Journal of System Simulation

Volume 29 | Issue 11 Article 3

6-5-2020

Pose Estimation Using Convolutional Neural Network with Synthesis Depth Data

Wang Song
1. Jiangnan University, Wuxi 214122, China;;

Fuchang Liu
2. Hangzhou Normal University, Hangzhou 311121, China;;

Huang Ji
2. Hangzhou Normal University, Hangzhou 311121, China;;

Weiwei Xu
3. Zhejiang University,Hangzhou 310058, China;

See next page for additional authors

Pose Estimation Using Convolutional Neural Network with Synthesis Depth Data

Abstract

Abstract: 3D scenes can be reconstructed more easily and rapidly with depth camera. However, it is difficult to retrieve items in 3D scenes from a single view depth image, especially for the pose estimation. In this paper, *we present a method of pose estimation using convolutional neural network with synthesis depth data*, which predicts the items' pose in 3D scenes by regression. This is achieved by (i) *synthesizing large amount of depth images with different pose for linear regression using 3D model,* (ii) *designing a class-dependent linear regression framework, which estimates the object's pose from different classes separately*, (iii) *reforming LeNet-5 model by representing the loss layer as a linear regression form*. The proposed algorithm is demonstrated on different data sets and achieves higher accuracy (average error 4.3°) than other algorithms.

Keywords

pose estimation, convolutional neural networks, depth images, scene reconstruction

Authors

Wang Song, Fuchang Liu, Huang Ji, Weiwei Xu, and Hongwei Dong

Recommended Citation

Wang Song, Liu Fuchang, Huang Ji, Xu Weiwei, Dong Hongwei. Pose Estimation Using Convolutional Neural Network with Synthesis Depth Data[J]. Journal of System Simulation, 2017, 29(11): 2618-2623.

第 29 卷第 11 期 系统仿真学报© Vol. 29 No. 11
2017 年 11 月 Journal of System Simulation Nov., 2017

基于卷积神经网络的深度图姿态估计算法研究

王松[1]，刘复昌[2]，黄骥[2]，许威威[3]，董洪伟[1]
（1.江南大学，江苏 无锡 214122；2.杭州师范大学，浙江 杭州 311121；3.浙江大学，浙江 杭州 310058）

摘要：随着深度相机的应用，三维场景的重建越来越简单、快速。从单视角的深度场景图像中检索出物体还是比较困难，特别是物体的姿态估计。提出了一种*基于卷积神经网络的深度图像姿态估计算法*。该算法采用了回归估计来实现姿态的估计。*通过 3D 模型合成大量不同姿态的深度图像样本*，从而解决回归估计需要稠密采样的训练数据问题。对于不同类别的物体，分别用线性回归估计来拟合姿态函数。*在基于 LeNet-5 模型上修改了卷积神经网络的结构，使得该网络适用于回归估计*。实验结果表明：我们的方法取得了平均误差约 4.3°的估计结果，优于其他文献的方法。

关键词：姿态估计；卷积神经网络；深度图像；场景重建
中图分类号：TP391.4 文献标识码：A 文章编号：1004-731X (2017) 11-2618-06
DOI: 10.16182/j.issn1004731x.joss.201711003

Pose Estimation Using Convolutional Neural Network with Synthesis Depth Data

Wang Song[1], *Liu Fuchang*[2], *Huang Ji*[2], *Xu Weiwei*[3], *Dong Hongwei*[1]
(1.Jiangnan University, Wuxi 214122, China; 2.Hangzhou Normal University, Hangzhou 311121, China;
3.Zhejiang University,Hangzhou 310058, China)

Abstract: 3D scenes can be reconstructed more easily and rapidly with depth camera. However, it is difficult to retrieve items in 3D scenes from a single view depth image, especially for the pose estimation. In this paper, *we present a method of pose estimation using convolutional neural network with synthesis depth data*, which predicts the items' pose in 3D scenes by regression. This is achieved by (i) *synthesizing large amount of depth images with different pose for linear regression using 3D model,* (ii) *designing a class-dependent linear regression framework, which estimates the object's pose from different classes separately*, (iii) *reforming LeNet-5 model by representing the loss layer as a linear regression form*. The proposed algorithm is demonstrated on different data sets and achieves higher accuracy (average error 4.3°) than other algorithms.

Keywords: pose estimation; convolutional neural networks; depth images; scene reconstruction

引言

随着深度相机的成本越来越低。不少领域都会使用该设备采集数据，例如：室内场景恢复、三维表情重建。此外，单视角的深度图像也被应用于场景检索。这里涉及到场景中物体的姿态估计问题。准确的姿态估计可以减少重建过程中的人机交互过程，使整个算法流程具有更好的自动性。因此，姿态估计是场景重建的重要环节。传统的多视角姿态估计算法主要基于两个不同视图之间的匹配，即对两个视图之间的转换关系做出估计。姿态估计问题可以分为 4 类：从二维到二维的估计问题、从三

收稿日期：2016-05-08 修回日期：2016-07-14；
基金项目：国家自然基金青年科学项目(61502133)，
浙江省自然基金一般项目(LY16F020029)；
作者简介：王松(1991-)，男，安徽阜阳，硕士生，研究方向为深度学习和计算机图形图像；刘复昌(1982-)，男，江苏南京，博士，讲师，研究方向为计算机图形图像与机器学习。

维到三维的估计问题、从二维透视投影到二维透视投影的姿态估计问题、从二维透视投影到三维的姿态估计问题。

与传统的多视角姿态估计不同，基于单视角深度图像的姿态估计主要有基于模型的姿态估计和基于学习的姿态估计两种方法。基于模型的姿态估计方法是通过比对真实图像和合成图像，进行相似度计算更新物体姿态。目前基于模型的方法为了避免在全局状态空间中进行优化搜索，一般都将优化问题先分解成多个局部特征的匹配问题，非常依赖于局部特征的准确检测。当噪声较大无法提取准确的局部特征的时候，该方法的鲁棒性受到很大影响。基于学习的思想借助于机器学习(Machine Learning)或深度学习(Deep Learning)方法，从事先获取的不同姿态下的训练样本中学习二维观测与三维姿态之间的对应关系，并将学习得到的决策规则或回归函数应用于样本，所得结果作为对样本的姿态估计，一般采用全局观测特征，可以保证算法具有较好的鲁棒性。然而这一类方法的姿态估计精度很大程度依赖于训练的充分程度。要想比较精确地得到二维观测与三维姿态之间的对应关系，就必须获取足够稠密的样本来学习决策规则和回归函数。而一般来说所需要样本的数量是随状态空间的维度呈指数级增加的，对于高维状态空间，事实上不可能获取进行精确估计所需要的密集采样。因此，无法得到密集采样而难以保证估计的精度与连续性，是基于学习的姿态估计方法无法克服的根本困难。

本文主要研究基于学习的姿态估计方法。我们通过渲染 3D 模型以获取大量的深度数据，并同时得到每张深度图的姿态信息，用来作为训练数据。这样就解决了基于学习方法的姿态估计无法得到精确的稠密采样的问题，并通过实验验证了合成数据对训练得到的模型对真实测试数据的有效性。通过卷积神经网络的方法训练得到一个能自动回归出对应类别的深度数据的姿态信息。神经网络训练的模型相对于人为设计的特征或者算法具有更加客观地结果。

1 相关工作

姿态估计是计算机视觉研究领域中的重要问题。基于学习的姿态估计方法源于姿态识别方法的思想。姿态识别需要预先定义多个姿态类别，每个类别包含了一定的姿态范围；然后为每个姿态类别标注若干训练样本，通过模式分类的方法训练姿态分类器以实现姿态识别。从基于学习的姿态估计方法的问题转本质来分，可将姿态估计问题转化为分类问题和回归问题两种。

基于分类的姿态估计，Poggio 等在文献中采用基于人工神经网络的二维视图匹配方法，选择最相似的二维视图的姿态参数作为姿态估计的结果，其中二维视图为三维模型在二维空间的投影[1]。Khotanzad 等在文献中采用基于人工神经网络对二维视图中的轮廓进行匹配，减少了二维视图中的冗余特征[2]。Savarese 等提取二维图像特征，通过匹配已知三维姿态参数的二维图像特征来确定三维姿态参数[3]。Fanelli 等在文献中采用随机森林算法从深度图像中分离出人脸，对人脸区域进行匹配[4]。Shao 等在文献中对三维模型的深度图像分块，采用随机森林算法对多个深度图像块进行匹配[5]。

基于回归的姿态估计，Lu 等在文献中首先提出了 OI 迭代算法[6]。冷大炜等在文献中提出了一种基于从二维到三维的轮廓点对应的三维刚体目标的迭代姿态估计算法，该算法研究的是单目观测下三维刚体目标，优化了传统迭代估计 OI 算法中收敛半径小、收敛速度慢等问题[7]。Haralick 等在文献中提出了全局收敛迭代法来解决由二维视图向三维空间估计的算法[8]。邱丽梅利用人脸特征点初步估计人脸姿态，以初步估计值为初始值，通过线性回归迭代算法，精确估计 3D 人脸空间姿态[9]。张浩鹏等提出了一种用于单目视觉成像系统的姿态估计方法，在卫星数据集上的实验结果证明了该方法用于航天器姿态估计的有效性[10]。

深度学习，近年来，深度学习也被用于解决分类与回归问题[11-13]。由于卷积神经网络参数众多，

https://dc-china-simulation.researchcommons.org/journal/vol29/iss11/3
DOI: 10.16182/j.issn1004731x.joss.201711003

第 29 卷第 11 期 系统仿真学报 Vol. 29 No. 11
2017 年 11 月 Journal of System Simulation Nov., 2017

需要大量的数据进行训练，所以卷积神经网络适用于大规模数据集处理。网络的输入为图像，输出为数值或者向量。Krizhevsky 等采用卷积神经网络识别 ILSVRC-2010 数据集图片所含物体，准确率为 83%[14]。Massa 等采用卷积神经网络对渲染三维模型得到的彩色图像进行姿态估计[15]。Simard 等采用卷积神经网络识别 MNIST 手写体数字，准确率达 99.6%[16]。Gkioxari 等采用卷积神经网络进行人脸姿态估计及动作分类识别，在 PASCAL VOC 数据集测试时获得了较高的准确率[17]。Ouyang 等采用多源深度模型进行人体姿态估计，在 LSP、PARSE、UIUC 数据集上的测试结果比最佳结果好 8.6%[18]。除此之外还有很多使用深度学习的例子[19-[21]，本文尝试将深度学习的方法用于姿态估计问题。

2 基于回归的姿态估计算法概述

本文算法采用的卷积神经网络的输入为三维模型的深度图，输出为该三维模型的姿态参数。基于学习的姿态估计方法源于姿态识别方法的研究。姿态识别需要预先定义多个姿态类别，每个类别包含了一定的姿态范围；然后为每个姿态类别标注若干训练样本，通过模式分类的方法训练姿态分类器以实现姿态识别。这一类方法并不需要对物体进行建模，一般通过图像的全局特征进行匹配分析，可以有效的避免局部特征方法在复杂姿态和遮挡关系情况下出现的特征匹配歧义性问题。然而姿态识别方法只能将姿态划分到事先定义的几个姿态类别中，并不能对姿态进行连续的精确的估计。为了使获取的姿态参数是连续的数值，而不是离散的数值，所以本文的姿态估计是一个回归问题。本文卷积神经网络的结构可以先考虑借鉴主流的网络结构，再通过控制变量法来改变各层神经元个数、梯度下降时的学习速率等参数以求优化目标函数。图 1 为本文的算法流程图，其过程分为两个阶段：训练阶段和测试阶段。在训练阶段，我们将训练样本中的深度图像作为卷积神经网络的输入，对应的姿态参数作为卷积神经网络的输出。设置合适的学习速率，通过梯度下降法来优化目标函数。当目标函数降到预设阈值时，训练阶段结束。此时，我们训练出比较合适的卷积神经网络模型。在测试阶段，我们将测试样本中的深度图像作为卷积神经网络的输入，通过训练得到的权重计算出实际输出，与对应的理想输出进行比较即可判断姿态估计结果的好坏。

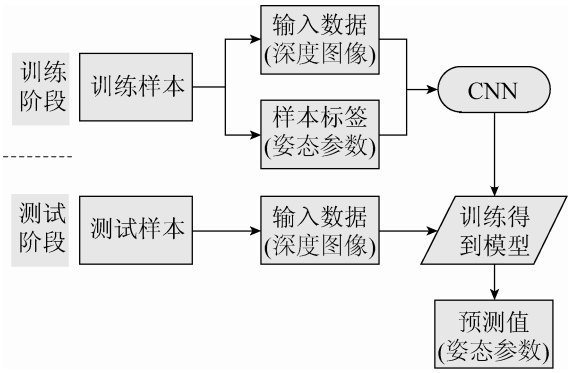

图 1 算法流程图
Fig. 1 Algorithm flow chart

3 基于卷积神经网络的回归框架设计

3.1 损失函数

本文所使用的损失函数是同一类别中所有样本估计值和预测值的欧氏距离平方的均值，也就是均方根误差(MSE)，公式如下：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i^{(j)}-Y_i^{(j)})^2$$

公式中 j 代表样本的类别。假设模型结果与测量值的误差满足正态分布，那么这个假设就是符合一般的客观统计规律的。

3.2 网络训练

本文的姿态估计问题是一种由二维视图向三维空间的估计(即 2–D perspective projection—3–D pose estimation)，它采用了卷积神经网络来对二维视图(即深度图像)进行姿态估计。卷积神经网络能够实现二维视图向三维空间的估计。本文中卷积神

经网络结构的设计参考 LeNet-5[23]经典网络结构，通过控制变量，逐渐改变调整各层之间的参数结构，使得调整后的模型更适用于本文的问题环境。在五层网络结构的基础上，增加了一层 pooling 层，修改了每层的 Feature Map 的数量，以及卷积核的大小，使得 LeNet-5 模型更适用于本文所应用的场景。同时我们修改了 LeNet-5 的 Loss 函数，把这个处理分类问题的模型修改成本文所处理的回归模型。图 2 为本文的网络结构图。

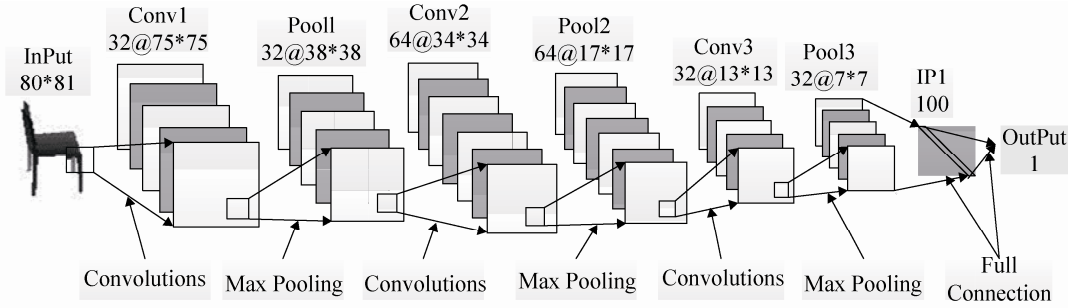

图 2 网络结构
Fig. 2 Network Structure

4 实验结果与分析

本文实验分为 4 个步骤：①获取三维模型数据集，通过渲染得到三维模型的深度图像；②可视化姿态的特征空间；③训练网络结构，获得权重参数；④显示结果，并分析误差。

4.1 数据获取

本文以渲染三维模型得到的深度图像为输入，姿态参数为输出。深度图及姿态参数的获取方法如下：在宽度为 w、高度为 h 的 OpenGL 窗口中渲染三维模型(三维模型位置固定)，坐标系如图 3 所示，xOz 平面为室内场景中的地面，z 轴正半轴所在的平行于 xOy 平面为三维模型的正面。设置视点 **V**，在视点 **V** 处模拟拍摄三维模型(读取深度缓存)得到深度图 **I**，记录此时的姿态参数 **P**。其中，姿态参数的角度归一化到 0~1 之间，即 $\theta':=\theta\div 360$。

4.2 评价标准

本文将实际值与预测值的误差作为评价标准。在实验中，通过卷积神经网络训练得到的模型，预测测试样本。求得测试集的平均误差作为衡量算法优劣的评价标准。

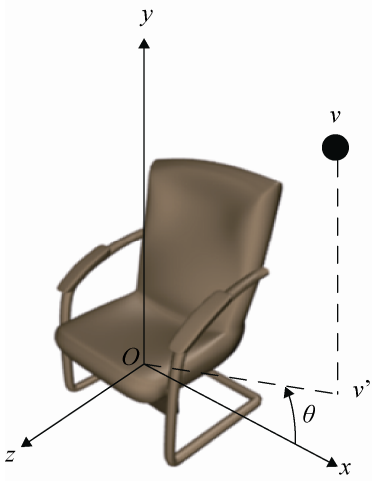

图 3 三维空间视点示意图
Fig. 3 Sketch map of 3D space viewpoint

4.3 部分实验结果展示及分析

以椅子模型为例，当椅子的各个部位充分展现时，预测效果较理想，如表 1 第 1、5、7 个实例所示；当椅子的部分部位被大面积遮挡时，预测效果欠佳，如表 2 第 2、6、8 个实例所示。据实验数据统计，本实例预测姿态参数的平均误差约 4.3°，低于文献[21]中的平均误差 7.3°。如果通过对深度图像中的物体进行场景的三维重建，重建的误差在视觉上影响较小。

第 29 卷第 11 期 系统仿真学报 Vol. 29 No. 11
2017 年 11 月 Journal of System Simulation Nov., 2017

表 1 理想预测结果

Tab. 1 Ideal prediction results

图例								
测试值	0.928 5	0.387 0	0.734 1	0.436 5	0.674 7	0.398 5	0.250 3	0.189 0
实际值	0.937 8	0.388 6	0.726 7	0.451 7	0.661 7	0.399 2	0.250 0	0.190 6
误差	0.009 3	0.001 6	0.007 4	0.015 2	0.013 0	0.000 7	0.000 3	0.001 6

表 2 不理想预测结果

Tab. 2 Unsatisfactory prediction results

图例								
测试值	0.899 2	0.225 7	0.093 8	0.671 2	0.170 0	0.392 2	0.635 1	0.603 5
实际值	0.963 3	0.010 0	0.029 1	0.596 9	0.020 6	0.446 1	0.001 7	0.493 3
误差	0.064 1	0.215 7	0.064 7	0.074 3	0.149 4	0.053 9	0.633 4	0.110 1

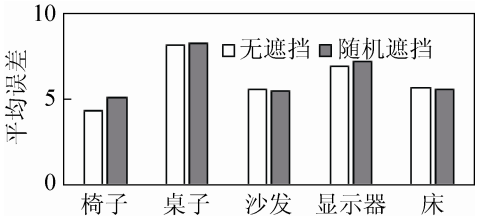

图 4 本文算法有无遮挡对比

Fig. 4 Occlusion contrast

实验中，我们对与遮挡问题做了几组对比实验。如图 4，我们将合成的训练数据做了随机遮挡处理，在测试中加入被遮挡的测试数据，结果显示，少量遮挡对实验结果影响不大。同时我们和已知的方法作了对比，如 Tulsiani，Malik[22]，和 Render for CNN[20]，实验中，我们按照文献[20]的评价标准并将准确率的计算规则更改为 $ACC_{\pi/18}$ 从而得到一组实验结果，如图 5，并且文献[20]和文献[22]的评价标准不变。可以看出本文算法相对于其他算法在精确度上更具有优势。

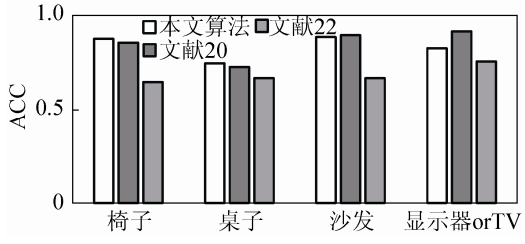

图 5 不同算法对比

Fig. 5 Comparison of different algorithms

5 结论

本文提出并实现了一种使用卷积神经网络对三维模型深度图像进行姿态参数预测的算法。首先，获取三维模型某种姿态下的深度图像，并记录该姿态下的姿态参数；然后，训练卷积神经网络结构中的参数(权重)，并使用梯度下降法优化目标函数；接着，使用完成训练的网络进行预测，将理想输出(记录的姿态参数)与实际输出进行误差分析。本实验结果表明，网络模型经过训练后能够较好地预测二维图像中物体的三维姿态。目前本实验处于初步阶段，未来我们面临的工作有：①在实际场景中考察姿态估计的效果；②实际场景中物体被其他物体遮挡的情况更加普遍，面对信息缺失的输入图像，我们可以考虑通过补全信息缺失部分(例如，使用形状玻尔兹曼机将图像有中遮挡的部分恢复完整[24])的思路实现准确的预测；③全连层的连接有望实现稀疏连接(即相邻两层之间的全连接权重中，绝大部分权重值为 0)，以进一步提高姿态预测的速度；④在保持当前的预测精确程度的前提下进一步简化网络结构，并寻找合适的方法调整学习速率，进一步减少训练时间及提高姿态估计的精确程度。

参考文献：

[1] Poggio T, Edelman S. A network that learns to recognize

3D objects[J]. Nature (S1476-4687), 1990, 343(6255): 263-266.

[2] Khotanzad A R, Liou J H. Neural network system for 3D object recognition and pose estimation from a single arbitrary 2D view[J]. Proceedings of SPIE - The International Society for Optical Engineering (S0277-786X), 1992, 1709: 107-118.

[3] Savarese S, Li F F. 3D generic object categorization, localization and pose estimation[C]// Proceedings / IEEE International Conference on Computer Vision. IEEE International Conference on Computer Vision. 2007: 1-8.

[4] Fanelli G, Weise T, Gall J, et al. Real Time Head Pose Estimation from Consumer Depth Cameras[M]// Pattern Recognition. Springer Berlin Heidelberg, 2011: 101-110.

[5] Shao T, Xu W, Zhou K, et al. An interactive approach to semantic modeling of indoor scenes with an RGBD camera[J]. Acm Transactions on Graphics (S0730-0301), 2012, 31(6): 439-445.

[6] Lu C P, Hager G D, Mjolsness E. Fast and globally convergent pose estimation from video images[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (S0162-8828), 2000, 22(6): 610-622.

[7] 冷大炜, 马洪兵, 孙卫东. 基于 2D-3D 泛轮廓点对应的三维刚体目标的迭代姿态估计[J]. 中国科学院大学学报, 2012，29(6): 821-828.(Leng D W，Ma H B，Sun W D． Rigid object based on general 2D-3D contour point Iterative pose estimation of a 3D correspondence［J］． Journal of Graduate University of Chinese Academy of Sciences，2012，29(6) : 821-828．)

[8] Haralick R M, Joo H, Lee C N, et al. Pose Estimation from Corresponding Point Data[J]. IEEE Transactions on Systems Man & Cybernetics (S0018-9472), 1989, 19(6): 1-84.

[9] 邱丽梅. 基于人脸特征点和线性回归的 3D 人脸姿态估计方法[J]. 三明学院学报, 2008, 25(4): 390-394.(Qiu L M. 3D Face Pose Estimation Based on Face Feature Points and Linear Regression[J]. Journal of Sanming University, 2008.)

[10] 张浩鹏, 姜志国. 基于姿态加权核回归的航天器姿态估计[J]. 北京航空航天大学学报, 2014, 40(4): 494-499.(Zhang H, Jiang Z. Spacecraft attitude estimation based on attitude-weighted kernel regression[J]. Beijing Hangkong Hangtian Daxue Xuebao/journal of Beijing University of Aeronautics & Astronautics, 2014, 40(4): 494-499.)

[11] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature (S1476-4687), 2015, 521(7553): 436-444.

[12] Schulz H, Behnke S. Deep learning[J]. KI-Künstliche Intelligenz(S1610-1987), 2012, 26(4): 357-363. .

[13] Schmidhuber J. Deep learning in neural networks: An overview[J]. Neural networks, 2015, 61: 85-117.

[14] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks [J]. Advances in Neural Information Processing Systems (S1049-5258), 2012, 25(2): 2012.

[15] Massa F, Aubry M, Marlet R. Convolutional Neural Networks for joint object detection and pose estimation: A comparative study[J]. Revista Brasileira De Farmacognosia (S0102-695X), 2014, 19(2a): 412-417.

[16] Simard P Y, Steinkraus D, Platt J C. Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis[C]// International Conference on Document Analysis and Recognition. IEEE Computer Society, 2003: 958.

[17] Gkioxari G, Hariharan B, Girshick R, et al. R-CNNs for Pose Estimation and Action Detection[J]. Computer Science (S1508-2806), 2014.

[18] Ouyang W, Chu X, Wang X. Multi-source Deep Learning for Human Pose Estimation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2014: 2337-2344.

[19] Smith S M, Brady J M. SUSAN—a new approach to low level image processing[J]. International journal of computer vision (S0920-5691), 1997, 23(1): 45-78.

[20] Su H, Qi C R, Li Y, et al. Render for CNN: Viewpoint Estimation in Images Using CNN s Trained with Rendered 3D Model Views[C]// IEEE International Conference on Computer Vision. IEEE, 2015.

[21] 王春雪. 基于图像的空间目标三维姿态估计研究[D]. 中国科学院大学(工程管理与信息技术学院), 2014.(C Wang.Three dimensional pose estimation of space target based on image[D].University of Chinese Academy of Sciences, 2014.)

[22] LeCun Y. LeNet-5, convolutional neural networks[J]. URL: http://yann. lecun. com/exdb/lenet, 2015.

[23] Tulsiani S, Malik J. Viewpoints and keypoints [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1510-1519.

[24] Eslami S M A, Heess N, Williams C K I, et al. The shape boltzmann machine: a strong model of object shape[J]. International Journal of Computer Vision (S0920-5691), 2014, 107(2): 155-176.