，$\varnothing$是一个适当大小的空矩阵。式（12）中只有两个参数是标准的正规化参数，两个网孔和顶点位置的变形在$\mathbb{R}^{3\times M}$中被定义为

$$\sum_{m=1}^{M}\min_{R\in SO(3)}\sum_{n\in\mathcal{N}(m)}\|(\mathbf{v}_n-\mathbf{v}_m)-R(\mathbf{v}'_n-\mathbf{v}'_m)\|^2, \quad (13)$$

$\mathcal{N}(m)$是一系列顶点 m 的邻接顶点。

在尽可能精确的正规化下，精确的变换并非是不利的，局部位置的非精确变换比大片区域的非精确变换要更加不好。注意：

$$D(V,\varnothing)=\sum_{m=1}^{M}\sum_{n\in\mathcal{N}(m)}\|(\mathbf{v}_n-\mathbf{v}_m)\|^2, \quad (14)$$

只是鼓励相近的顶点重合。在这种情况下，当 k>=2 时，V_k代表与平均网格V_1的偏移，因此这转换为我们所希望的顶点偏移量时平滑的。

式【12】利用四个参数生成了尽可能精确的正规化函数的改进版本，促进核心网孔中骨架位置跟附近的顶点（特别是定点环）保持一致。我们对于每根骨头 b 表示一系列的顶点索引值为$C_b\subset\{1,\ldots,M\}$。一堆骨头的位置矩阵$L,L'\in\mathbb{R}^{3\times B}$由列$\{\mathbf{l}_b\}_{b=1}^{B}$，$\{\mathbf{l}'_b\}_{b=1}^{B}$和网孔顶点矩阵$V.V'\in\mathbb{R}^{3\times M}$和列$\{\mathbf{v}_m\}_{m=1}^{M}$ $\{\mathbf{v}'_m\}_{m=1}^{M}$，$D(L,V,L',V')$表示为：

$$\sum_{b=1}^{B}\min_{R\in SO(3)}\sum_{m\in C_b}\|(\mathbf{v}_m-\mathbf{l}_b)-R(\mathbf{v}'_m-\mathbf{l}'_b)\|^2. \quad (15)$$

类似的，我们得到

$$D^{\dagger}(L_k,V_k,\varnothing,\varnothing)=\sum_{b=1}^{B}\sum_{m\in C_b}\|(\mathbf{v}_m-\mathbf{l}_b)\|^2, \quad (16)$$

对于普通的成分 k>=2，同样鼓励一根骨头的偏移跟与骨头相接的顶点的偏移是相似的。

3.1.3. 形状先验

我们规范形状参数β使用方程：

$$E_{\text{shape}}(\beta)=(1-\beta_1)^2+\sum_{k=2}^{K}\beta_k^2 \quad (17)$$

用最少的定点和骨头位置偏移鼓励用户特定手模型来保持相关性。

3.1.4. 姿势先验

我们强烈的惩罚任何姿势变形违反了人们的物理约束通过给姿势添加障碍约束θ使用方程：

$$E_{\text{pose}}(\theta)=\sum_i\begin{cases}(\theta_i-\theta_i^{\min})^4 & \text{if } \theta_i<\theta_i^{\min}\\ (\theta_i^{\max}-\theta_i)^4 & \text{if } \theta_i>\theta_i^{\max}\\ 0 & \text{otherwise}\end{cases} \quad (18)$$

$\theta^{\min}$和$\theta^{\max}$是手部关节旋转角度的最小是和最大值近似值。

3.1.5. 皮肤权重先验

为了保证每个定点 m 的皮肤权重总和为 1，我们发现增加另一个能量$E_{\text{skin}}(\Upsilon)=\sum_{m=1}^{M}\|\sum_{b=1}^{B}\alpha_{bm}-1\|^2$能够减少$\lambda_{\text{skin}}$（见式 9）大的权重偏差。为了确保皮肤权重保持非负，我们用取对数的形式$\tilde{\alpha}_{bm}=\log(\alpha_{bm})$对 b 根骨头的 m 个顶点进行参数化。

4. 优化

为了优化你的模型的能量方程，我们把它提升为一个简单的能量，通过最新的一组变量来定义，使用标准的非线性优化器来优化。

4.1. 改进能量

正如前面所定义的那样，我们的

能量$E(\Upsilon)$是一个复杂的形式，包含了很多最小化项。根据文献【28】，我们注意到下面的两个关于 x1 和 x2 的实值函数$f(x)$和$g(x)$：

$$\min_x f(x) + \min_x g(x) = (\min_{x_1} f(x_1) + \min_{x_2} g(x_2)) = \min_{x_1,x_2}(f(x_1)+f(x_2)) \le f(x_1)+f(x_2) \quad (19)$$

最少数量的变量除以总的变量能够被标记且通过求和。我们的能量通过引进一系列的形状参数$\mathcal{B} = \{\beta^s\}_{s=1}^S$和位置参数$\Theta = \{\theta_{sf} : s \in \{1,...,S\}, f \in \{1,...,F_s\}\}$能够被改进，对应$\mathcal{U} = \{\mathbf{u}_{sfn} : s \in \{1,...,s\}, f \in \{1,...,F_s\}, n \in \{1,...,N_{sf}\}\}$，同时尽可能精确的规范化旋转[2] $\mathcal{R} = \{R_m\}_{m=1}^M \cup \{R_b^l\}_{b=1}^B$。这引进了一个新的能量$E'(\Upsilon,\mathcal{B},\Theta,\mathcal{U},\mathcal{R})$，比如

$$E(\Upsilon) = \min_{\mathcal{B},\Theta,\mathcal{U},\mathcal{R}} E'(\Upsilon,\mathcal{B},\Theta,\mathcal{U},\mathcal{R}) \le E'(\Upsilon,\mathcal{B},\Theta,\mathcal{U},\mathcal{R}) \quad (20)$$

对任何一组$\mathcal{B},\Theta,\mathcal{U},\mathcal{R}$有上式。我们在补充材料里列入了改进了的能量的整个形式，不过可以假设一种简单的形式$\lambda_{\text{arap}} = \lambda_{\text{skin}} = \lambda_{\text{shape}} = \lambda_{\text{pose}} = 0$，我们的数据项就简化为

$$E_{\text{data}}^{sfn}(\theta,\beta,\Upsilon) = \min_{\mathbf{u}\in\Omega} \|\mathbf{x}_{sfn} - \mathcal{S}(\mathbf{u};\mathcal{P}(\theta;\Phi(\beta,\Upsilon)))\|^2 . \quad (21)$$

所以改进过后的能量为

$$E'(\Upsilon,\mathcal{B},\Theta,\mathcal{U},\mathcal{R}) = \sum_{s=1}^{S}\sum_{f=1}^{F_s}\sum_{n=1}^{N_{sf}} E_{\text{data}}'^{sfn}(\mathbf{u}_{sfn},\theta_{sf},\beta_s;\Upsilon) \quad (22)$$

同时

$$E_{\text{data}}'^{sfn}(\mathbf{u},\theta,\beta,\Upsilon) = \|\mathbf{x}_{sfn} - \mathcal{S}(\mathbf{u};\mathcal{P}(\theta;\Phi(\beta,\Upsilon)))\|^2 . \quad (23)$$

改进过后这个公式去除了内部最小化对于$\mathbf{u} \in \Omega$。

4.2. 非线性优化

我们使用列文伯格-马夸尔特法来优化这个能量同时利用 Ceres solver【1】来自动处理大型、动态、稀疏问题。我们的优化计划（见下文）将会利用非线性子程序$(N, \mathcal{F})$对 N 个列文伯格-马夸尔特法步骤优化除了$\mathcal{F} \subset \{\Upsilon,\mathcal{B},\Theta,\mathcal{U},\mathcal{R}\}$之外的参数，一种 Ceres. 支持的运算。

4.3. 离散优化

为了跳出局部极小值，我们同时利用副程式的离散修正目的是提高 u 的对应率通过寻找一组离散的参与者。特别的，我们考虑一组例子$\mathcal{U}_{\text{prop}} = \mathcal{U} \cup \mathcal{U}_{\text{samp}} \subseteq \Omega$当$\mathcal{U}_{\text{samp}}$是一组固定的表面参数组合，在$\Omega$寻找粗糙的一致的样例。然后我们对目标 s，帧 f 和数据点 n 考虑执行一组循环来寻找一组新的表面坐标

$$\mathbf{u}'_{sfn} = \arg\min_{\mathbf{u}\in\mathcal{U}_{\text{prop}}} E_{\text{data}}'^{sfn}(\mathbf{u},\theta_{sf},\beta^s,\Upsilon) . \quad (24)$$

作为结果的$\mathcal{U}' = \{\mathbf{u}'_{sfn}\}$的值保证没有超过能量$(i.e.\ E'(\Upsilon,\mathcal{B},\Theta,\mathcal{U}',\mathcal{R}) \le E'(\Upsilon,\mathcal{B},\Theta,\mathcal{U},\mathcal{R}))$。

4.4. 初始化

我们手动初始化位置Θ为了使模板按照点云的形式大致排列。尽管我们可以考虑自动初始化的方法，这个行为并非很频繁执行，只需要初始化一次就够了。相似的，每个$\beta^s \in \mathcal{B}$被初始化，这样β_1^s符合粗略大小的的目标 s 同时对于 k>=2 都有$\beta_k^s = 0$。我们用粗略的手模板(见第五小节)初始化 V1 和 L1，同时用 0 初始化其余的出事成分。所有的尽可能精确地正规化循

[2] 尽可能规范化的旋转对很多项来说牵扯到基本的成分，因此不需要参数化

环在 R 范围内被初始化为识别，U 是离散优化的调用。

4.5. 优化调度

初始化后，我们需要执行一个优化调度（见算法 1）交叉离散更新不断优化，同时不断解除参数的限制。我们发现对各个阶段的算法进行排序找到一个好的最小值是的算法更具有鲁棒性，这样两个阶段之间精确地转换时间就变得很少。

算法 1　优化调度	
局部最优化(4, $B\cup\Gamma$)	
转换成 GM()	转换成 Geman McClure,健壮的误差函数
局部最优化(4, $B\cup\Gamma$)	
局部最优化(4, Γ)	
局部最优化(4, $\varnothing$)	
函数　局部最优化	
for i=1:N:N do	
NoLinear(25,F)	
DiscreteUpdate()	

5. 评估

我们现在描述进行的设置和各种实验来评估我们的方法。

手模板：我们使用三维模型软件 Blender 来操纵模板模型。

参数设置：数据项$\rho(e)$和$\rho^{\perp}(e)$健壮的内核被初始化为柯西内核$\rho(e;\sigma)=\sigma^2\log(1+e^2/\sigma^2)$约束了异常的数值。然后当参数合理的接近一个好的解决方案我们转换为非常健壮的 GM 内核$\rho(e,\sigma)=e^2/(e^2+\sigma^2)$来避免拟合大多数的异常数值。见算法 1。

5.1. 数据库

为了评估和比较我们的模型，我们使用三种数据库：(i) 3 维合成数据，一组仿真的数据集包含三维的数据点云覆盖了整个手部表面。(ii) 2.5 维的合成数据，一组深度图像的仿真数据。(iii) 真实的数据，从 Kinect V2 传感器获取的一组深度图像的实数数据库。

三维的合成数据通过流行的模型工具 Poser 生成。Poser 支持生成超过 100 混合形

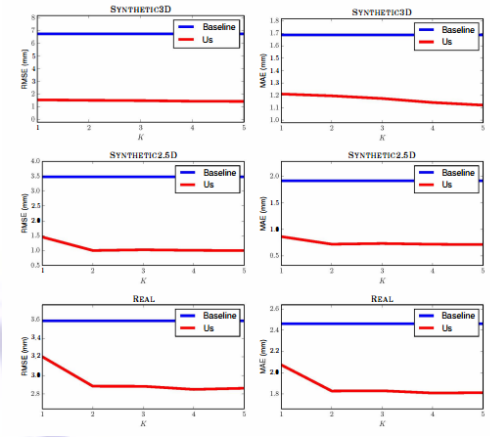


图 3

RMSE（左）和 MAE（右）K 个基础组建的基准（蓝）和我们的模型（红）。

状能够被用来生成一组大量的不同类型的真实的手模型。我们因此用 15 个不同的姿势随意抽取 50 个明显手形状的权值。在数据库中，所有的三维顶点都被用到，即使有一部分顶点在一

部真实的深度相机中无法被显示出来。2.5 维合成数据，从另一个方面来讲，用一个去除掉任何相机捕捉不到点的虚拟的深度相机在一个固定的角度来生成项目需要的 3 维数据。3 维和 2.5 维的合成数据是无噪声的数据，我们用这些数据来测试模型的表现力，而完全不用担心陷入局部最小值。我们也测试增加人工噪声对拟合过程的影响。

真实的数据需要用 Kinect V2 传感器来获取。我们记录一组 50 个不同的对象：17 个女士，31 个男士和 2 个儿童，每个对象被要求在深度相机前表演各式各样的手关节动作。我们对每个对象选择了平均 15 个各不相同的手型。与合成数据不同的是，真实的数据包含由于深度的不连续和多路径交叉（见图 5）产生的一组合适数量的噪声和异常的像素值。

5.2. 基准

我们跟一个建立在个性化程序（详见文献【28】）上的基准的方法作比较。通过对 S 个对象分别应用个性化，我们能获得一组个性化的网孔 $\{V^s\}_{s=1}^S$，骨架 $\{L^s\}_{s=1}^S$ 和大小 $\{\beta_0^s\}_{s=1}^S$。对于对象 s，我们能链接和平滑这些矩阵并且去除掉像比例 $p^s = \frac{1}{\beta_0^s}(\vec{V}^{s\top}\ \vec{L}^{s\top})^\top$ 这样的因子。通过对向量 $\{p^s\}_{s=1}^S$ 应用 PCA，我们获得了平均向量 $\bar{p}$ 以及一系列原始方向 $\{p_k\}$。特别的，每一个输入的网孔 p 都有一个向量 α^s 与其对应，如：

$$p^s = \bar{p} + \sum_k \alpha_k^s p_k \,. \qquad (25)$$

注意如果我们把 K 维的 PCA 方向截短成最小值

$$\min_{\{p^s\}_{s=1}^S} \| \min_{\alpha} (\bar{p} + \sum_{k=1}^{K} \alpha_k p_k - p^s) \|^2 \,. \qquad (26)$$

相比之下，我们的模型通过细分表面、皮肤和线性形状基在观测到的点和模型表面之间最小化了三维误差。

5.3. 结论

根据所有的数据库数据，我们使用 30 个对象的数据来训练（学习形状基参数 Υ 跟每个对象的形状系数 β^s 和每帧的姿势 θ_{sf}），20 个对象的数据来测试（固定 Υ 不变，优化 β^s 和 θ_{sf}）。所有报告的定量和定性的结果，包括图形，都是测试集提供的。

定量分析上，我们对测试集和皮肤表面上每个数据点计算 3 维残差，同时使用均方根误差（RMSE）和平均绝对误差（MAE）做总结。由于人造 3 维和 2.5 维数据缺少噪声，均方根误差是一个合理的度量方法。然而，对于真实的数据，平均绝对误差对异常值的鲁棒性更好。误差级用毫米（mm）来表示。

在图 3 中，我们展示了均方根误差和平均绝对误差对三组数据库中基成份数量的影响。对 PCA 基准，我们固定 K，目的是说明训练网孔上至少 90%的变化。这导致对于真实的数据和人造 2.5 维数据 k=4 时以及人造 3 维数据 k=5 包含平均向量。此外基础组件利于错误率，当 k=3 时精度出现饱和。

在图 4 中，我们展示了在一个给定的阈值下用平均方差求得点的饱和的百分比。类似的，在 k=1 时我们的结果比基准的要好，尽管基准使用了 4~5 个基本组件。我们使用附加的基础组件提高精确度，但是当 k=2 时会发生饱和。

我们在图 5 中说明了对真实数据库进行处理的一些定性的拟合结果。这也展示了我们的模型可以处理的各种各样的手型和手的姿态，还有对异常值的鲁棒性时可靠的。在图 6 中，显示了形状系数对于所有真实数

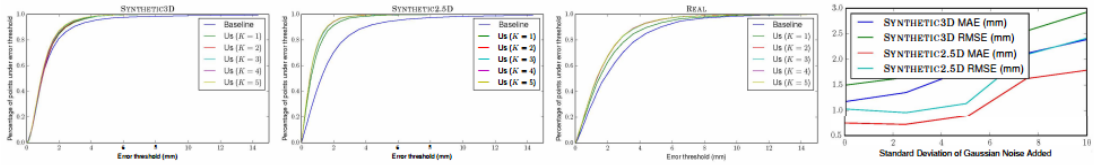


图 4　左边三通道；误差低于基准阈值的数据点的百分比以及我们的模型有 k 个基本组件。

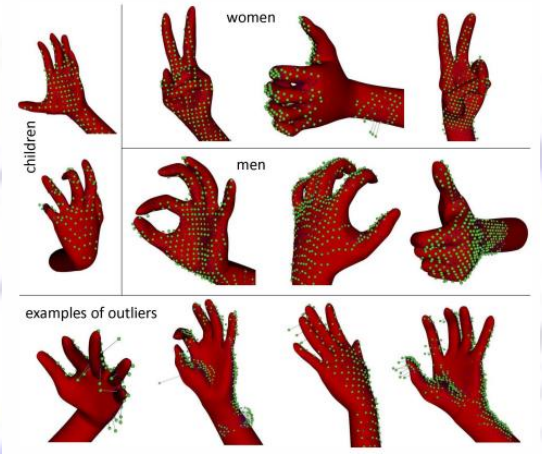


图 5 我们的表面（红）模型根据不同的帧来拟合数据，每个数据点 $\mathbf{x}_{sfn} \in \mathbb{R}^3$ （绿）有一个关联的表面点 $S(\mathbf{u}_{sfn}; \ldots)$ （蓝）。异常值的例子展示了拟合当前噪声的鲁棒性。

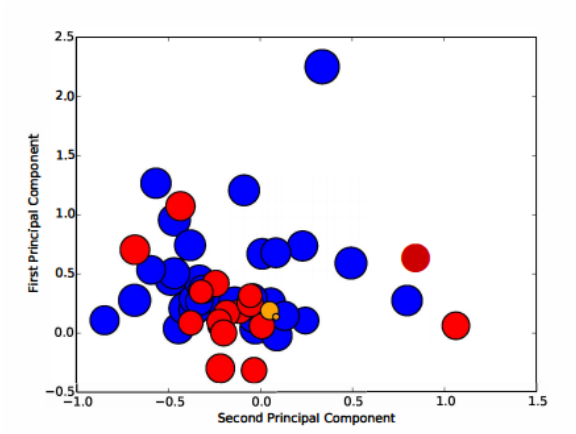


图 6　当我们模型跟真实数据集相拟合时学习一组形状系数 $\{\beta_s\}_{s=1}^{S}$，投影到前两个原方向。每个点编码规模系数 $(i.e.\ \beta_1)$ 的大小。大家可以看到儿童的手较小与此同时男士对象（蓝）的手更大。

右在合成数据集上进行控制噪声试验。

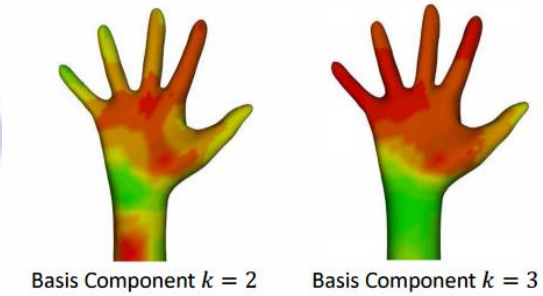


图 7　对于前两个偏移基础组件在真实数据集上进行学习，我们通过偏移向量模给表面涂色。绿色和红色分别代表偏移量小还是大。

据集上的对象投射为前两个主要成分[3]。图 7 通过偏移向量的模形象的展示了前两个偏移基础组件（k=2，3）。

[3] 这个向量纯粹是为了可视化的目的，对于基准没有什么用处

左边的基础成分在手指和手腕的序列上有很大的影响，同时右边的基础组件通过拇指影响到最远的手指，扩展到整个手掌。

最后，我们使用合成 3 维和 2.5 维的数据集进行噪声控制实验（见图 4）。我们使用 K=4 固定数量的基础组件，我们增加了不同的三维高斯噪声标准差。在这个实验中噪声只添加到训练数据中来展示我们优化方案的鲁棒性。

6. 讨论

我们已经展示了一个骨架驱动的可变模型如何从稀疏噪声数据中学习，同时这个方法大大超过了基准的方法。

我们模型的测试时间非常有效，是线性基础组件的数量级，只需要几个组件来准确的描述人类各种各种的手型。一旦形状参数 β 为一个给定的使用者推测出来，这个形状模型能够被“烤”和控制网孔细分有限次数的标准 LBS 网孔模型。这可以暂时利用详细的 LBS 模型【22】，进行手追踪，因为他越来越普遍的用于个性化形状模型的手追踪【25，26】。

作为未来的工作，在追踪的过程中个性化手模型交互和实时拟合形状参数 β 将会变得很自然。同时，我们想调查姿势形状的明确编码依赖的有效项。当前模型是一个相当粗糙的方案，但是这将是有趣的，看看这种方法能否产生一个超解模型。最后，我们希望把我们的技术应用于人体和动物的其他方面。

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. https: //code.google.com/p/ceres-solver/. 6

[2] I. Albrecht, J. Haber, and H.-P. Seidel. Construction and animation of anatomically based human hand models. In Proc. Eurographics, 2003. 2

[3] B. Allen, B. Curless, and Z. PopoviC. Articulated body deformation from range scan data. In ACM Transactions on Graphics (TOG). ACM, 2002. 2

[4] B. Allen, B. Curless, and Z. PopoviC. The space of human body shapes: reconstruction and parameterization from range scans. ACM Trans. Graphics, 22(3), 2003. 2

[5] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape completion and animation of people. ACM Trans. Graphics, 24(3), 2005. 1,2,3

[6] L. Ballan, A. Taneja, J. Gall, L. V. Gool, and M. Pol1efeys. Motion capture of hands in action using discriminative salient points. In Proc. ECCV, 2012. 1,2

[7] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In Proc. SIGGRAPH, 1999. 1,2

[8] A. Buryanov and V. Kotiuk. Proportions of hand segments. International Journal of Morphology, 28(3):755-758, 2010. 2

[9] T. Cashman and A. Fitzgibbon. What shape are dolphins? building 3D morphable models from 2D images. IEEE Trans. PAMI, 2013. 2

[10] Y. Chen, Z. Liu, and Z. Zhang. Tensor-based human body modeling. In Proc. CVPR, June 2013. 2

[11] M. de La Gorce, N. Paragios, and D. J. Fleet. Model-based hand tracking with texture, shading and self-occlusions. In Proc. CVPR, 2008. 2

[12] O. Freifeld and M. l. Black. Lie bodies: A manifold representation of 3D human shape. In Proc. ECCV, 2012. 1

[13] N. Hasler, C. Stoll, M. Sunkel, B. Rosenhahn, and H.-P. Seidel. A statistical model of human pose and body shape. Computer Graphics Forum, 2009. 1,2

[14] T. Helten, A. Baak, G. Bharaj, M. Milller, H.-P. Seidel, and C. Theobalt. Personalization and evaluation of a real-time depth-based full body tracker. In 3DV, 2013. 2

[15] D. A. Hirshberg, M. Loper, E. Rachlin, and M. l. Black. Coregistration: Simultaneous alignment and modeling of articulated 3D shape. In Proc. ECCV, 2012. 1

[16] H. Li, R. W. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. Proc. SGP, 2008. 2

[17] M. Liao, Q. Zhang, H. Wang, R. Yang, and M. Gong. Modeling deformable objects from a single depth camera. In Proc. ICCV, 2009. 2

[18] C. Loop. Smooth Subdivision Surfaces Based on Triangles. Master's thesis, The University of Utah, Aug. 1987. 4

[19] T. Neumann, K. varanasi, N. Hasler, M. Wacker, M. Magnor, and C. Theobalt. Capture and statistical modeling of armmuscle deformations. In Proc. Eurographics, 2013. 2

[20] T. Rhee, U. Neumann, and J. P. Lewis. Human hand modeling from surface anatomy. In Proceedings of the 2006 symposium on Interactive 3D graphics and games, pages 27-34. ACM,2006. 2

[21] N. Robertini, T. Neumann, K. varanasi, and C. Theobalt. Capture of arm-muscle deformations using a depth camera. In Proceedings of the 10th European Conference on Visual Media Production, 2013. 2

[22] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. vinnikov, Y. Wei, D. Freedman, P. Kohli, E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, robust, and flexible realtime hand tracking. In Proc. CHI, 2015. 8

[23] S. M. Software. Poser. http://my.smithmicro . com/ poser-3d-animation-software.html.7

[24] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In Proc. SGP, 2007. 5

[25] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In 2013 IEEE International Conference on Computer Vision (ICCV), Dec. 2013. 8

[26] S. Sridhar, H. Rhodin, H.-P. Seidel, A. Oulasvirta, and C. Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In Proceedings of the International Conference on 3D Vision (3DV), 2014. 8

[27] C. Stoll, Z. Karni, C. RossI, H. Yamauchi, and H.-P. Seidel. Template deformation for point cloud fitting. In Proc. Eurographics, 2006. 2

[28] J. Taylor, R. Stebbing, V. Ramakrishna, C. Keskin, J. Shotton, S. Izadi, A. Hertzmann, and A. Fitzgibbon. User-specific hand modeling from monocular depth sequences. In Proc. CVPR,2014. 1,2,4,5,7

[29] A. Tsoli, N. Mahmood, and M. l. Black. Breathing life into shape: capturing, modeling and animating 3D human breathing. ACM Trans. Graphics, 33(4), 2014. 2

[30] M. Wand, B. Adams, M. Ovsjanikov, A. Berner, M. Bokeloh, P. lenke, L. Guibas, H.-P' Seidel, and A. Schilling. Efficient reconstruction of nonrigid shape and motion from real-time 3d scanner data. ACM Trans. Graphics, 2009. 2

[31] A. Weiss, D. Hirshberg, and M. J. Black. Home 3D body scans from noisy image and range data. In Proc. ICCV, 2011. 2