室内场景语义分割方法研究

徐晓刚，龚小谨

（浙江大学，信息与电子工程学院，浙江，杭州，310027）

Abstract: 近年来随着人工智能技术的飞速发展，在*RGB-D*场景下的语义分割技术取得了巨大的进步．本次*SRTP*科研训练的主要研究课题在于室内场景下的语义分割技术的探索。基于整个课题，提出的主要框架分为两个部分：1. 基于*Kernel* 描述符和马尔科夫随机场的框架(*MRF*)，以下简称*RGB-D*框架；2. 基于全卷积神经网络(*FCN*)的框架，以下简称*FCN*框架。*RGB-D*框架主要利用*Kernel*描述符能够有效地提取图像特征的特点，结合处理上下文语义的马尔科夫随机场和分割树模型，完成室内场景的像素级别场景分割并且标注。该框架区别与传统的计算机视觉方法之处在于分别对轮廓检测器产生的分割树的特征和*Kernel*描述符产生的超像素分别使用*SVM*算法进行标注和分割。该框架在*NYU-V2*数据集上进行测试，实验最后表明取得71%的准确率。*FCN*框架主要使用目前流行的深度学习框架*Caffe*，将传统的卷积神经网路的最后的全连接层改为反卷积层，设计一种*end-to-end*的图像语义分割模型。其中主要区别于*CNN*的关键在于将全连接层改为卷积层的方法。最后的实验依旧在*NYU-V2*的数据集上进行，实验最后表明在像素级别取得了61%的准确率。

Key words: 场景语义分割；RGB-D；马尔科夫随机场；FCN

Research for Indoor Scene Semantic Segmentation Methods

Xiaogang Xu, Xiaojin Gong

(*Zhejiang University, College of Information Science and Electronic Engineering, Zhejiang, Hangzhou, 310027*)

Abstract: With the development of artificial intelligence, great improvement in the field of image semantic segmentation under the situation of RGB-D. The main topic for this SRTP focus on the research of the technology of semantic segmentation under the circumstances of indoor. We propose two frameworks based on this subject: 1. The framework based on Kernel descriptor and Markov random field(MRF), we called this RGB-D Metric; 2. The framework based on fully convolution neural network(FCN), we called this FCN Metric. The RGB-D Metric use the characteristics of Kernel descriptor that this can help to extract the image feature effectively。 The context semantic model is implemented by the models of MRF and segmentation tree. We achieve the goal of object labeling indoor in the level of pixels. The main difference between this framework and traditional computer vision's methods is that this metric utilize machine learning methods which is called Support vector machine (SVM) to label the objects. The features used for classification is obtained by segmentation tree and the super pixels produced by Kernel descriptor. The results of our experiments on the database of NYU-v2 indicating that the accuracy of this metrics can reach 71%. The FCN Metric is implemented on the platform of Caffe which is a popular stage for deep learning. FCN models change the full-connection layers of CNN into convolution layers. An end-to-end framework for image semantic segmentation is realized by this metric. The main difference between CNN and FCN is that FCN turn the full-connection layers into convolution layers. The experiments for this metric is also implemented on the database of NYU-v2. The results illustrate that the accuracy of this metrics can reach 61%.

Key words: scene semantic segmentation; RGB-D; Markov random field; FCN

1 引言

随着计算机视觉技术的发展，如何自动从一张图片中进行物体分类已经得到长久的研究。但是场景理解一直都还是图像处理与计算机视觉领域研究的热点。任务的主要目的是给室内场景每个像素点提供一个预定义的语义类别标签。根据不同的标注基元量化级别，相应的方法大致可以分为区域级别的语义标注方法和像素级别的语义标注法。前者的标注效率较高而且整体的视觉效果较好，但是后者的标注层次化细节较高，但是标注效率相对较低。在这里我们主要研究的是像素级别的语义分割与标注。

文献[1] 基于高斯核线性组合成对项势能的快速稠密全连通 *CRFs* 概率图模型，提出了 *RGB* 图像像素级别的语义标注推断方法。而近几年来由于深度传感器的进步，使其具有强大的捕获场景结构的能力，学者们倾向于将场景集合深度融入语义标注。文献[2]提出了解析 *RGB-D* 室内场景中区域界别的主要平面和物体的框架，并且基于此推断物体支撑关系。文献[3]使用了一种反馈式的前向神经网络作为判别分类器，从 *RGB* 图像，深度图像以及经过旋转处理后的 *RGB* 图像中提取尺度不变特征转换 *SIFT* 特征描述。而深度学习的兴起，又为图像语义的分割提供更加强大的框架，主要是利用卷积神经网络的提取特征的能力来提升算法的性能。卷积神经网络在经过多年的发展之后，表现出简单高效的特征提取能力。文献[4]提出了基于多尺度 *RGB* 图像和深度图像的卷积神经网络实现 *RGB-D* 场景语义标注。深度卷积神经网络(*DCNN*)的出现更加提高了标注的准确性，物体检测任务中加州大学伯克利分校的 *Ross Girshick* 等人提出的 *R-CNN*，和随后微软亚洲研究院的 *Fast R-CNN*、*Faster R-CNN* 在 *PASCAL VOC* 数据集上， 准确率和识别速度上不断提高。文献[5]提出了一种利用 *DCNN* 来做像素级别标注的框架，称之为全卷积神经网络(*FCN*)。*FCN* 对图像进行像素级别的分类，从而解决了语义级别的图像分割问题。与经典的 *CNN* 在卷积层之后使用全连接层来得到固定长度的特征向量进行分类（全联接层＋*SOFTMAX* 输出）不同，*FCN* 可以接受任意尺寸的输入图像，采用反卷积层对最后一个卷积层的 *feature map* 进行上采样， 使它恢复到输入图像相同的尺寸，从而可以对每个像素都产生了一个预测， 同时保留了原始输入图像中的空间信息， 最后在上采样的特征图上进行逐像素分类。

本课题存在如下的难点：

1）室内场景由于其物品种类多，物品之间的颜色信息接近，难于进行分割，再加上对比度信息小，遮挡现象频繁等原因，使得对于室内场景的语义分割有较高的难度。

2）此项目相比较与以往工作的创新点即在于它的识别中存在大量物品以及需要我们将彩色
图像和深度图像结合进行识别。

3）场景语义标注方案中存在难以合适选择标注基元尺度，标注方案难以充分挖掘几何深度信息对上下文推理过程的贡献的问题。

本文提出基于 *RGB-D* 和马尔科夫随机场(*MRF*)的 *RGB-D* 场景语义理解框架，和基于卷积神经网络的 *FCN* 图像像素级别分割的框架。与传统方法相比，所提出的语义分割方法无论在主观上还是客观评估上，均具有较好的效果。

2 相关工作

图像场景理解作为计算机视觉研究中的一项必要工作，自 20 世纪 60 年代以来，一直都是相关领域研究的热点和难点。笼统地讲，图像场景理解就是一个对任意输入图像，通过一系列对视觉信息和知识信息的处理与分析，最终得到图像语义解释的过程；是一门以图像为对象，知识为核心，研究图像中何位置有何目标、图像中目标之间的相互关系。目前国内外许多著名的计算机视觉研究社区已提出诸多关于通用图像语义标注方法的解决方案。而近些年来，越来越多的计算机视觉研究社区开始逐步给予室内场景语义标注问题以更多关注。导致上述现象产生的部分原因很可能是由于*NYU Depth* 系列 *RGB-D* 数据集的构建。*NYU Depth* 系列 *RGB-D* 数据集是国际上第一个针对室内场景语义标注问题单独构建。我们的课题验证主要是受益于这一公开的 *RGB-D* 数据集。作为利用深度信息进行室内场景语义标注的先驱，已经有许多有效的算法被提出，比如随机场概率图模型优化框架等。而分类的机器学习算法在近些年来也表现出其强大的生命力，一般常用的包括支持向量机(*SVM*)，决策树等。

3 算法概述

1) *RGB-D* 框架

室内场景下的语义分割与标注在目前还是一个具有挑战性的任务，一般的流程在于首先提取有效地 *RGB-D* 的特征，对其进行编码使得我们在之后可以使用相对应的机器学习分类算法来对整幅图像的不同像素进行分类。算法的流程概述如下：

（1）运用 *Kernel descriptors*(*KDES*)来描述图像中的局部特征,例如重力，颜色，深度信息以及平面的表面法向等。选择这种方法是因为它已经表现出良好的精确度。而在初步的特征提取之后，进一步使用高效匹配核函数（*EMK*）对特征进行进一步处理。我们利用 *EMK* 将图像的局部特征映射到低维度的特征空间中，对同一张图中的不同局部特征进行匹配，将具有高匹配度的部分块设为一个超像素。超像素的获得使得我们在进行分类的时候的算法复杂度和效果更好。

（2）在此之后，我们需要结合 *GPB* 检测器算法，分别在 *RGB* 图像和深度图中运行。以线性加权的方法将它们相加结合使用。由此我们可以获得图像中边界轮廓的检测。之后通过设定多个阈值，每一个阈值均能获得轮廓图。而接下来超度量轮廓图（*UCM*）可以在边界图，以及一组等级镶嵌的区域之间建立对偶关系，产生一棵嵌套的分割树。我们可以从中选取一个效果最好的树，舍弃有过分割等现象的。之后可以在一颗分割树的每一层中对超像素进行 *SVM* 多分类算法，得到一个分割结果。

（3）在对每一张图像获得一棵分割树之后，在上下文语义处理中，我们首先对叶子层的每一个超像素来构造一个 *feature*，这个 *feature* 是用到从叶子一直回溯到根的路径上的节点的特征集合。并且在这一层上应用 *SVM* 多分类进行标注其余层也可以进行同样的方法获得每一个超像素的特征集合。另一个上下文语义的方法是用 *MRF* 的方法，主要能够对边缘的信息做更好地修正，所以获得效果也更好。

2) *FCN* 框架

FCN 的方法区别于传统的 *CNN* 网络，与传统用 *CNN* 进行图像分割的方法相比，*FCN* 将传统 *CNN* 中的全连接层转化成一个个的卷积层。如图 1 所示，在传统的 *CNN* 结构中，前 5 层是卷积层，第 6 层和第 7 层分别是一个长度为 4096 的一维向量，第 8 层是长度为 1000 的一维向量，分别对应 1000 个类别的概率。FCN 将这 3 层表示为卷积层，卷积核的大小(通道数，宽，高)分别为(4096, 1, 1)、(4096, 1, 1)、(1000, 1, 1)。所有的层都是卷积层，故称为全卷积网络。

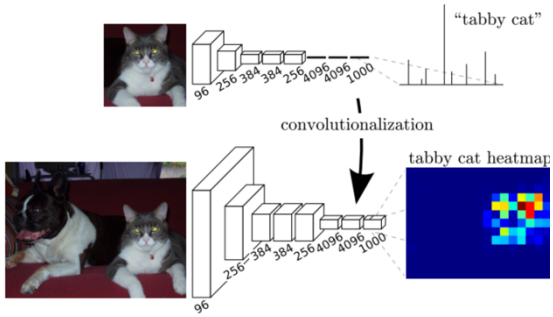


图 1

FCN 网络结构有两大明显的优点：一是可以接受任意大小的输入图像，而不用要求所有的训练图像和测试图像具有同样的尺寸。二是更加高效，因为避免了由于使用像素块而带来的重复存储和计算卷积的问题。

4 RGB-D 框架详述

1) *Kernel* 描述符与有效匹配 *Kernels*(EMK)

在进行语义分析的时候，一个关键的点在于如何提取图像中能够反映像素级别的特征。*Kernel* 是一个针对于局部特征描述的一致框架：它对于任意的像素级别的相似函数，都能够将其转换成一个 *patch* 上的描述符。在这里我们使用了六个不同方面的 *kernel* 描述符，包括：梯度，颜色，局部二元特征，深度图梯度，旋转/曲面法线以及自相似性。在这里为说明问题，我们简单描述梯度描述符的概念。我们首先将图像转换成灰度图像，并且在每一个像素点计算其灰度的梯度值，而我们的梯度描述符可以从梯度相似函 k_0 来进行构造：

$$F_{grad}^{t}(Z)=\sum_{i=1}^{d_o}\sum_{j=1}^{d_n}\alpha_{ij}^{t}\{\sum_{z\in Z}\tilde{m}_z k_o(\tilde{\theta}_z,p_i)k_s(z,q_j)\}$$

其中 F_{grad} 我们构造的梯度描述符，z 是归一化之后的在深度 *patch* 中的相对位置。$\tilde{\theta}_z$ 和 $\tilde{m}_z$ 是在像素 z 上的归一化之后的深度梯度的方向和幅度。在这里表示方向的 *Kernel*：

$$k_o(\tilde{\theta}_z,\tilde{\theta}_x)=\exp(-\gamma_o\|\tilde{\theta}_z-\tilde{\theta}_x\|^2)$$

用来计算梯度方向之间的相似性。而位置 *Kernel*：

$$k_s(z,x)=\exp(-\gamma_s\|z-x\|^2)$$

主要是用来衡量两个像素在空间上的接近程度。

$\{p_i\}_{i=1}^{d_o},\{q_j\}_{j=1}^{d_s}$ 是从支持区域中一致采样出来的，其他的 *Kernel* 描述符也是通过相类似的像素级别的相似函数计算出来的。

除了 *KDES* 提供的特征之外，我们也提供了典型的先验特征，主要是关于位置，面积，周长等。对于 *RGB-D*，我们另外添加了相对深度的信息。至此我们已经从一个图像方格中提取出了我们需要的 Kernel 描述符。我们还需要使用有效匹配 *Kernel*(*EMK*)来将描述符进行转化。*EMK* 的主要作用在于将图像的局部特征映射到低维度的特征空间中，使得我们可以对同一张图中的不同局部特征进行匹配。这一步是我们用来获得超像素的特别步骤。

我们用 $\phi(s)$ 来表示我们已经获得的特征，在这里是 *KDES* 的特征加上先验特征。对于分割树中的每一层，我们都使用 $1-vs-all$ 的

SVM 训练方法，也就是说对于每一个在高度为 t 上面的语义类别 $c\in\{1,...,C\}$，我们都可以得到一个线性的分数函数：

$$f_{t,c}(s)=w_{t,c}^{T}\phi(s)+b_{t,c}$$

在这里我们还需要考虑的是如何来为每一个数据实例来选择一个权重。因为这些超像素都是具有不同的面积的，我们现在假设 c 是类别 s 的实际类别 A_s 在这里指的是 s 的面积，Q_s 在这里代表的则是所有类别 c 中的超像素，我们可以给 f(s)分配权重为：

$$A_s/(\sum_{q\in Q_c}A_q)^p$$

2) 上下文模型

在确定了整个的特征提取以及为此进行了 *SVM* 的初步分类之后，得到的效果其实不能够满足我们的要求，特别是我们没有考虑能够应对语义的上下文模型。针对上下文模型，我们所需要做的主要还是在于实现分类的目标。在这里我们需要进行两个级别的 *SVM*。首先我们需要对提取出来的特征进行超像素级别的 *SVM* 分类。之后我们需要充分利用超像素级别的 *MRF*, 以及分割树中的路径形成的特征。*MRF* 在这里的使用条件在于我们使用了线性的超像素级别的 *SVM* 和 *GBP* 轮廓检测器之后。我们经过实验之后发现我们结合他们一起使用可以获得一个更加好的性能。

1. 在分割树中使用路径分类策略

我们已经使用 *GBP* 轮廓检测器获得了一棵有固定高度的分割树。我们最直接的学习方法就是将每一层用 *SVM* 计算之后的输出级联起来，针对每一个超像素 *s*，我们都产生一个树的特征：

$$Tree(s)=\{f_{t,c}(s_t)\},\forall t,c$$

在这里 *s* 指的是底层的超像素值。$\{s_t\},t\in\{1,...,T\}$ 在这里指的是 s 的祖先。在这里我们还是可以针对树的特征，结合实际的标签来进行训练。我们发现针对此的训练有较好的效果。如果我们仅仅将特征在每一条路径上累加，那么虽然分割的精确度可以增加，但是却只是只有很少的分割。

2. 结合 *GPB* 的超像素 *MRF*

我们提出的第二个协助的框架就是马尔科夫随机场。*MRF* 模型提供了不确定性描述与先验知识联系的纽带, 并利用观测图像, 根据统计决策和估计理论中的最优准则确定分割问题的目标函数, 求解满足这些条件或消费函数的最大可能分布, 从而将分割问题转化为最优化问题。其具有以下几个鲜明的特点:*MRF* 模型可以将像素的空间关系紧密地结合在一起, 将像素间的相互作用加以传播, 因而在图像分割中可以用低阶的 *MRF* 来描述像素间的作用关系;*MRF* 模型既能反映图像的随机性，又能反映图像的潜在结构，这样可以有效地描述图像的性质;*MRF* 模型即从物理模型出发，又与图像数据(灰度值或特征)拟合直接联系起来。

在这里我们使用图论中割的方法去寻找标注，它的主要目标函数是能够最小化 *MRF* 的能量：

$$E(y_1,...,y_{|S|})=\sum_{s\in S}D_s(y_s)+\sum_{\{s,r\}\in N}V_{s,r}(y_s,y_r)$$

在这里 y_s 是超像素 *s* 的标签。N 是所有邻接的对。对于每一个数据项 D_s, 我们都使用每一层的经过面积加权之后的 *SVM* 输出，$-f_{t,c}$ ，而对于每一对的数据项 $V_{s,r}$，我们都可以使用：

$$V_{s,r}=\beta\exp(-\gamma\cdot gPb\_rgbd(s,r))$$

因为 *RGB-D* 的 *GPB* 检测器能够将群体的特征证据融合到一个简单的值，*MRF* 在这里只有两个参数，也十分容易进行交叉验证。我们发现将超像素级别的 *MRF* 和树路径特征结合起来，能够有更好的效果。

5 FCN 框架详述

1）传统卷积神经网络 *CNN*

卷积神经网络(*Convolutional Neural Network, CNN*)是深度学习技术中极具代表的网络结构之一，在图像处理领域取得了很大的成功，在国际标准的 *ImageNet* 数据集上，许多成功的模型都是基于 CNN 的。*CNN* 相较于传统的图像处理算法的优点之一在于，避免了对图像复杂的前期预处理过程（提取人工特征等），可以直接输入原始图像。图像处理中，往往会将图像看成是一个或多个的二维向量，如现在训练比较常用的 *MNIST* 手写体图片就可以看做是一个 28 × 28 的二维向量（黑白图片，只有一个颜色通道；如果是 *RGB* 表示的彩色图片则有三个颜色通道，可表示为三张二维向量）。传统的神经网络都是采用全连接的方式，即输入层到隐藏层的神经元都是全部连接的，这样做将导致参数量巨大，使得网络训练耗时甚至难以训练，而 *CNN* 则通过局部连接、权值共享等方法避免这一困难。

图像最后用来做分类的原理一般在于 *SOFTMAX* 的使用，能够在最后使用全连接层将其作为 *SOFTMAX* 层的输入，之后再进行分类输出，就可以得到最后的分类结果。由于之前已经用卷积操作等对图像的特征进行提取，在最后的输出一般具有较高的准确度。但是一般来说，输入的图片的大小需要固定，而且最后只能够对区域级别的物体进行分类识别，如果要进行像素级别的识别，那么将要耗费许多的训练时间。

2）全卷积神经网络 *FCN*

FCN 对图像进行像素级的分类，从而解决了语义级别的图像分割（*semantic segmention*）问题。与经典的 *CNN* 在卷积层之后使用全连接层可以得到固定长度的特征向量进行分类不同，*FCN* 可以接受任意尺寸的输入图像，采用反卷积层对最后一个卷积层的 *feature map* 进行上采样，使它恢复到输入图像相同的尺寸，从而可以对每个像素都产生了一个预测，同时保留了原始输入图像中的空间信息，最后在上采样的特征图上进行逐像素分类。最后逐个像素计算 *SOFTMAX* 分类的损失，相当于每一个像素对应一个训练样本。简单的来说，*FCN* 与 *CNN* 的区域在把于 *CNN* 最后的全连接层换成卷积层，输出的是一张已经 *Label* 好的图片。

其中主要的一个特点就在于全连接层变为卷积层。全连接层和卷积层之间唯一的不同就是卷积层中的神经元只与输入数据中的一个局部区域连接，并且在卷积列中的神经元共享参数。然而在两类层中，神经元都是计算点积，所以它们的函数形式是一样的。因此，将此两者相互转化是可能的，因为如下的原因：对于任一个卷积层，都存在一个能实现和它一样的前向传播函数的全连接层。权重矩阵是一个巨大的矩阵，除了某些特定块，其余部分都是零。而在其中大部分块中，元素都是相等的。相反，任何全连接层都可以被转化为卷积层。比如，一个 $K=4096$ 的全连接层，输入数据体的尺寸是 $7\times7\times512$，这个全连接层可以被等效地看 $F=7,P=0,S=1,K=4096$ 的卷积层。换句话说，就是将滤波器的尺寸设置为和输入数据体的尺寸一致了。因为只有一个单独的深度列覆盖并滑过输入数据体，所以输出将变成 $1\times1\times4096$，这个结果就和使用初始的那个全连接层一样了。其中更加重要的步骤在于上采样的步骤也就是如何使得输出图像能够达到足够大的尺寸。

在这里我们要注意的是 *FCN* 的缺点：首先是我们得到的结果还是不够精细。进行 8 倍上采样虽然比 32 倍的效果好了很多，但是上采样的结果还是比较模糊和平滑，对图像中的细节不敏感。而且整个框架是对各个像素进行分类，没有充分考虑像素与像素之间的关系。忽略了在通常的基于像素分类的分割方法中使用的空间规整步骤，缺乏空间一致性。

3） 基于区域的全卷积网络 *R-FCN*

FCN 虽然在像素级别分割的情况下已经取得很好的效果，但是目前还存在些许问题，主要是存在这么一个事实：分类需要特征具有平移不变性，检测则要求对目标的平移做出准确响应。现在的大部分 *CNN* 在分类上可以做的很好，但用在检测上效果不佳。*SPP*，*Faster R-CNN* 类的方法在 *ROI pooling* 前都是卷积，是具备平移不变性的，但一旦插入 *ROI pooling* 之后，后面的网络结构就不再具备平移不变性了。因此，*R-FCN*[10] 想提出来的 *position sensitive score map* 这个概念是能把目标的位置信息融合进 *ROI pooling*。

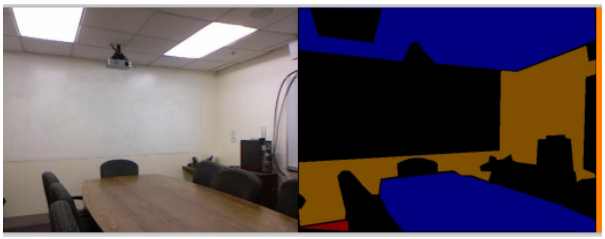
为消除上述存在的问题，在网络的卷积层间插入 *ROI* 池化层。这种具体到区域的操作在不同区域间跑时不再有平移不变性。然而，该设计因引入相当数目的按区域操作层（*region-wise layers*）而牺牲了训练和测试效率。 *R-FCN* 采用全卷积网络结构作为 *FCN*，为给 *FCN* 引入平移变化，用专门的卷积层构建位置敏感分数地图（*position-sensitive score maps*）。每个空间敏感地图编码感兴趣区域的相对空间位置信息。 在 *FCN* 上面增加 1 个位置敏感 *ROI* 池化层来监管这些分数地图。

区域建议网络 *RPN* 给出感兴趣区域，*R-FCN* 对该感兴趣区域分类。*R-FCN* 在与 *RPN* 共享的卷积层后多加 1 个卷积层。所以，*R-FCN* 与 *RPN* 一样，输入为整幅图像。但 *R-FCN* 最后 1 个卷积层的输出从整幅图像的卷积响应图像中分割出感兴趣区域的卷积响应图像。*R-FCN* 最后 1 个卷积层在整幅图像上为每类生成 k^2 个位置敏感分数图，有 C 类物体外加 1 个背景，因此有 $k^2(C+1)$ 个通道的输出层。k^2 个分数图对应描述位置的空间网格。比如，$k\times k=3\times 3$，则 9 个分数图编码单个物体类 $\{top-left, top-center, top-right, ..., bottom-right\}$。*R-FCN* 最后用位置敏感 *ROI* 池化层，给每个 *ROI* 1 个分数。

6 实验结果与分析

1）*RGB-D* 框架的实验结果

对于 *RGB-D* 框架算法， 我们在公开的数据集 *NYU-V2* 上进行。*NYU-V2* 数据集包含典型的室内场景图片，并且在室内场景图片中有较多的物品遮挡的现象，以及存在不同物品之间颜色差异小的特点，也是我们这个任务的最大难点之一。在最后的验证阶段，对于这个数据集，我们 60%的数据来做训练，并且使用 40%的数据来做验证。针对每一张图片，我们都对其做 8 分类的设置，最后的结果的有效性在于我们验证集上的每一张图片中的超像素的分割点的正确的标注值与我们预测的标注类别之间的差别的计算来获得准确率。 部分结果如下：



最后我们发现使用 *RGB-D* 框架的结果的准确率在 71%左右。

可以看到我们最后的框架的效果并不是很高，我们分析了其中可能存在的一些问题：

1> 效果需要继续提高。在室内的场景中，由于不同物品之间靠的很近，而且不容易分割开来，特别是一些颜色相近的物体会被识别为一体，最后会出现识别精度的下降。

2> 我们的 *Kernel descriptors* 的描述特征虽然被证明是高效的，而且有 *EMK* 算法进行超像素产生，但是可能我们对于如何提取特征，以及如何优化 *GPB* 检测器进行边缘检测，还需要继续思考。

3> 在每一层的分割以及语义标注问题上，我们现在的 *SVM* 算法效果不太理想，需要我们改进训练的量，或者换用其他的分类学习方法

2） *FCN* 框架的实验结果

图像语义分割不同于以上任务，这是一个空间密集型的预测任务，换言之，这需要预测一幅图像中所有像素点的类别。以往的用于语义分割的 *CNN*，每个像素点用包围其的对象或区域类别进行标注，但是这种方法不管是在速度上还是精度上都有很大的缺陷。全卷积网络(*FCN*)的概念，针对语义分割训练一个端到端，点对点的网络，达到了 *state-of-the-art*。这是第一次训练端到端的 FCN，用于像素级别的预测。*FCN* 是在 *Caffe* 的深度学习架构下按照网上公开的 *python* 代码例程进行实现的，由于之前在论文已经论述过，对比 3 种性能较好的几种 *CNN*：*AlexNet, VGG16, GoogLeNet* 进行实验，我们在这里选择 *VGG16*。在这里由于我们实验时间的限制，我们主要针对的是 *FCN-8S*，下面给出我们在这个数据集上选取的几张图片的实验结果：

从图中可以看出，虽然在论文上实现的效果较好，但是在我们实际的训练中，效果明显还是不如之前的结果。最后结合我们训练的几张图片，发现这个框架的准确率在 61%左右。

FCN 这个框架的缺点，我们讨论之后发现有以下几点：

1> 可以直观地看出，这种方法和 *Ground truth* 相比，容易丢失较小的目标

2> 得到的结果还是不够精细。进行 8 倍上采样虽然比 32 倍的效果好了很多，但是上采样的结果还是比较模糊和平滑，对图像中的细节不敏感。

3> 对各个像素进行分类，没有充分考虑像素与像素之间的关系，忽略了在通常的基于像素分类的分割方法中使用的空间规整步骤，缺乏空间一致性。

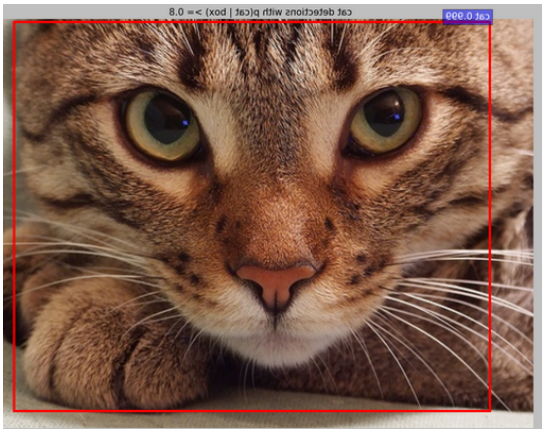
当然它的优点我们也可以有所总结，我们讨论的几点如下：

1> 与传统的用 *CNN* 进行图像分割的方法相比，这种框架可以接受任意大小的输入图像，而不用要求所有的训练图像和测试图像具有同样的尺寸。

2> 这种框架是更加高效，因为避免了由于使用像素块而带来的重复存储和计算卷积的问题。

3） *R-FCN* 的使用和测试结果

R-FCN 主要是用来提取我们所感兴趣的区域，我们发现这个框架是主要基于 *FCN* 和 *Faster R-CNN*，我们在这里也尝试这个框架，也是我们日后所需要进行的一个研究方向。但是在这里我们没有使用 *NYU* 的数据集进行测试，我们选取的是和其论文中使用的类似的数据集进行验证，原理已经在上文中有所阐述，在实际的代码实现和查看之后的结果如下（本代码还是基于 *Caffe* 的 *python* 版本）：



6 结论

本文提出基于 *RGB-D* 和马尔科夫随机场(*MRF*)的 *RGB-D* 场景语义理解框架，和基于卷积神经网络的 *FCN* 图像像素级别分割的框架。*FCN* 框架主要使用目前流行的深度学习框架 *Caffe*，将传统的卷积神经网路的最后的全连接层改为反卷积层，设计一种 *end-to-end* 的图像语义分割模型。与传统方法相比，所提出的语义分割方法无论在主观上还是客观评估上，均具有较好的效果。可以看出深度学习框架的兴起，能够达到和传统的学习方法相同的效果，甚至更好，这也是我们之后所需要研究的内容。深度学习与传统标注方法之间可以相互学习，因为深度学习对于参数设置的要求比较高，而且看重实验，我们需要加强其理论层面的设计，以期可以得到更加好的效果。

参考文献：

1. K. P, K. V, “Efficient inference in fully connected CRFs with Gaussian edge potentials,” in Conf. 25th Annual Conference on Neural Information Processing Systems, Granada, Span, pp.109-117, 2011
2. S. N, H. D, K. P, “Indoor segmentation and support inference from RGBD images,” in Conf. 12th European Conference on Computer Vision, Firenze, Italy, pp.746-760, 2012
3. S. N, F. R, “Indoor scene segmentation using a structured light sensor,” In Conf. IEEE international Conference on Computer Vision Workshops. Barcelona, Spain, pp.601-609, 2011
4. Chen L C, P. G, K. I, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,” arXiv preprint arXiv:1412.7062, 2014
5. J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.
6. Ningbo Wang, Xiaojin Gong and Jilin Liu “A New Depth Descriptor for Pedestrian Detection in RGB-D Images,” in Conf. Pattern Recognition, Jan.2012

7. Nathan Silberman，Derek Hoiem and Push meet Kohli,“Indoor Segmentation and Support Inference from RGBD Images,” Springer Berlin Heidelberg, vol.7576, no.1 pp.746-760,2012
8. J Xiao，A Owens，A Torralba, “SUN3D: A Database of Big Spaces Recon structed using SfM and Object Labels”, in Conf. IEEE International Conference on Computer Vision, pp.1625-1632, 2013
9. Ren X, Bo L, Fox D. “RGB-(D) scene labeling: Features and algorithms，” IEEE Conference on Computer Vision & Pattern Recognition, vol.157, no.10, pp.2759 – 2766, 2012
10. Dai, J., Li, Y., He, K., Sun, J, “R-fcn: Object detection via region-based fully convolutional networks,” arXiv preprint arXiv:1605.06409 (2016)
11. S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” In NIPS, 2015
12. He, K., Zhang, X., Ren, S., Sun, J, “ Identity mappings in deep residual networks,” arXiv preprint arXiv:1603.05027 (2016)
13. Hu, P., Ramanan, D, “ Bottom-up and top-down reasoning with convolutional latent-variable models,” arXiv preprint arXiv:1507.05699 (2015)
14. S. O, D. S, Y. K, “3D object recognition using spin-images for a humanoid stereoscopic vision system,” in Conf. IEEE International Conference on Intelligent Robots and Systems, Beijing, China, 2006
15. M. Maire, S. X. Yu, and P. Perona, “Object detection and segmentation from joint embedding of parts and pixels,” In ICCV, 2011
16. D. Pathak, P. Kraehenbuehl, and T. Darrell, “Constrained convolutional neural networks for weakly supervised segmentation,” In ICCV, 2015
17. P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation,” *IEEE Trans. PAMI*, 2010
18. L. Bo, X. Ren, and D. Fox, “Kernel descriptors for visual recognition,” In *NIPS*, 2010
19. L. Bo, X. Ren, and D. Fox, “Depth Kernel Descriptors for Object Recognition,” In *IROS*, 2011
20. R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, “Liblinear: A library for large linear classification,” The Journal of Machine Learning Research, 9:1871–1874, 2008