第 36 卷 第 6 期
2021 年 12 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 36 No. 6
Dec. 2021

DOI:10.13364/j.issn.1672-6510.20210030

基于改进的 YOLOv3 实现手势识别的人机交互方法

苏 静，刘兆峰，王 嫄，冯柯翔，王晓薇
（天津科技大学人工智能学院，天津 300457）

摘 要：随着人工智能技术的飞速发展，人机交互方法也发生了巨大的变化. 鉴于目前主要的人机交互方式仍是使用键盘、鼠标和触控板组合的传统交互方式，本文提出一种基于改进的 YOLOv3 实现手势识别的人机交互方法，通过 K-means 对标签的边界框进行聚类，然后运用 Mosaic 数据增强丰富小目标，最后采用自定义最小化边界框中心点距离的 GCDIoU 损失函数优化模型参数. 在自建数据集上进行实验验证，该模型针对手势小目标的检测准确率达到 98.87%，召回率达到 99.98%. 结果表明，Mosaic 数据增强应用于小目标检测具有很好的效果，而 GCDIoU 损失函数则加快了模型的收敛.

关键词：人工智能；人机交互；手势识别；YOLOv3
中图分类号：TP391.7 **文献标志码**：A **文章编号**：1672-6510(2021)06-0049-06

Human-Computer Interaction Method for Gesture Recognition Based on Improved YOLOv3

SU Jing，LIU Zhaofeng，WANG Yuan，FENG Kexiang，WANG Xiaowei
(College of Artificial Intelligence, Tianjin University of Science & Technology, Tianjin 300457, China)

Abstract: With the rapid development of artificial intelligence technology, human-computer interaction has also undergone tremendous changes. In view of the fact that most people are still using the traditional interaction method of keyboard, mouse and touchpad, in this article we propose a method of real-time gesture recognition and human-computer interaction based on the improved YOLOv3. Specifically, the bounding box of the label was first clustered through K-means. Then the small targets were enriched by Mosaic data enhancement, and finally the GCDIoU loss function that minimizes the distance between the center points of the bounding box was used. Through the above methods, the final model's detection accuracy for small gesture targets reached 98.87% , and the recall rate reached 99.98% . Therefore, Mosaic data enhancement has a good effect when applied to small target detection, while GCDIoU loss function speeds up the convergence of the model, and the effect is better.

Key words: artificial intelligence; human-computer interaction; gesture recognition; YOLOv3

计算机视觉技术正逐步进入人们日常生活，并推动着人机交互领域的升级和进化. 基于手势的人机交互中一个非常重要的环节是手势识别. 目前已实现的手势识别算法包括基于模板匹配、基于数据手套以及基于隐马尔科夫模型等. 但是，这些算法存在工序复杂、设备昂贵、计算量大等问题，导致训练模型的泛化能力差，因此很难达到实时检测的目的. 基于计算机视觉的手势识别方法可以有效克服传统方法的弊端，YOLO(you only look once)算法的出现，在目标检测领域取得极佳的检测效果[1].

目标检测是一个从深度学习快速发展中受益匪浅的领域，近年来人们实现了许多目标检测算法，包括 Faster RCNN 算法[2]、SSD 算法[3]、Mask-RCNN 算法[4]和 RetinalNet 算法等[5]. YOLO 算法是目标检测中 one-stage 的开山之作，开辟性地将物体检测任务直接作为回归问题处理，直接将候选和检测两个阶段

收稿日期：2021-01-30；**修回日期**：2021-08-14
基金项目：天津科技大学创新创业训练计划项目(202010057204)
作者简介：苏 静（1977—），女，天津人，副教授，sujing@tust.edu.cn

合而为一，“只需要看一眼”就可以知道在每张图片中有哪些物体以及物体的位置. 相较于之前的物体检测方法，YOLO 算法在保证一定的检测准确率的前提下达到了实时检测的速度.

在实验中将原始版本的 YOLOv3 算法[6]应用于手势实时识别领域，虽然可以取得较好的效果，平均准确率（mAP）达到 96.36%，准确率为 98.06%，但召回率仅为 81.59%，在检测的精度和速度上还有一定的提升空间，并且由于训练集数量和 GPU 显存的限制，导致训练时间过长.

为了解决上述问题，本文提出对 YOLOv3 算法进行改进：首先，YOLOv3 算法中的先验框是利用 K-means[7]聚类算法在 COCO 数据集预计算定义，但由于目标检测结果往往需要缩放到原始尺寸，自定义数据集中目标对象的尺寸与 COCO 数据集的并不完全相同，因此需通过对所有标签的边界框进行重新聚类计算先验框，提高模型识别的准确率和召回率；之后，针对 GPU 显存的限制，采用在输入端进行 Mosaic 数据增强[8]的方式解决，通过随机选取 4 张图片进行随机缩放、随机裁减、随机分布的方式进行拼接，一次训练 4 张图片可以大大提高训练效率，个人电脑单 GPU 就可以达到很好的效果，并且极大缩短了训练时间；最后，在输出端调整损失函数，结合 IoU、GIoU、CIoU、DIoU[9]定义 GCDIoU 损失函数，加快训练过程中的位置损失的下降速度，提高模型的训练效率.

1　目标检测算法

YOLOv3 是完全卷积网络（fully convolutional networks for semantic segmentation，FCN）[10]，有 75 个卷积层，还有跳跃连接层和上采样层，没有池化层，使用步长为 2 的卷积层对特征图进行采样. 作为完全卷积网络，YOLOv3 是可以处理各种输入图像的大小，但在实际实验中还要保持一个恒定的输入大小，其中一个很大的问题是，如果想进行批处理图像，那么就需要固定图像的宽高，这是将多个图像连接成一个大批处理所需要的.

网络通过步长因子对图像进行下采样. 例如，如果网络的步长是 32，则大小为 416×416 的输入图像将产生大小为 13×13 的输出图像. 一般来说，网络中任何一层的步长都等于该层的输出小于网络输入图像的因子.

在深度方面，特征图中有 $S\times S\left(B\times(5+C)\right)$ 个元素，B 表示每个单元格可以预测的边界框的数量，每个边界框都有 $5+C$ 属性，这些属性描述每个边界框的中心坐标、尺寸、对象性得分和 C 类置信度. YOLOv3 为每个单元格预测 3 个边界框.

YOLOv3 网络对输入图像进行下采样直到第一检测层，其中使用步长为 32 的层的特征映射进行检测，以因子 2 向上采样，然后与具有相同特征图大小的先前层的特征图连接. 另一个检测层的步长为 16，重复相同的上采样程序，并在步长为 8 的层进行最终检测. 上采样可以帮助网络学习细粒度特征，这些特征有助于 YOLOv3 检测小物体.

网络在 3 个不同的尺度上进行预测，检测层用于对 3 种不同尺寸（32、16、8）的特征图进行检测. 也就是说，当输入是 416×416 的情况下，模型在 13×13、26×26 和 52×52 的尺度上进行检测.

1.1　制作数据集

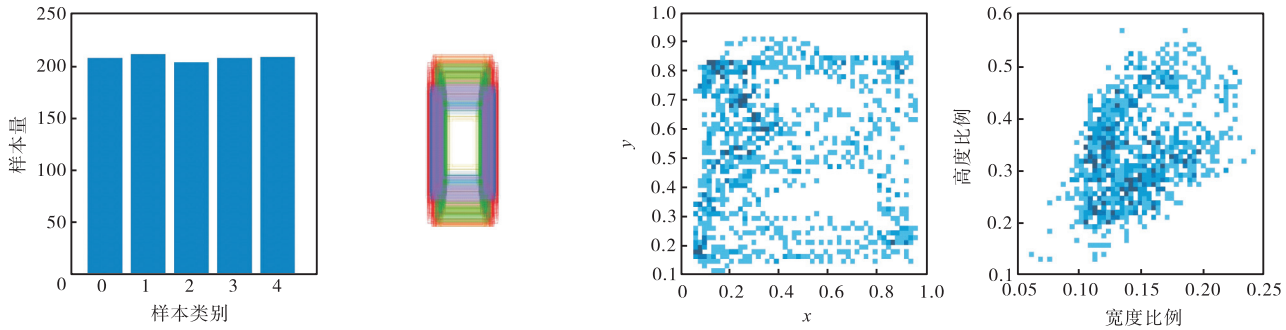
由于特定手势控制数据集的稀缺，导致互联网上暂时未有符合项目要求的公开数据集，因此实验中采用众包方式收集大量手势录像. 对于通过使用摄像头和计算机视觉技术捕获的 RGB 手势录像，首先要将其按照 10 帧/秒进行切分，然后按照 YOLO 算法数据集的格式通过 CVAT 进行人工标记手势区域，标注内容包括手势的类别信息以及位置信息.

标注工作完成后要对所有的标签进行 K-means 聚类处理，结果如图 1 所示. 图 1 中：(a) 为选择的 5 种手势图片数量，0—4 分别代表鼠标的单击、双击、右键、上滑和下滑；(b) 为通过 K-means 聚类算法得到的先验框，由于数据集中大部分都是手势 box，所以表现为竖直框；(c) 为手势位置在图片中的分布，空白部分为面部和胸前位置；(d) 为 box 的宽高所占图像宽高比例分布，呈现线性相关性.

1.2　数据增强

数据增强是由现有的、有限的训练数据通过变换创建新的训练样本，通过调整图像的参数推广到其他情况，允许模型适应更广泛的情况. 用于手势识别的数据集并不大，因此采用了广泛的数据增强，包括几何畸变、光照畸变、色彩空间调整和图像遮挡等等. 除此之外，在手势识别的项目训练中，由于手势大部分都是小目标，而 YOLOv3 针对小目标的检测效果并不理想，因此需要对数据进行一些处理. Mosaic 数据增强是参考 CutMix 数据增强[11]的方式，但 CutMix 数据增强只使用了 2 张图片进行拼接，而

Mosaic 数据增强则采用了 4 张图片，实验在此基础上又添加了色彩空间的调整（图 2），通过随机缩放、随机裁减、随机分布的方式进行拼接，并且随机添加边缘空白.



（a）不同类别样本　（b）聚类先验框可视化　（c）目标框坐标比例分布　（d）目标框宽高比例分布

图 1　K-means 聚类处理结果

Fig. 1　Clustering results of K-means

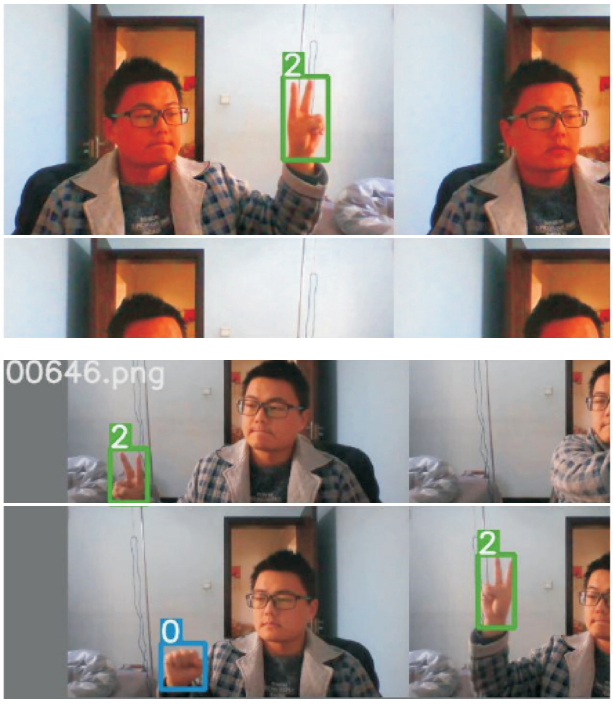


图 2　Mosaic 数据增强和色彩空间调整

Fig. 2　Data enhancement and color space adjustment

随机缩放增加了很多小目标，让网络检测效果更好，对小目标检测的准确率更高，并且通过 Mosaic 数据增强训练时可以直接计算 4 张图片的数据，使得小批次训练大小并不需要特别大，个人电脑的 GPU 就可以达到很好的效果.

1.3　损失函数

IoU_Loss[12]由旷视科技提出，把 4 个点（x，y，w，h）构成目标的真实边界框（box）看成整体进行回归，对两个 box 计算其交并比，但 IoU_Loss 有两个缺点：第一，当预测框和目标框不相交时，IoU = 0，不能反映距离的远近，此时损失函数不可导，因此无法优化两个 box 不相交的情况；第二，如果预测框和目标框的大小确定，只要两个 box 的相交值是确定的，其 IoU 值就是确定的，但并不能反映两个 box 是如何相交的.

基于 IoU 存在的问题，一个好的位置损失函数应该考虑重叠面积、中心点距离、box 宽高比这 3 个重要的几何因素.

GIoU 损失函数是通过 IoU 减去两个边界框闭包区域中不属于两个框的区域占闭包区域的比重表示. DIoU 和 CIoU 损失函数是将目标与先验框之间的距离、重叠率、尺度和 box 宽高比均考虑进去，使得目标框回归变得更加稳定，不会像 IoU 和 GIoU 一样出现训练过程中发散等问题.

综合考虑各种 IoU 的计算优劣势，定义 GCDIoU_Loss 为损失函数 $L_{\mathrm{GCDIoU\_Loss}}$ 进行优化.

$$L_{\mathrm{GCDIoU\_Loss}}=1-\mathrm{IoU}+\frac{|A_c-U|}{|A_c|}+\frac{\rho^2(b,b^{\mathrm{gt}})}{c^2}-\alpha v \tag{1}$$

式中：b 和 b^{gt} 分别为预测框 B 和预测框 B^{gt} 的中心点，$\rho(\cdot)$ 为欧式距离，c 为预测框 B 和预测框 B^{gt} 的最小外接矩形的对角线距离，v 用来度量宽高比的相似性.

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{\mathrm{gt}}}{h^{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2 \tag{2}$$

$$\alpha=\frac{v}{(1-\mathrm{IoU})+v} \tag{3}$$

1.4　pynput 鼠标控制

训练的模型可以通过调用电脑的默认摄像头获

取图像，并返回屏幕中识别的手势位置和标签，然后通过 pynput 库进行人机交互.

pynput 库是一个可以控制和监听输入设备，通过 pynput.mouse 包含用于控制和监听鼠标或触控板的类，通过 pynput.keyboard 包含用于控制和监听键盘的类，并且可以通过设置环境变量将其值作用于适合当前平台的键盘或鼠标后端.

为了确保在 Windows 上监听器和控制器之间的坐标一致，当系统缩放比例增加到 100% 以上时，最新版本的 Windows 支持运行旧版本的应用程序，这使旧的应用程序可以缩放，尽管比较模糊，但这种缩放导致鼠标监听器和控制器的坐标信息不一致，监听器接收物理空间坐标，但是通过控制器设计必须使用缩放的坐标. 要解决以上问题，需要设置 Windows 应用支持定位精度 DPI，并且这是一个全局设置，因此需要通过 ctypes 模块启用 DPI 感知.

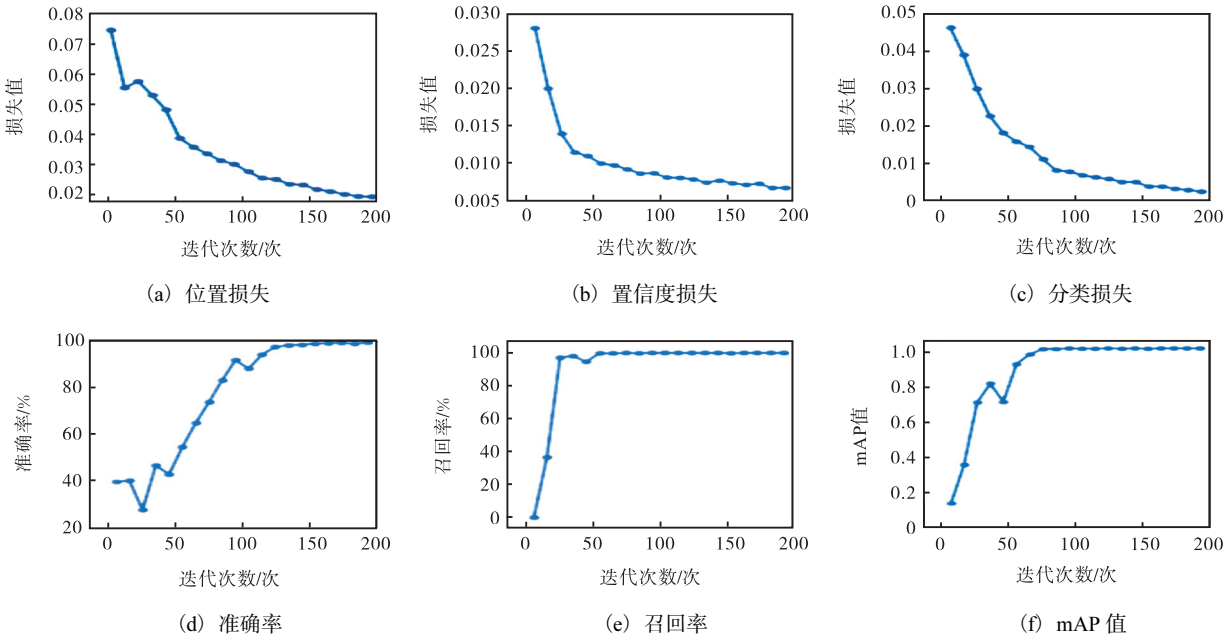
pynput 方法参考：click(button, count = 1)表示在当前位置发出按钮单击事件，move(dx, dy)表示将鼠标指针从其当前位置移出多个像素，position 表示返回鼠标指针的当前位置，press(button)表示在当前位置发出按钮按下事件，release(button)表示在当前位置发出按钮释放事件，scroll(dx, dy)表示发送滚动事件.

2　实验结果及处理

2.1　实验结果

实验环境：Window 10、Intel(R)Core(TM)i7-7700HQ CPU@2.80 GHz、NVIDIA GeForce GTX 1060 with Max-Q Design、Python 3.8、PyTorch 1.7.1、CUDA 10.2、cudnn 7.0 个人电脑.

实验结果如图 3 所示. 由图 3 可以发现，实验中所采用的方法得到了预期的效果. GCDIoU 由于增加了包含预测框和真实框的最小矩形框，计算每个预测框之间的欧氏距离，并且综合考虑重叠率和宽高比，让模型加速收敛. 迭代次数为 100 时，位置损失低于 0.03，置信度损失低于 0.01，分类损失低于 0.01. 最终检测的准确率为 98.87%，接近 100%. 由于 Mosaic 数据增强人为制造了大量的小目标，提升了对手势检测的召回率，在迭代次数为 100 时，就已经稳定在 99.98%.



(a) 位置损失　(b) 置信度损失　(c) 分类损失

(d) 准确率　(e) 召回率　(f) mAP 值

图 3　训练集实验结果

Fig. 3　Experimental results of training sets

2.2　IoU 修改提升

对于深度学习的神经网络模型来说，损失函数的设计尤为重要，对比 YOLO-V1 和 YOLOv3 中的定位损失，采用的是平方和的损失计算方法，虽然可以考虑不同尺度对回归损失的影响，但没有考虑到位置坐标的相关性，并不能很好地反映预测框与真实框的重合程度.

通过结合 IoU、GIoU、CIoU、DIoU 定义的 GCDIoU 损失函数，充分考虑了预测框和真实框的重叠面积、中心点距离、宽高比，在训练过程中加快了

位置损失的下降速度，以此提高模型的训练效率. 如图 4 所示，GCDIoU 的损失值相比于其他 4 种基本处于较低水平.

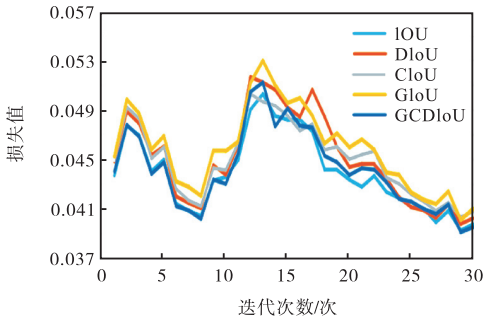


图 4　不同 IoU 损失值下降对比

Fig. 4　Comparison of different IOU loss values

2.3　对比实验

为检测模型改进效果，实验中对比改进前与改进后模型在测试集的 mAP 值对比（图 5），结果表明改进效果较好.

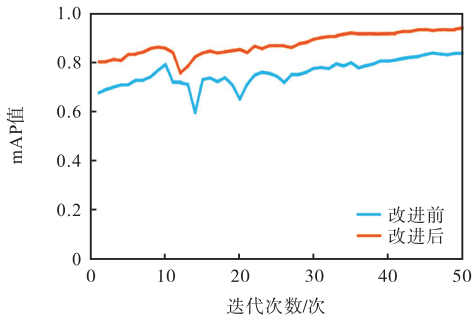


图 5　模型改进前后测试集 mAP 值对比

Fig. 5　Comparison of mAP value of test sets before and after model improvement

3　结　语

本文提出了一种基于改进的 YOLOv3 实现的手势实时识别人机交互方法. 数据集通过对录像按照 10 帧/秒进行切分获得图片，借助 CVAT 标注工具进行标注. 底层特征提取器采用迁移 Darknet-53 模型，通过 K-means 聚类算法对先验框进行优化，训练时通过 Mosaic 数据增强提高手势检测准确率，最后通过训练出的模型结合 pynput 模块实现手势识别控制鼠标达成人机交互. 改进后的人机交互方法在识别准确率、召回率和速度上都取得了很好的效果，可用于基于视觉手势交互场景的实时识别，对视频中的多目标手势进行快速、准确的识别.

人机交互方式的改进可以带来工作效率的提升，本文提出的手势识别与人机交互方法只是针对静态手势进行识别，对于视频中帧与帧之间的时序信息没有加以利用，未来改进的方向是识别和理解连续的、动态的手势.

参考文献：

[1] REDMON J，DIVVALA S，GIRSHICK R B，et al. You only look once：unified，real-time object detection[C]//IEEE. 2016 IEEEE Conference on Computer Vision and Pattern Recognition（CVPR）. New York：IEEE，2016：779–788.

[2] REN S，HE K，GIRSHICK R B，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Transactions on pattern analysis and machine intelligence，2015，39：1137–1149.

[3] LIU W，ANGUELOV D，ERHAN D，et al. SSD：single shot multibox detector[C]//ECCV. Computer Vision-ECCV 2016. Berlin：Springer，2016：21–37.

[4] HE K，GKIOXARI G，PIOTR D，et al. Mask R-CNN[C]//IEEE. 2017 IEEE International Conference on Computer Vision（ICCV）. New York：IEEE，2017：2980–2988.

[5] LIN T，GOYAL P，GIRSHICK R B，et al. Focal loss for dense object detection[C]//IEEE. 2017 IEEE International Conference on Computer Vision（ICCV）. New York：IEEE，2017：2999–3007.

[6] REDMON J，FARHADI A. YOLOv3：an incremental improvement[EB/OL]. [2021–01–25]. https://arxiv.org/abs/1804.02767V1.

[7] 王千，王成，冯振元，等. K-means 聚类算法研究综述[J]. 电子设计工程，2012，20（7）：21–24.

[8] BOCHKOVSKIY A，WANG C，LIAO H. YOLOv4：optimal speed and accuracy of object detection[EB/OL]. [2021–01–25]. https://arxiv.org/abs/2004.10934.

[9] ZHENG Z，WANG P，LIU W，et al. Distance-IoU loss：faster and better learning for bounding box regression[EB/OL]. [2021–01–25]. https://arxiv.org/abs/1911.08287.

[10] SHELHAMER E，LONG J，DARRELL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on pattern analysis and machine intelligence，2017，39：640–651.

[11] YUN S，HAN D，OH S J，et al. Cutmix：regularization sstrategy to train strong classifiers with localizable features[C]//IEEE. 2019 IEEE CVF International Confer-

ence on Computer Vision(ICCV). New York: 2019: 6022–6031.

[12] WU S, LI X, WANG X. IoU-aware single-stage object detector for accurate localization[J]. Image and vision computing, 2020, 97: 103911.

[13] 张勇, 张强, 徐林嘉, 等. 一种基于 YOLOv3 的静态手势实时识别方法: 201811137932.5[P]. 2019–02–12.

[14] ZHANG Q, ZHANG Y, LIU Z, et al. Real-time hand gesture recognition method based on improved YOLOv3[J]. Computer engineering, 2020, 46(3): 237–245.

[15] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//IEEE. 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). New York: IEEE, 2017: 8100173.

[16] 李光华. 基于计算机视觉的手势识别技术研究与应用[D]. 成都: 电子科技大学, 2014.

责任编辑: 郎婧

(上接第 48 页)

统 : 1–8[2021–05–27]. http://kns.cnki.net/kcms/detail/21.1106.TP.20210517.1243.006.html.

[7] 印国成. 基于 K-means 的语义协同过滤推荐算法[J]. 扬州大学学报(自然科学版), 2018, 21(1): 46–49.

[8] KANIMONZHI S. Effective constraint based clustering approach for collaborative filtering recommendation using social network analysis[J]. Bonfring international journal of data mining, 2014, 1(1): 12–17.

[9] TSAIC F, HUNG C. Cluster ensembles in collaborative filtering recommendation[J]. Applied soft computing, 2012, 12(4): 1417–1425.

[10] 李顺勇, 张钰嘉, 张海玉. 基于 NKL 和 K-means 聚类的协同过滤推荐算法[J]. 河南科学, 2020, 38(1): 6–12.

[11] 岳希, 唐聃, 舒红平, 等. 基于数据稀疏性的协同过滤推荐算法改进研究[J]. 工程科学与技术, 2020, 52(1): 198–202.

[12] FENG C J, LIANG J Y, SONG P, et al. A fusion collaborative filtering method for sparse data in recommender system[J]. Information sciences, 2020, 521: 365–379.

[13] KOOHI H, KIANI K. User based collaborative filtering using fuzzy C-means[J]. Measurement, 2016, 91: 134–139.

[14] HERLOCKER J, KONSTAN J A, RIED J. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms[J]. Information retrieval journal, 2002, 5(4): 287–310.

[15] 杨晓君. K-means 聚类算法研究及在股票投资的应用[D]. 重庆: 重庆大学, 2019.

[16] 吴金李, 张建明. 基于二分 K-means 的协同过滤推荐算法[J]. 软件导刊, 2017, 16(1): 26–29.

[17] 武建伟, 俞晓红, 陈文清. 基于密度的动态协同过滤图书推荐算法[J]. 计算机应用研究, 2010, 27(8): 3013–3015.

责任编辑: 郎婧