基于隐式神经网络的 8K 三维光场显示实时双手交互研究

邓新堯，邢树军
（北京邮电大学电子工程学院）

摘要：在本文中，我们提出了一种实时三维光场显示交互系统。该系统使用户能够在三维光场光场的实时显示中准确地与 3D 对象进行交互。我们的双手势识别模块实时获取双手的骨骼信息，并通过归一化和位置编码处理关键点坐标，从而利用多层感知器（MLP）网络识别手势信息。3D 渲染模块实时处理手势和位移信息，并联合分析 3D 渲染场景信息，以实现光场的实时交互。我们的定向定量实验结果表明，我们提出的方法实现了针对不同显示器和各种 3D 渲染引擎的实时双手势交互。实验结果表明，我们提出的方法具有更好的交互速度和准确性。它有效地克服了双手势交互中常见的遮挡和速度慢的局限性，同时也改善了三维光场光场显示相对单一的交互模式。这项技术在实际应用中的潜力巨大，为三维光场光场显示交互的发展做出了重要贡献。

关键词：人工智能，双手交互实时交互，三维光场显示
中图分类号：TP399

Real-time Double hand Interaction for 8k 3D Light Field Display Based on implicit neural network

Deng Xinyao, Xing Shujun
(School of Electronic Engineering,Beijing University of Posts and Telecommunications)

Abstract: In this paper, we present a real-time 3D display interaction system. The system enables users to accurately interact with 3D objects in real-time display of 3D light fields. Our double-handed gesture recognition module obtains skeleton information of both hands in real-time and processes key point coordinates by normalization and positional encoding, so gesture information is recognized using the MLP network. The 3D rendering module processes hand gesture and displacement information in real-time, and jointly analyzes 3D rendering scene information to enable real-time interaction of the light field. The results of our directional quantitative experiments show that our proposed method achieves real-time double-handed interaction for different displays and various 3D rendering engines. Our proposed method offers better interaction speed and accuracy than previous methods. It effectively overcomes the limitations of occlusion and slow speed that are often associated with two-handed interaction and also improves the relatively single interaction mode of the 3D light field display. This technology's potential for real-world applications is significant, making it a valuable contribution to the development of 3D light field display interaction.

Key words: Artificial intelligence, real-time interaction with two hands, three-dimensional light field display.

0 引言

手势交互作为人机交互领域最为直观自然的交互方式之一，其灵活、直观、非接触性及符合人类日常交流习惯的特点使之为繁琐的界面设备交互提供了一种自然的替代方案[1]。随着机器学习，深度学习技术的发展，手势交互技术已成功应用许多领域，如 VR/AR[2]、三维光场显示[3]等。

作者简介：邓新堯（2000-），男，主要研究方向：计算机视觉，智能交互
通信联系人：邢树军（1985-），男，助理研究员，硕导，主要研究方向：三维可视化，高速光场渲染，三维交互引擎. E-mail: xsj@bupt.edu.cn

基于 RGB-D 图像的代表性方法锚—节点回归网络[4]（A2J）通过一个 ResNet50[5]作为主干网络，主干网络后衔接三个分支网络，分别来预测锚点与关节的平面偏移量，估算关节的深度和给出锚点权重建议。

随着三维显示技术的发展，手势交互与三维光场显示技术的结合发展迅速，Li[6]等在手臂和悬浮 3D 模型进行接触检测后，通过蓝牙传递信号使指套振动，从而完成力触觉交互，但只是进行了坐标转换，判断手指与三维显示物体是否触碰到，没有额外功能。

目前，应用于光场真三维显示设备的交互手段多为单手交互，与单手交互相比，双手更符合人类的交互直觉，并能完成更多的复杂交互动作，在光场上打造更好的交互体验。

一些方法进行了双手交互的探索，但由于双手会互相遮挡造成信息干扰，以往的双手识别多借助于去遮挡算法的优化进行去干扰，效果并不理想，并且有较高的算力需求。Aboukhadra 等人提出一种新颖的基于 GCN[7]和 Transformer[8]的网络框架 THOR-Net[9]，该框架能够实现现实场景中双手和物体的三维重建，并且采用了自监督的训练方法建立手部表面纹理。

尽管目前国内外进行了一些交互式手势估计的研究，但基于 RGB 或 RGB-D 图像的手势识别技术占用算力较高，实现三维光场显示同时实时手势检测的需求较为困难；面对光场显示设备多数仅考虑到单手交互情景，不满足双手精准交互的需求；现有的双手手势识别工作识别速度慢，识别准确度低等问题亟待解决。

本文提出一种基于隐式神经网络的 8K 三维光场显示实时双手交互系统，通过 Leap motion 提取手部骨骼信息精简网络模型输入，在三维光场显示渲染同时进行实时双手手势检测工作。可精准实现双手大范围内与光场三维显示设备中 3D 物体的实时交互。交互帧率达到 35 帧/s。

1 方法

整体工作流程如图 1 所示，在 leap motion 终端检测到手部信息后，对信息先后进行坐标处理和位置编码的拓展信息维度操作，将处理后数据进入 MLP 网络进行实时识别，结合三维光场渲染信息得出交互逻辑，三维光场显示进行渲染完成实时交互。

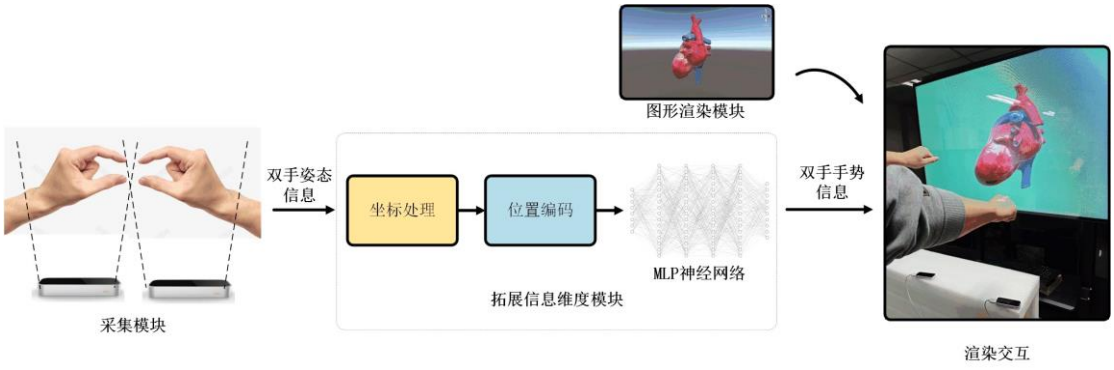

图 1 手势交互流程图

Fig. 1 Flow chart of gesture interaction

本文中方法应用的手部关键点如图 2 所示，分别取每根手指的掌骨，近端指骨，中间

指骨和远端指骨的远心端。其中大拇指由于骨骼分布不同记三个点，结合掌心点共 20 个三维点坐标，将这 20 个点记为集合 P，其中 $P = \{(x_i, y_i, z_i) \mid i = 1, 2, ..., 20\}$。

图 2 采集的手部关键点示例

Fig. 2 Example of the collected hand keypoints

1.1 坐标处理

为了更精确地获得手势的局部信息，将手部骨骼关键点坐标集合 P 中各点经过平移旋转操作，转化为以手掌掌心为原点，垂直手掌方向为 Y 轴，X 轴和 Z 轴在手掌的水平面上的新坐标系中坐标集合 $P' = \{(x'_i, y'_i, z'_i) \mid i = 1, 2, ..., 20\}$。

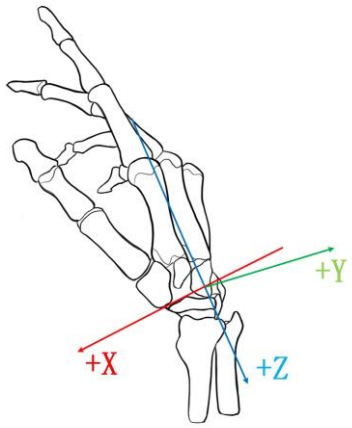

图 3 以掌心为原点的新坐标系

Fig. 3 The new coordinate system with the palm center as the origin

以掌心为原点的新坐标系如图 3 所示，以掌心点作为原点 O'，选取两个互相垂直的手部向量作为手部坐标系的 Y 和 Z 轴，关键点坐标集合 P 在不同坐标系间的变换可分解为平移和旋转两个过程，如图 4 所示。

对于平移操作，使用一个平移向量 T=(tx, ty, tz)T 来表示在 x、y 和 z 轴上的平移量，其中 T 为手部坐标系原点 O'到原坐标系原点 O 的向量 O'O。对集合 P 中的每个元素进行平移操作，得到平移后的新集合 $P^T = \{(x_i + tx, y_i + ty, z_i + tz) \mid i = 1, 2, ..., 20\}$。对每个元素的坐标进行平移操作时，分别将平移向量的对应分量加到原始坐标的对应分量上，如图 4 所示，通过对 P 中的每个元素平移操作，我们可以将集合 P^T 看作中间坐标系 A 下的坐标集合。

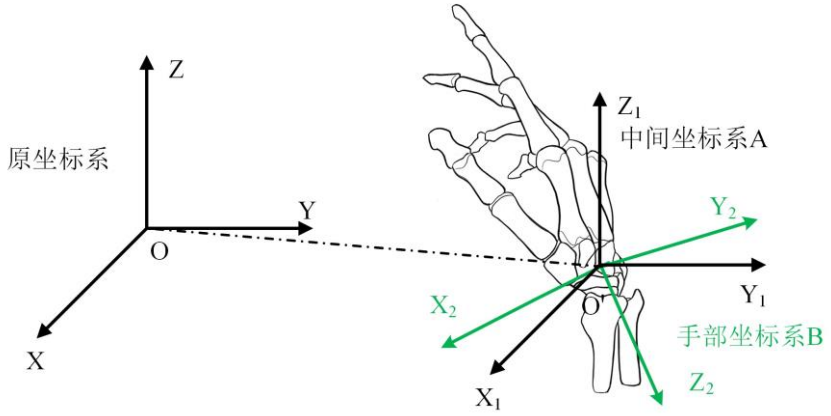

图 4 原坐标系，中间坐标系 A 和手部坐标系 B

Fig. 4 Original coordinate system, intermediate coordinate system A and hand coordinate system B

接下来，对平移后的新集合 PT 进行旋转操作。通过定义旋转矩阵 R，将旋转矩阵应用于平移后的新集合 PT，得到最终的旋转后的坐标 $P' = \{(R * (x_i + tx, y_i + ty, z_i + tz)) \mid i = 1, 2, ..., 20\}$。

新坐标系 A 原点与手部坐标系 B 原点重合，坐标系 A 各轴与 Leap motion 坐标系相同，坐标系 B 的各坐标轴向量基于坐标系 A 表示，因此坐标系 A 为世界坐标系，旋转矩阵 R 的求解为坐标系两次投影变换过程，记点 P 在坐标系 A 上的坐标为 P1（Xa,Ya,Za），在坐标系 B 上的坐标为 P2（Xb,Yb,Zb）。

由于坐标系 B 的各坐标轴向量基于坐标系 A 表示，坐标系 B 下的坐标可通过如图 5 所示的两次投影过程转化为坐标系 A 下的坐标，其逆投影为坐标系 A 坐标集合转化为坐标系 B 下坐标集合的过程，即为需要的旋转矩阵 R。

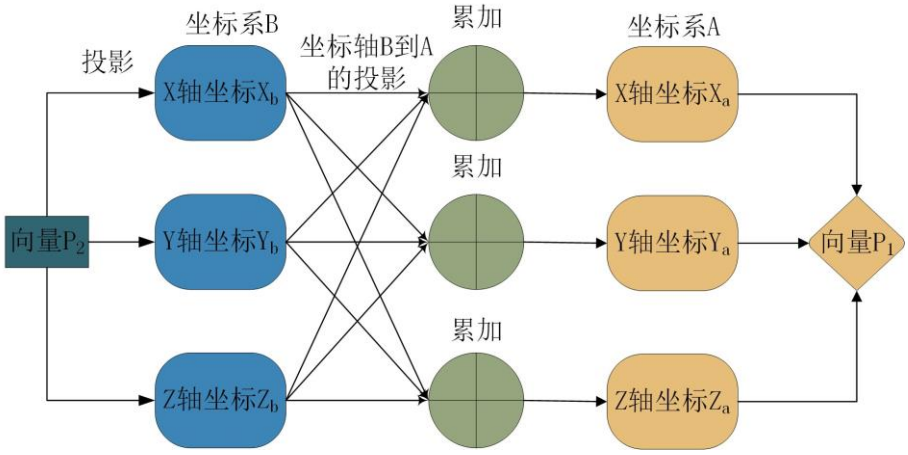

图 5 坐标系变换两次投影过程

Fig. 5 The projection process of the coordinate system transformation twice

坐标系 B 到坐标系 A 的投影表示为坐标系 B 各轴单位向量组成的旋转矩阵 ${}^{A}_{B}R$ ，其中 ,, 为坐标系 B 各轴单位向量在坐标系 A 下的表示

$$ {}^{A}_{B}R = [[X_2][Y_2][Z_2]] = \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix} \quad (1) $$

${}_B^A R$ 的每一行依次可视为坐标系 B 中各轴在坐标系 A 中的投影，坐标系 B 内向量到坐标系 A 内向量的坐标投影过程用公式表示

$$
\begin{aligned}
{}_B^A R \times P_2 &= \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix} \times \begin{bmatrix} X_b \\ Y_b \\ Z_b \end{bmatrix} \\
&= \begin{bmatrix} X_a \\ Y_a \\ Z_a \end{bmatrix}
\end{aligned} \tag{2}
$$

其中 P_2 表示 P 点在坐标系 B 中坐标，P_1 表示 P 点在坐标系 A 中坐标，从坐标系 A 内向量到坐标系 B 内向量可视为一个逆投影过程，即

$$
P_2 = {}_B^A R^{-1} \times P_1 \tag{3}
$$

${}_B^A R$ 为单位正交矩阵，根据单位正交矩阵性质 ${}_B^A R^{-1} = {}_B^A R^T$，逆投影过程为

$$
\begin{aligned}
P_2 &= {}_B^A R^{-1} \times P_1 \\
&= {}_B^A R^T \times P_1 \\
&= \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix} \times \begin{bmatrix} X_a \\ Y_a \\ Z_a \end{bmatrix} \\
\text{旋转矩阵} R &= \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix}
\end{aligned} \tag{4}
$$

将左右手的关键点集合 P 各点通过平移向量 T 和旋转矩阵 R 进行处理，得到归一化后坐标集合 $P' = \{(R * (x_i + tx, y_i + ty, z_i + tz)) \mid i = 1, 2, ..., 20\}$。使网络更关注于局部特征信息。

1.2 位置编码

位置编码是一种将位置信息嵌入到神经网络中的技术，通常用于处理与位置相关的任务，例如 3D 渲染[10][11]、自然语言处理[8]等。

由于标准 Mlp 不适合这些低维基于坐标的视觉和图形任务[12]，并且 Mlp 难以学习高频函数，这种现象被文献称为"谱偏差"（Spectral bias）[13][14]。一些研究发现，对输入坐标进行位置编码可以允许 Mlp 网络表示更高频率的内容[10][11]，这是由于它能够为输入数据添加一些空间信息，使得神经网络可以更好地理解输入数据之间的关系。

本文中的方法应用 Nerf 中的位置编码方法，对双手归一化三维坐标集合 P'各点进行编码，将其映射到高维坐标集合 P"，从而提高各手势信息的区分度，实验证明这是有效的。

$$
\begin{aligned}
\gamma(p) = (&\sin(2^0 \Pi p), \cos(2^0 \Pi p), \ldots \\
&\sin(2^{L-1} \Pi p), \cos(2^{L-1} \Pi p))
\end{aligned} \tag{5}
$$

函数 γ(·)分别应用于三维坐标中的三个坐标值得到三维坐标的高维映射。

在本方法的实验中，L=11 时，网络运行得到最优结果。

1.3 Mlp 网络

Mlp 网络采用一个五层的 Mlp 网络，采用交叉熵损失函数作为损失函数，对网络进行 Adam 优化。位置编码处理后高维坐标集合作为网络输入，网络输出各手势的判别概率，概率权重最高即为网络预测手势。

1.4 三维光场显示交互

三维光场显示设备接收手势动作信息及双手移动信息结合此刻三维光场渲染场景信息联合做出渲染决策，实时改变三维光场渲染，完成双手控制模型旋转、放大与分解等操作。

2 结果

实物系统如图 6 所示，上方为我们所使用的三维光场显示设备，用于检测手部骨骼信息的 leap motion 被置于三维光场显示器下方平面。

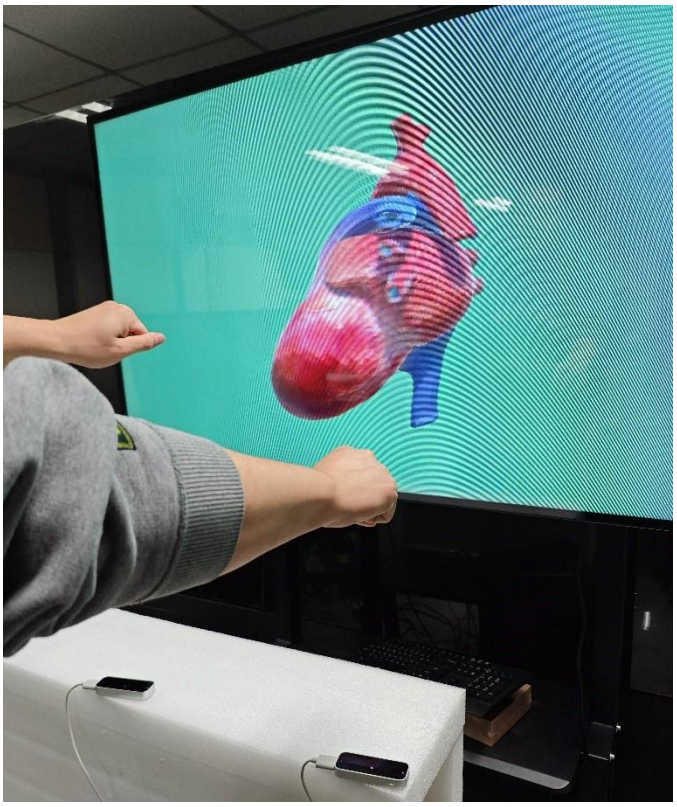

图 6 系统实物图

Fig. 6 Physical picture of the system

三维光场交互效果如图 7 所示。心脏模型的旋转交互实验如图 7(a)和 7(b)所示，当双手手势为旋转指令手势时，心脏模型随右手移动实时旋转；模型放大交互实验如图 7(c)和 7(d)所示，当双手为模型放大缩小指令时，模型随双手方位关系实时做出相应模型放大缩小操作。为了验证本文方法在双手遮挡情况下的识别效果，选取 3 维城市景观场景作为展示，场景缩放控制如图 7(e)和 7(f)所示，场景随双手位置关系进行放缩操作，如图 7(f)所示，在双手发生明显的遮挡关系时，系统仍可以实现精准的交互效果。头骨模型分解展示如图 7(g)和 7(h)所示，当双手做出分解指令时，头骨模型随双手运动做出精准的分解操作。我们的实

验表明，所提系统可实现双手对 3D 场景的精准控制交互，在 3D 显示分辨率为 3840×2160 像素下，交互速率可达 35 帧/s，交互精度 0.7mm。

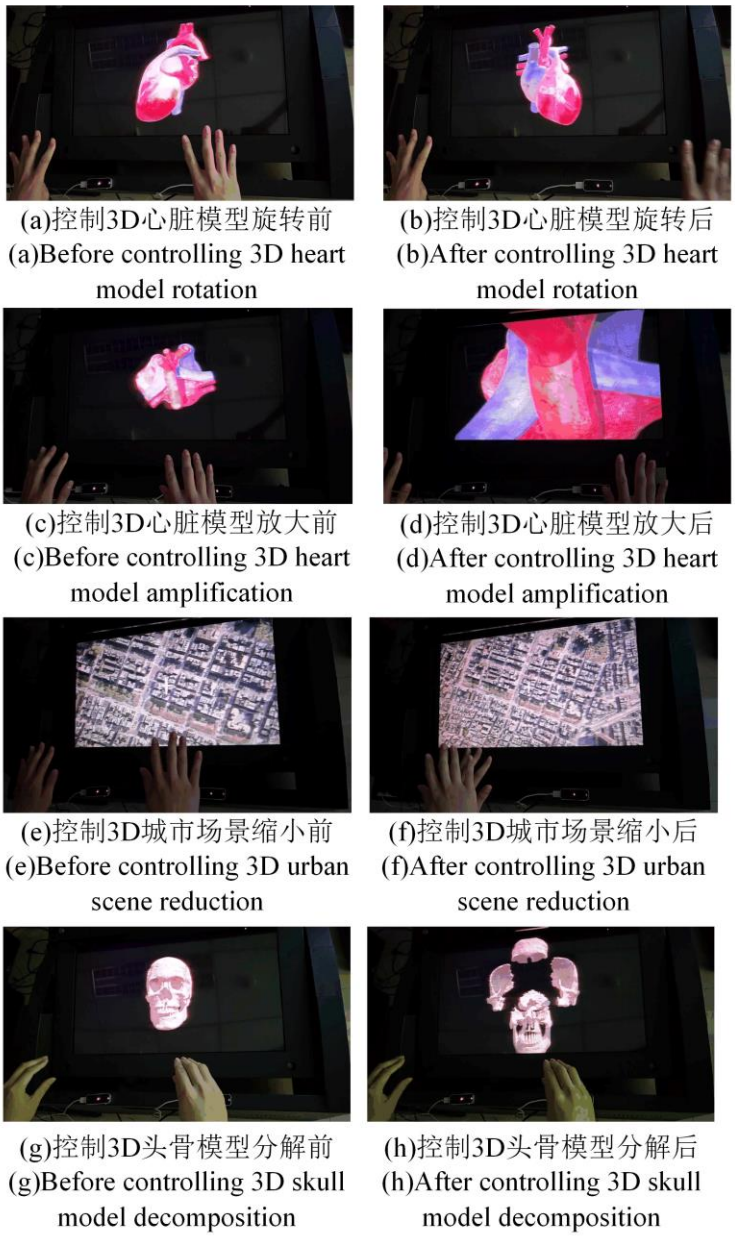

图 7 手势交互效果图

Fig. 7 Effect of gesture interaction

为了验证方法的有效性，我们开展了一项研究，旨在探究将多种 3D 渲染引擎、不同的 3D 光场显示器以及我们提出的手势识别方法相结合的有效性。如图 8 所示，结果非常成功，我们在 75 英寸的光场显示器上部署了我们的方法，并实现了出色的交互效果。

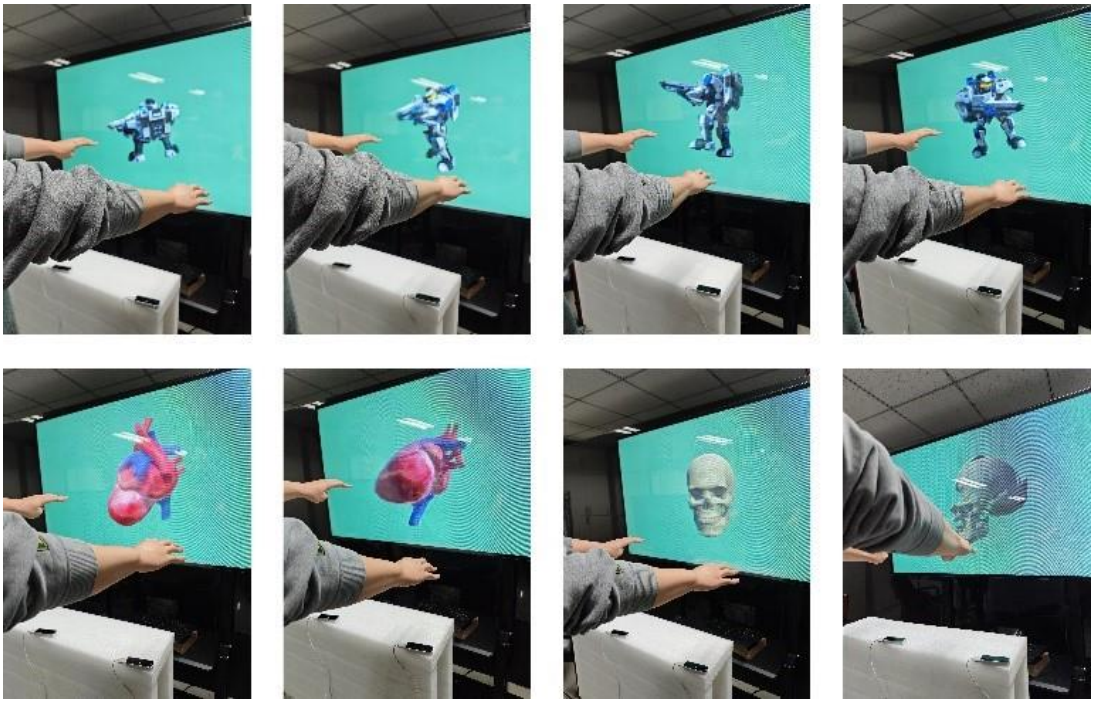

图 8 我们的方法在 75 英寸光场显示器上的交互效果

Fig. 8 The interaction effect of our method on a 75-inch light field display

为了验证识别的可行性，本研究在自建包含十种双手手势在不同方位（手掌距 leap motion 高度 5-50cm，水平范围 80cm 的）的数据集上进行了验证，本文方法取得了 88.3287% 的准确率。

为了验证坐标处理操作的有效性，本研究分别对未坐标处理与坐标处理后的骨骼关键点数据输入网络进行处理。为了数据的可靠，每次将实验网络训练十次获得平均正确率。如图 9 所示，本研究团队发现，经过坐标处理处理后实验均产生了更加可靠的实验结果，这是由于坐标处理操作使网络更关注于手势的局部信息，从而产生更为正确的结果。

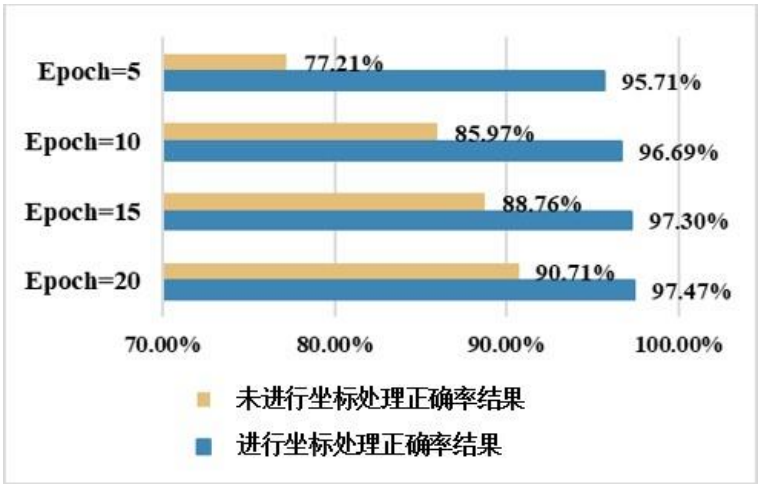

图 9 坐标处理影响准确率

Fig.9 Coordinate processing influences the accuracy results

为了验证位置编码操作的有效性，本研究进行了实验对比了进行了归一化操作后，进行或不进行位置编码时手势检测正确率的结果，为了寻找位置编码的最佳参数，对数据进行不同的位置编码处理。实验结果如图 10 所示，可以看到，对数据进行位置编码操作提升了网络分类的正确率，这证明了位置编码将手势信息映射到高维坐标，从而提高各手势信息的区

分度的有效性；同时，不同的位置编码参数对网络运行效果有不同的影响，在 L=11 时，网络达到最高准确率，综合考虑网络运行效率，采用 L=11 作为位置编码参数。此时，每一个三维坐标被映射为一个 66 维坐标。

为了评估该方法在识别双手手势方面的有效性，选取了 8 种不同的手势，与当前最先进的目标检测算法之一 ——YOLOv8 进行对比。如图 11 所示，实验结果表明，在识别双手手势时，所提出的方法在准确率和稳定性方面均优于 YOLO。

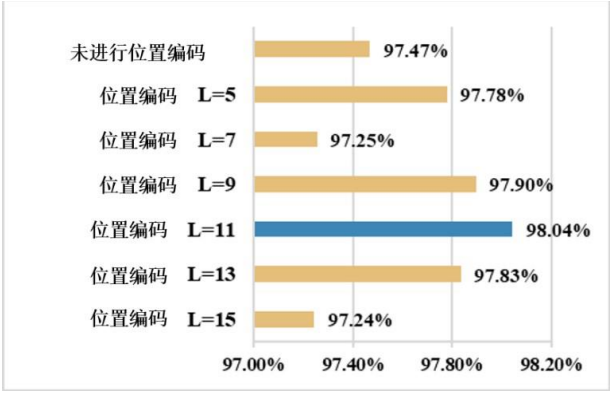

图 10 位置编码影响准确率

Fig.10 Position encoding affects the accuracy

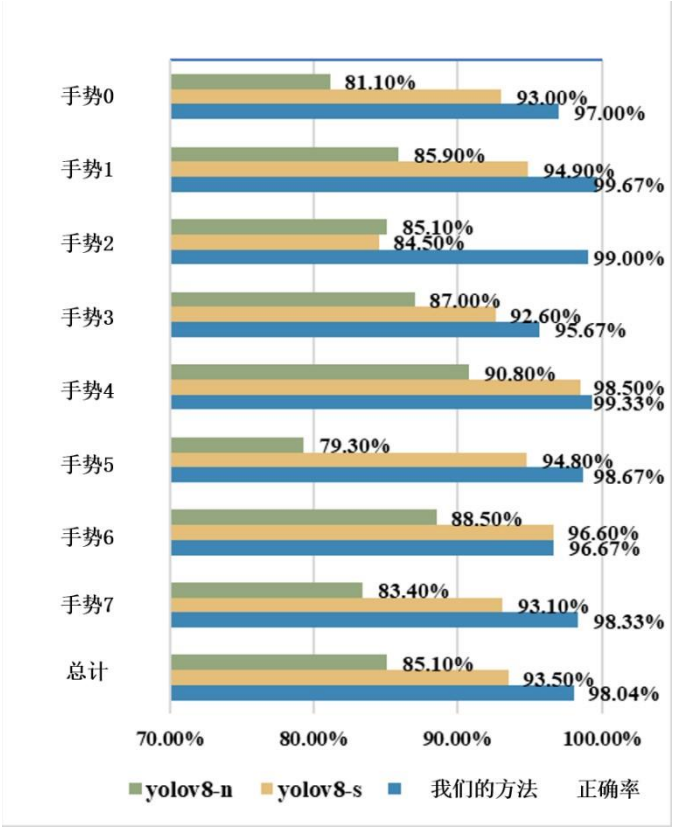

图 11 不同方法的手势识别效果

Fig.11 Gesture recognition effect of different methods

更详细的结果在如图 12 的混淆矩阵中给出，混淆矩阵显示，本文方法对每个手势分类都拥有较好的鲁棒性，并显示出了一定的泛化能力，使之可以使用在其他应用骨骼数据识别的场景。

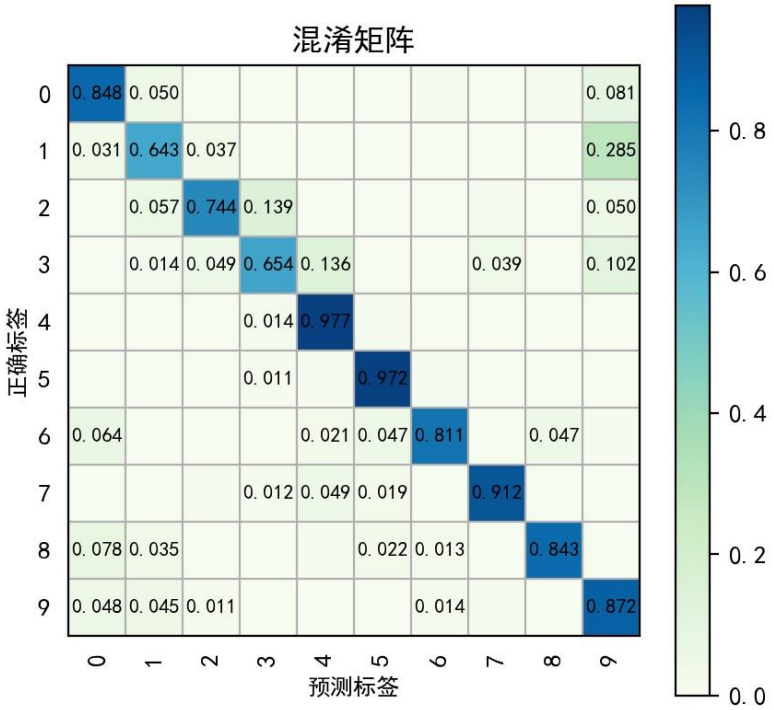

图 12 本方法在 10 个手势组成数据集下的混淆矩阵

Fig 9 Confusion matrix of the proposed method under a dataset composed of 10 gestures

3 结论

本文提出一种光场三维显示下基于位置编码与全连接网络的双手大范围精准实时交互方法，通过使用两个 Leap motion 分别获取双手骨骼信息来进行双手姿态检测，从源头上去除了双手检测工作中的遮挡与干扰。同时，我们的方法通过一个类归一化操作，使网络更加关注于手势的局部关键信息；通过位置编码拓展输入数据维度，增加检测工作的鲁棒性。经过实验证明，该方法有效实现了光场三维显示下双手大范围精准实时交互，交互帧率达到 35 帧/s；提升了双手光场交互工作的检测准确率与交互效果。所提系统为三维光场显示双手交互提供了更多可能性。

[参考文献] (References)

[1] GUO L, LU Z, YAO L. Human-Machine Interaction Sensing Technology Based on Hand Gesture Recognition: A Review[J/OL]. IEEE Transactions on Human-Machine Systems, 2021, 51(4): 300-309.
[2] LI Y, HUANG J, TIAN F. Gesture interaction in virtual reality[J/OL]. Virtual Reality & Intelligent Hardware, 2019, 1(1): 84-112.
[3] LIN X yu, XING Y, ZHANG H le. Real-time floating 3D display interaction system based on gesture recognition by leap motion[J/OL]. Chinese Journal of Liquid Crystals and Displays, 2022, 37(5): 654-659.
[4] XIONG F, ZHANG B, XIAO Y. A2J: Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation from a Single Depth Image[A/OL]. arXiv, 2019[2023-04-17].
[5] HE K, ZHANG X, REN S, 等. Deep Residual Learning for Image Recognition[A/OL]. arXiv, 2015[2023-04-17].
[6] LI. Suspended Three-Dimensional Display Haptic Interaction Based on Leap Motion[J]. Acta Optica Sinica: 40.
[7] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[A/OL]. arXiv, 2017[2023-04-17].
[8] VASWANI A, SHAZEER N, PARMAR N. Attention Is All You Need[A/OL]. arXiv, 2017[2023-04-17].
[9] ABOUKHADRA A T, MALIK J, ELHAYEK A. THOR-Net: End-to-end Graformer-based Realistic Two Hands and Object Reconstruction with Self-supervision[A/OL]. arXiv, 2022[2023-04-17].
[10] MILDENHALL B, SRINIVASAN P P, TANCIK M. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis[A/OL]. arXiv, 2020[2023-04-18].

[11] ZHONG E D, BEPLER T, DAVIS J H. Reconstructing continuous distributions of 3D protein structure from cryo-EM images[A/OL]. arXiv, 2020[2023-04-22].
[12] JACOT A, GABRIEL F, HONGLER C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks[A/OL]. arXiv, 2020[2023-04-22].
[13] RAHAMAN N, BARATIN A, ARPIT D. On the Spectral Bias of Neural Networks[A/OL]. arXiv, 2019[2023-04-22].
[14] BASRI R, GALUN M, GEIFMAN A. Frequency Bias in Neural Networks for Input of Non-Uniform Density[A/OL]. arXiv, 2020[2023-04-22].