计算机视觉实验期末报告

E202008000106 邹雪丹
201708010207 杨佳政
201708010214 唐协成
201508010628 李鑫

1、 任务分析

任务需求：从大量的视频数据中自动发掘常见的抓取手势类型集，其在机器人学和医疗康复等领域内有重要的应用教职。这次编程的任务是通过设计计算机视觉算法，对包括手部动作的图像集合进行处理和分析，从中自动发现不同抓取手势类型。

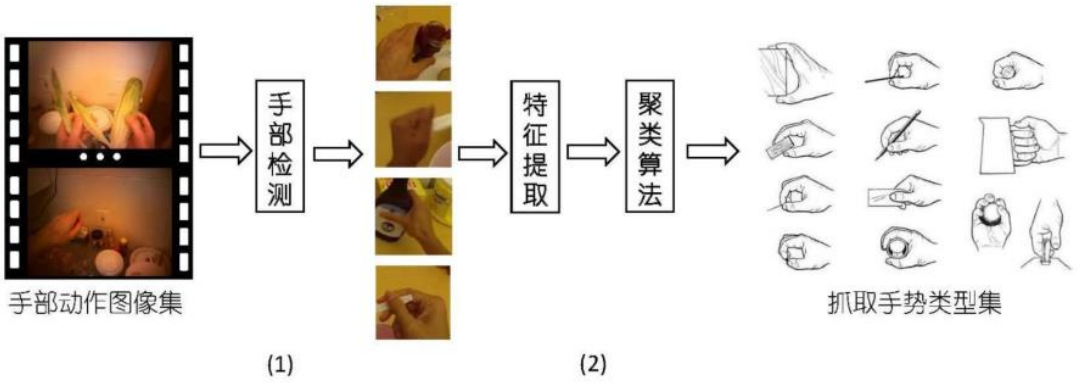
图一：任务的部分图像截图

我们观察给出的图片数据集，总共有 1500 余张。我们发现所有的图片均是在第一人称视角拍摄的（First Person Video），进一步的调查发现这其实是 Gatech 大学发布的 EGTEA Gaze+数据集[1]。这个数据集是依靠第一人称录制工具拍摄的一组厨房煮菜视频，被广泛地应用在第一人称视频相关任务的研究中。

我们进一步地观察给出的图片数据集发现，其中的部分图片手只出现在了边缘角落，在较差的光照条件下导致和背景色差异不明显，是很难纯粹靠颜色来划分的。在另一些图片中还存在着诸如动态模糊，同时出现两只手等问题。我们随机地挑选了我们的数据图片在百度 AI 平台[2]上做了手部测试和关键点标记算法的测试，事实证明即使是已经商业化的百度 AI 也没能很好地从数据集中分判出手部图片。

基于以上结果，我们认为这个问题是一个研究级难度的问题。我们需要识别第一人称摄像头拍摄的各种手部抓取姿势并将它们分类。

我们认为这个问题可以分解为识别手部并分割，以及对分割到的手部图像特征提取并分类或聚类，这两个大的部分。对于每一个部分的解决方法是相对独立的，我们可以单独地提升其中每个步骤的效果。



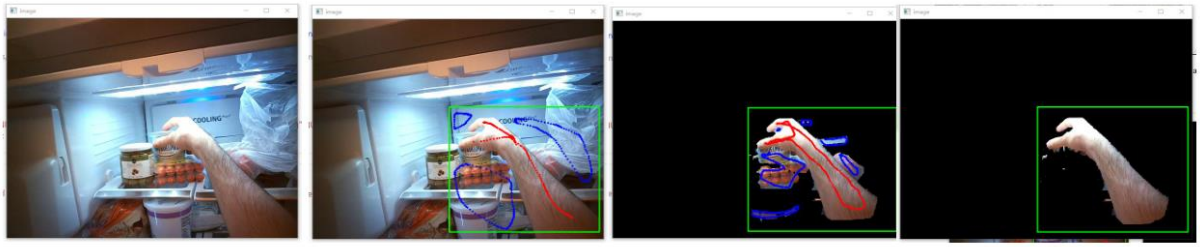
图二：任务的过程划分

2、 手部检测和分割

2.1 方法分析

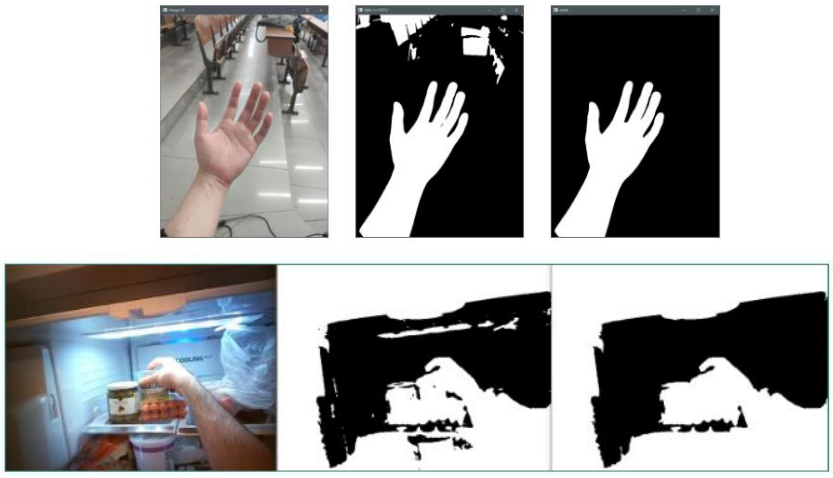
常见的手部检测和分割方法大致包括：（1）Grabcut （2）基于肤色的检测方法 （3）Yolo 神经网络 （4）R/Fast/Faster-CNN 方法 （5）深度图方法 （6）手动设计特征法，再结合一些辅助的图像处理操作如分水岭算法和图像的腐蚀膨胀处理。我们这里结合实际情况对 Grabcut，基于肤色检测和 CNN 类方法做进一步讨论。

我们首先对我们的图像使用了经典的 Grabcut [3]方法进行了测试。如果要精确地划分，Grabcut 方法需要用户提前交互性地选取出图片中的部分前景区和背景区。我们的测试表面大部分图片需要 2 到 3 次的交互才能较好地分判出手部区域。基于此，用这个方法来分割所有的 1500 多张图片是较为繁琐的，但是可以被考虑来作为部分图片打标签的方法。



图三：用 Grabcut 方法交互式地分割一张数据图片的手部

接下来我们测试了常见的基于 Jones 等人提出的肤色检测方法[4]，这种方法的原理是基于手部肤色和其它物件的颜色在 YUV 色彩空间的分布上存在差异，由此常见的操作是将图片转化到 YUV 色彩空间上，然后对其中的 U 分量进行高斯滤波，最后再对结果做二值化阈值 OSTU 法（大津算法）。在做了上述处理后我们大致可以得到包含可能是手部区域的多个图像轮廓，可以考虑使用现在 opencv 自带的 Satoshi 等人提出的 findContour()函数[5]来寻找所有的轮廓，保留较大的轮廓得到结果。我们的测试发现，这种方法对简单的手部明显的图像有一定的效果，但是对于任务数据集中的复杂情况仅仅依靠颜色阈值来分割手部是行不通的。



图四：基于常规的肤色检测的方法的分割结果

我们还注意到了一些其他的传统方法，包括 Mittal Arpit 等人提出的将手部识别分割任务做成三个分类器（手形，语意和皮肤）然后评估整体的得分[6]。以 Li Cheng 等人提出的将肤色检测方法改善成一个分类器的问题[7]。其它方法还包括在先验知识上做光流处理（DTR）,或者构建决策树森林，LSVM 等等。

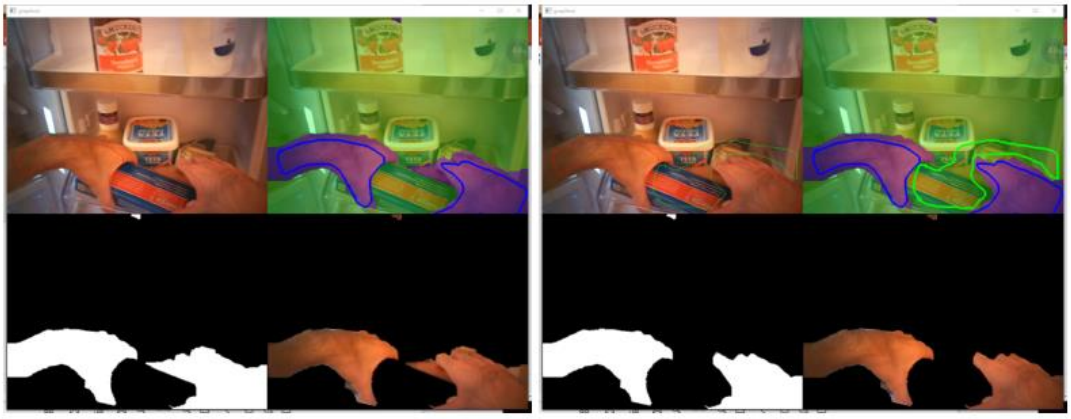
最近的基于深度学习的卷积神经网络 CNN 提出的一些网络模型也被运用在了处理 FPV 的手部识别问题上，这类方法通常是 two-stage 的，即先定位图像中手部的区域，再从区域中提取出手部的图像。诸如 Bambach Sven

等人提出的 CNN 方法首先利用神经网络确定手部的位置区间，再做传统的 Grabcut 分割，依旧是传统的分割方法中的 Two-stage 方法[8]。而 Cai Minjie 等人提出的基于贝叶斯的 CNN 网络则只需要一次性地输入手部图像，不需要任何标签就可以训练得到分割手部图像的网络[9]。

2.2.1 基于肤色检测和推荐系统的方法

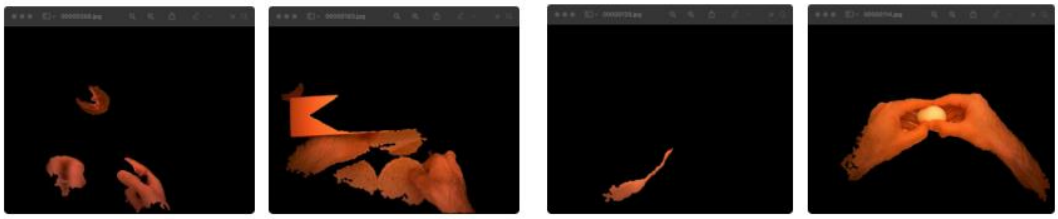
这一部分介绍了我们对 Li Cheng 等人[7]提出的基于肤色检测的推荐系统方法分割手部的具体实践。我们采用了他们在 github 上提供的开源代码，这部分的代码都在 OSX 系统上运行，并搭载了 OpenCV2 环境。

他们的方法构造了一个分类器，为了训练这个分类器需要我们预先给定打了标签的同学作为训练样本。我们按照 10:1 原则对 150 张图片使用他们提供的基于 Grabcut 方法的打标器打上了手部标签。大部分情况下，在 3 次迭代以内，通过标记出前景区和背景区可以较好地标记出图片中手部区域。



图五：使用 Grabcut 方法交互式地划分前景和背景区域来给图片打标

之后，我们使用了打好标记的图片训练他们提出的模型，并最终喂入我们的目标图片来做手部的识别分割。我们修改了源代码中的输出部分使得原先基于视频的分割能够运用在我们的静态图片上。可以发现，这种方法虽然能够在一定程度上分割手部图像，但是也有相当的情况只能分割出无意义的碎片。

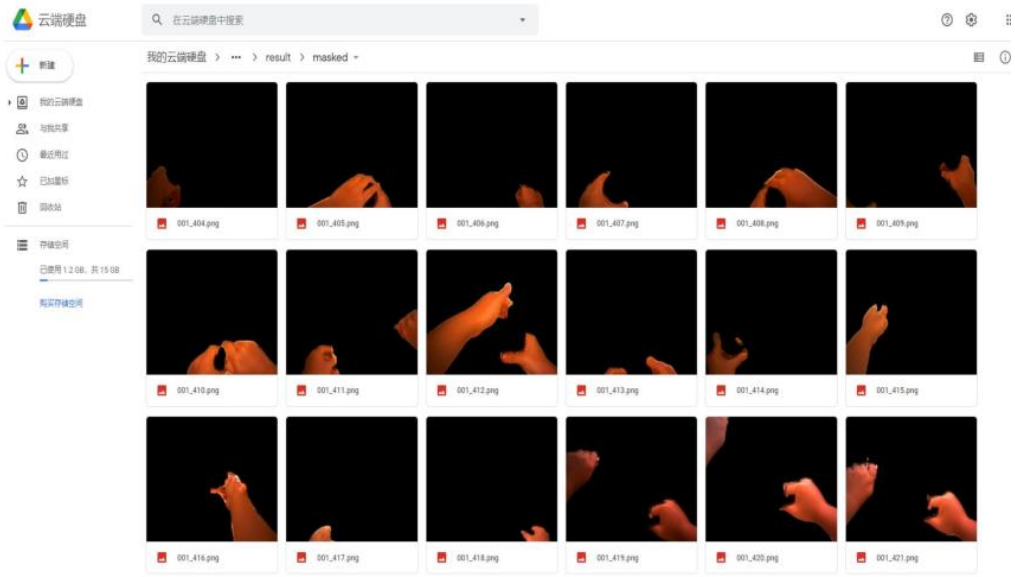


图六：部分分割效果不佳的结果

2.2.2 基于贝叶斯 CNN 的方法

我们接着尝试了 Cai Minjie 等人于 2020 年提出的一种贝叶斯 CNN 神经网络方法。他们的方法只需要输入无标签的第一人称视频图像就能够很好地提取其中的手部图像。

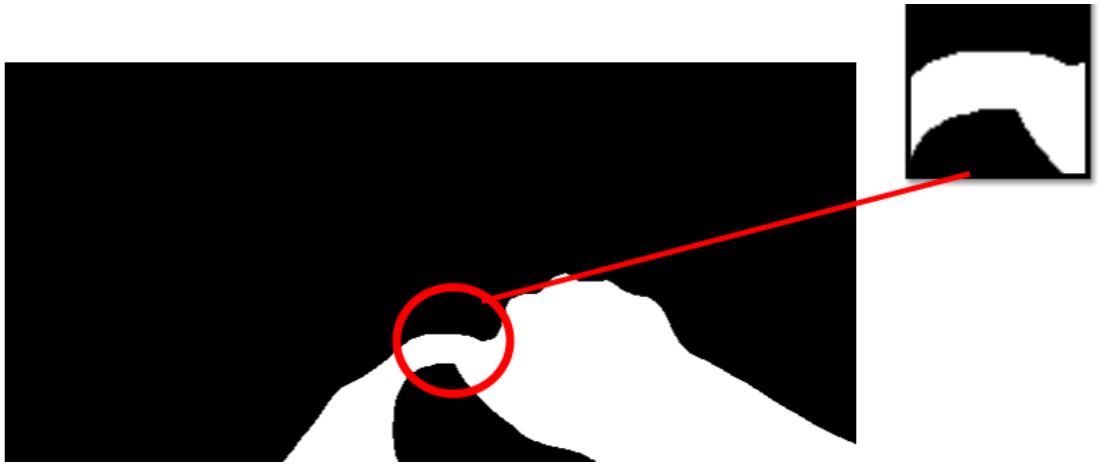
我们在 Google Colab 上，修改了源代码中的训练部分，注意到其预训练模型正是基于 EGTEA Gaze+数据集，我们直接使用了预训练模型，修改后得到了测试代码来讲我们的数据集丢入网络得到分割结果。可以看到，这个方法的识别风格结果比 Li,Cheng 等人的方法结果好上不少。



图七：使用深度学习方法分割得到的部分结果

接下来，我们对分割的结果进行必要的处理以尽可能地去除掉错误划分出来的像素块并分离左右手。我们使用了先前提到的基于 Satoshi 等人提出的[5]OpenCV 自带的 findContours()函数来提取到图像所有的轮廓。我们设置了一个阈值以去除残余的错误轮廓，然后选择最多两个最大轮廓，由此得到所有的手部区域的单独划分。我们最终得到了将近 3000 张单独的手部图片。

不过，我们提取的手部轮廓有时也存在一定的问题。在某些特殊情况下，由于左右手接触过于紧密，靠贝叶斯 CNN 网络也无法将左右手单独提取出来。我们考虑过使用腐蚀的方法先尽可能让手部的轮廓分离再进行轮廓提取，但是我们后续发现对于很多手部信息本来就不是特别清晰的图片数据，使用腐蚀方法只会让图片完全丢失手的手指部分关键信息，权衡之下我们放弃了这种处理，并保留了双手粘合的情况。



图八：双手紧贴导致左右手无法被很好地单独分割

3、 手部聚类和分类

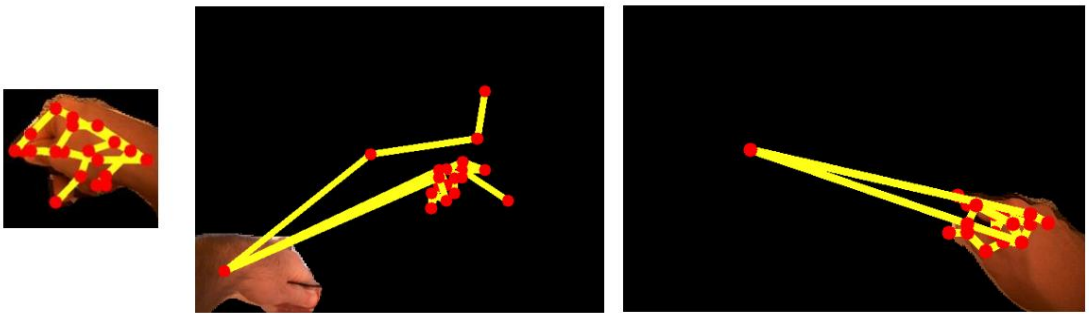
3.1 方法分析

对于手部的特征提取和聚类，或者手部分类问题，常见的手部特征提取方法包括基于深度学习，使用各种 CNN 的方法，以及利用手动设计特征诸如 Gabor, HOG, SIFT, DTR，模板匹配，统计分析等方法。近年，T Simon 等人提出的一种通过标记少量多视角下手部图像作为训练集训练的 CPM 网络可以对单张二维手部图像进行关键点标记，其方法将手部分成了累计 21 个关键点[10]。可以考虑利用这些关键点，再结合经典的基于划分的聚类方法 K-means 算法来对手部姿势进行聚类。

相对于聚类，为了更好地获取不同类别的具体意义，可以考虑采用分类的方法。而对于手部姿势的分类问题，可以根据 T Feix 等人提出的一种抓取姿势分类表[11]来对手势进行分类。Cai Minjie 等人基于这种分类表，结合计算手部的 HOG，手部交互物体的 SIFT 特征，训练了一个一对多多类别分类器来分析抓取手势的类别[12]。同样，具有自然空间顺序的图像分类非常适合于 CNN，其在学习数据充足时对图像的分类问题表现稳定。考虑到手部的动作有较强的一致性特征，也可以使用传统的 CNN 网络来对手部图像分类。

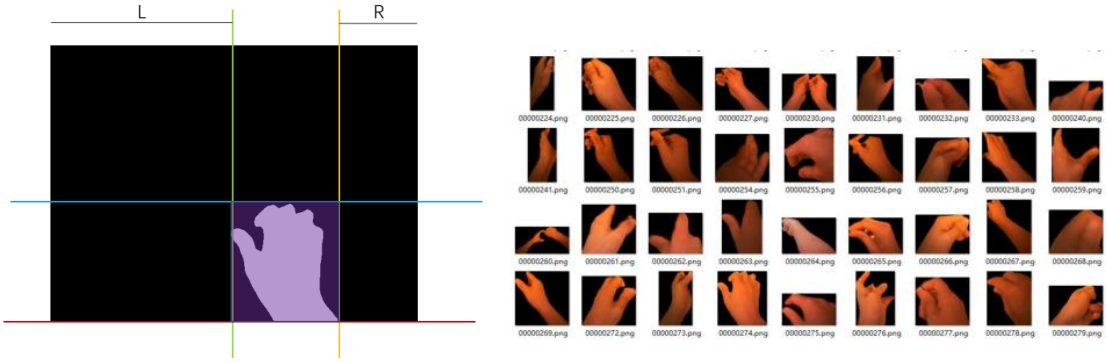
3.2.1 基于特征提取再聚类的方法

我们在 OSX 搭载 OpenCV3 环境尝试利用 T Simon 等人提出的 CPM 网络[10]对手部进行关键点提取。不过我们很快发现这种方法对于没有单独提取手留有大量黑色区域的图片效果很差。为了提升方法的正确率，我们不得不对分割结果的手部区域进行提取。



图九：手部关键点提取与错误的案例

提取最小手部矩形的方法很简单，只需要遍历整张图，找到非黑色的 mask 区域的最左，最右，最上和最小坐标。用最左上，最右下坐标组合得到的 rect 矩形区域就是最小的手部矩形区域。类似地，注意到左手大部分情况在图像的左侧，而右手在大部分图像的右侧，用 rect 矩形的左边缘和右边缘坐标还可以大致判断是左手还是右手。实际上，由于数据集的复杂性，这种简单的特征分辨左右手正确率并不算特别高。



图十：手部最小全包围矩形区域的提取

之后我们再利用 T. Simon 等人的方法[10]作用于我们提取后的手部局部图片上，成功得到了每张图的手部特征点坐标。为了方便后续聚类操作不再花费额外的时间重新提取每张图的特征点，我们让每张图生成的手部特征点坐标同时保存在对应的 txt 文本上并记录图片编号。

对于最后的聚类部分，我们采用了 K-means 聚类，直接将所有点的横纵坐标求均值，得到的几何中心点做聚类，即对 3000 余个 2 维向量做 K-means 聚类。我们规定了初始产生 10 个随机的聚类中心，由于聚类中心的随机性，存在有一类没有任何图片的情况。我们反复执行了多次 K-means 直到得到了每个类别都有一定聚类图片的结果。

事实证明我们直接计算每个手部关键点几何中心并聚类的方法效果不好。原因在于我们在计算几何中心的时候丢失了每个关键点的潜在信息。实际上应该考虑将每个手 21 个关键点的坐标展开成 42 维的向量进行聚类，这样聚

类结果才可以充分地利用每个关键点的信息，提升最后的聚类效果。

图十一：部分手部抓取姿势的聚类结果

3.2.2 基于 CNN 分类的方法

为了得到更加具有语义的分类结果，我们考虑采用经典的 CNN 方法进行分类。我们在 Google Colab 上实现了我们的 CNN 分类部分。我们参考 T. Feix 于 2009 年提出的手抓取姿势分类表[11]并做了简化，将我们的手势分为左右手各 9 种类别，加上双手的情况和其他情况共计 20 种分类。

我们按照 10:1 原则对近 300 张图片做了人工打标分类，为了让我们的 CNN 具有更好的泛化能力，我们同时对原始的训练样本做了左旋，右旋以及镜像图片的左旋，右旋这种图像增强操作。最终我们丢入我们的待分类图片到训练好的 CNN 网络中得到最终的分类结果。

图十二：按照文件夹排列的最终分类结果

4、 总结

对于给定的较为困难的第一人称视频录像的图像，我们通过调研文献，在网上寻找开源代码，手动搭建运行环境，分别围绕传统方法和深度学习方法两个角度完成了我们的手部提取分割和分类聚类任务。可以很明显地感受到，在手部图像的识别分割部分，采用基于深度学习的方法最终取得的效果明显优于传统机器学习方法训练得到的分割器。这里的深层次原因还有待探讨，不过这也能一定程度上反映出深度学习虽然在目前还缺少足够的数学理论解释，其在图像处理任务上表现出来的优越性能是显而易见的。

5、 参考文献

[1] http://cbs.ic.gatech.edu/fpv/#egtea_gaze_plus
[2] https://ai.baidu.com/tech/body/hand
[3] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. In ACM SIGGRAPH 2004 Papers (SIGGRAPH '04). Association for Computing Machinery, New York, NY, USA, 309– 314.
[4] Jones, Michael J., and James M. Rehg. "Statistical color models with application to skin detection." International Journal of Computer Vision 46.1 (2002): 81-96.
[5] Satoshi Suzuki and others. Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing, 30(1):32– 46, 1985.
[6] Mittal, Arpit , A. Zisserman , and P. Torr . "Hand detection using multiple proposals." British Machine Vision Conference 2011
[7] Li, Cheng, and Kris M. Kitani. "Pixel-level hand detection in ego-centric videos." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013
[8] Bambach, Sven, et al. "Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions." Proceedings of the IEEE International Conference on Computer Vision. 2015.
[9] Cai, Minjie, Feng Lu, and Yoichi Sato. "Generalizing Hand Segmentation in Egocentric Videos With Uncertainty-Guided Model Adaptation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020
[10] T. Simon, H. Joo, I. Matthews and Y. Sheikh, "Hand Keypoint Detection in Single Images Using Multiview Bootstrapping," 2017 IEEE Conference on Computer Vision and Pattern Recognition.
[11] T. Feix, R. Pawlik, H.-B. Schmiedmayer, J. Romero, and D. Kragic, “A comprehensive grasp taxonomy,” in Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation, 2009, pp. 2– 3.
[12] Cai, Minjie , K. M. Kitani , and Y. Sato . "A scalable approach for understanding the visual structures of hand grasps." Proceedings IEEE International Conference on Robotics & Automation 2015(2015):1360-1366.