# 第二次课程作业文档

2017211416 张志博

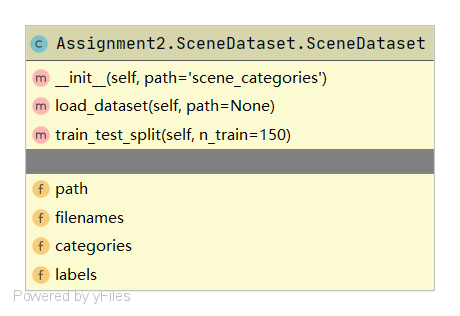
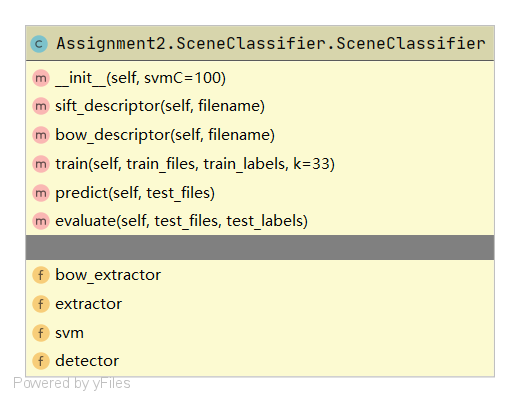
## 作业说明

1. **任务：**编写一个图像分类系统，能够对输入图像进行类别预测。具体的说，利用数据库的2250张训练样本进行训练；对测试集中的2235张样本进行预测。
2. **数据库说明：**scene\_categories数据集包含15个类别（文件夹名就是类别名），每个类中编号前150号的样本作为训练样本，15个类一共2250张训练样本；剩下的样本构成测试集合。
3. **使用知识点：**SIFT特征、K-means、BOW、SVM

## 算法流程

1. **数据集预处理：**按要求读取并将数据分类成训练集和测试集。
2. **提取SIFT特征：**提取2250张训练样本的SIFT特征（实际使用了更好的SURF特征）。
3. **K-means聚类：**因为SIFT特征的特征点个数对于每张图片都是不一样的，因此没有办法直接进行SVM分类，所以使用K-means方法对SIFT特征进行聚类，结果就作为基于BOW的特征提取器的词汇。
4. **SVM：**使用基于BOW提取的特征结果训练SVM分类器。
5. **分类输出：**先提取基于BOW的特征，再使用SVM进行预测即可。

## 函数说明

1. Python类**Scene Dataset**
   1. 示意图：
   2. 属性
      1. **path**：用于加载数据的scene\_categories数据集的路径。
      2. **filenames**：所有数据样本的路径。
      3. **labels**：所有数据样本的类别，以0到14这15个数字表示。
      4. **categories**：数据样本类别的具体名称。
   3. 方法
      1. **load\_dataset**
         1. 输入参数：用于加载数据的scene\_categories数据集的路径path。
         2. 功能：根据所给路径，按指定格式处理，得到并返回filenames、labels、categories。
      2. **train\_test\_split**
         1. 输入参数：训练集的大小n\_train。
         2. 功能：按题目要求，将每类的前n个样本（filenames以及labels）划分至训练集，其他划分至测试集，返回训练集和测试集。
2. Python类**Scene Classifier**
   1. 示意图：
   2. 属性
      1. **detector**：用于提取关键点。
      2. **extractor**：用于提取描述符（特征）。
      3. **bow\_extractor**：基于BOW的描述符提取器。
      4. **svm**：分类器。
   3. 方法
      1. **sift\_descriptor**
         1. 输入参数：需要分析样本的路径filename。
         2. 功能：返回描述符。
      2. **bow\_descriptor**
         1. 输入参数：需要分析样本的路径filename。
         2. 功能：返回基于BOW的描述符提取器计算的描述符。
      3. **train**
         1. 输入参数：训练数据集train\_files，训练标签集train\_labels，K-means参数k。
         2. 功能：提取数据集SIFT特征，进行K-means聚类获得BOW的词汇，再提取基于BOW的描述符，最后配合利用标签集训练SVM。
      4. **predict**
         1. 输入参数：测试数据集test\_files
         2. 功能：提取基于BOW的描述符，再使用SVM对其分类，返回测试标签集。
      5. **evaluate**
         1. 输入参数：测试数据集test\_files，测试标签集test\_labels
         2. 功能：根据测试集的预测结果计算出混淆矩阵，plots输出混淆矩阵图像，打印输出macro、weighted、micro总3种F1-score，返回混淆矩阵。

## 实验设置

注：下面表格中F1-Score(micro)即为论文中使用的Accuracy-Score，未标注的则默认采用F1-Score(micro)，即Accuracy-Score。

1. SIFT or SURF
   1. 实验设置及结果
      1. 无关变量：SVM使用RBF核，参数C=100，参数Gamma=1

|  |  |  |  |
| --- | --- | --- | --- |
|  | K-means的参数K | F1-Score(macro) | F1-Score(micro) |
| SIFT | 100 | 0.491198869 | 0.523937360 |
| SIFT | 1000 | 0.549631654 | 0.568680089 |
| SURF | 100 | 0.508513569 | 0.546756152 |
| SURF | 1000 | 0.560643730 | 0.597762864 |

* + 1. 结论：SURF效果优于SIFT（时间效率上也是）。

1. K-means的参数K
   1. 实验设置及结果
      1. 无关变量：使用SURF特征，SVM使用RBF核，参数C=100，参数Gamma=1

|  |  |  |
| --- | --- | --- |
| K-means的参数K | F1-Score(macro) | F1-Score(micro) |
| 33 | 0.480419618 | 0.515436242 |
| 50 | 0.473135600 | 0.511409396 |
| 100 | 0.508513569 | 0.546756152 |
| 500 | 0.540352371 | 0.578970917 |
| 1000 | 0.560643730 | 0.597762864 |

* + 1. 结论：K=33时可以取得局部较优的结果，K远大于33时，K越大效果越好。

1. SVM
   1. 实验设置及结果
      1. 无关变量：使用SURF特征
      2. 注：此部分比较方式不够严谨，仅供大致比较和示意。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K-means的参数K | SVM核 | SVM的参数C | SVM的参数Gamma | F1-Score(macro) | F1-Score(micro) |
| 100 | RBF | 100 | 1 | 0.508513569 | 0.546756152 |
| 1000 | RBF | 100 | 1 | 0.560643730 | 0.597762864 |
| 100 | LINEAR | 100 |  | 0.493116707 | 0.535570470 |
| 1000 | LINEAR | 100 |  | 0.545162083 | 0.578970917 |
| 100 | RBF | 10 | 1 | 0.472162080 | 0.513646532 |
| 100 | RBF | 1000 | 1 | 0.469760993 | 0.504697987 |
| 100 | RBF | 100 | 0.01 | 0.324843290 | 0.355704698 |
| 100 | RBF | 100 | 0.1 | 0.471007479 | 0.511856823 |
| 100 | RBF | 100 | 10 | 0.478184389 | 0.511409396 |

* + 1. 结论：SVM中，RBF 核优于LINEAR核，C=100和Gamma=1是一组不错的参数。

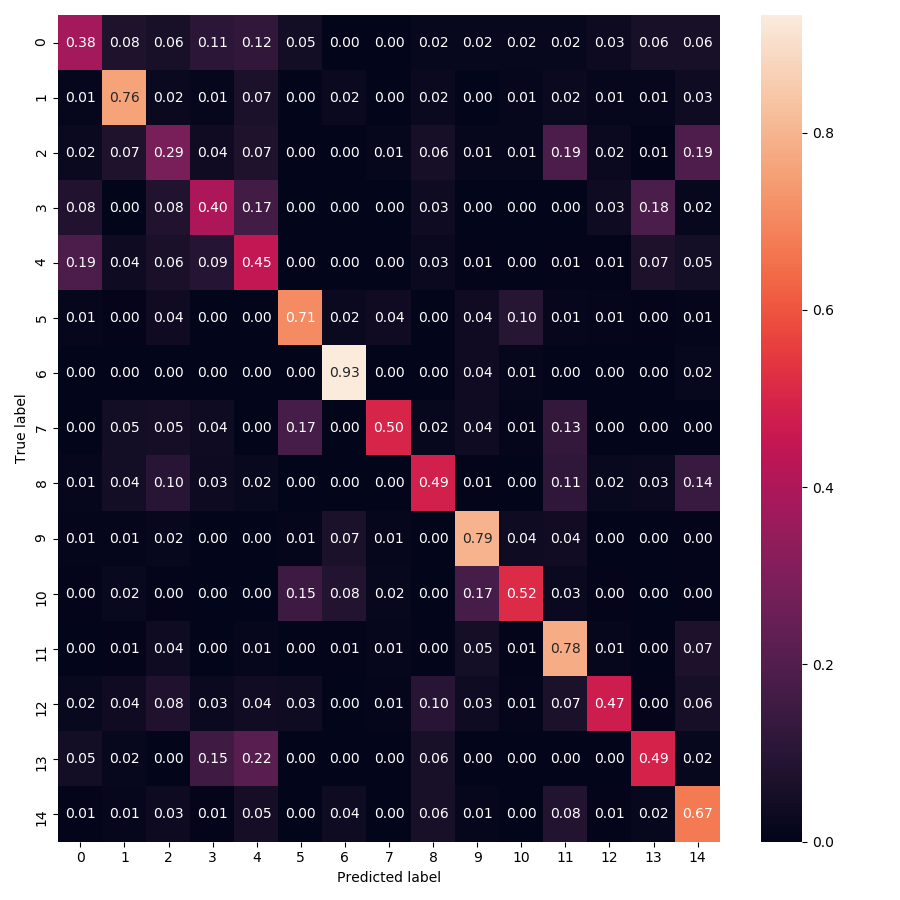
1. 与论文
   1. 结果如下表，性能略微有所提升，但在K较高时效果仍没有明显改善。

|  |  |  |  |
| --- | --- | --- | --- |
| K（或M） | 我的版本 | 论文版本 | 复现效果 |
| 16 | 0.476 | 0.543 | 87% |
| 200 | 0.558 | 0.722 | 77% |
| 400 | 0.571 | 0.748 | 76% |
| 1000 | 0.598 |  |  |

* 1. 结论：我的版本在K相同时性能与论文有差距，这个差距会随K增大而增大。

## 结果分析

1. 混淆矩阵示例



1. 实验中各个环节和各个参数对最终性能的影响
   1. SIFT or SURF
      1. SURF相较于SIFT，特征类似，但鲁棒性更强，最重要的是速度快，在大规模运算时相比SIFT具有压倒性的优势，实际测试也是如此。
   2. K-means
      1. K-means的参数K对试验结果影响很大，本以为可以找到一个合适的数值，比如(训练样本数量/2)^0.5即33或34，可以保证k较小即计算量较小的情况来得到较好的实验结果，但在实验中发现K几乎是越大效果就越好，尽管这会花费更多训练时间。如果想尝试改善性能，还可以考虑其他聚类方法。
      2. 但是在论文中，使用了一个更小的训练集（每类100，不清楚是否是随机划分），K从16提升至200就有一个很大的提升，但我的实现中K从16提升至1000的提升幅度仍很有限，暂不清楚是否有其他影响因素。
   3. SVM
      1. 在SVM的Kernel中，RBF核是广泛使用且效果较好的核，但也有说法称当样本特征较多时，LINEAR核的计算更加高效，但经过测试，发现计算量减少难以掩盖训练效果较差的结果，因此最终还是选择使用了RBF核。参数C经实验发现在100时效果较好，Gamma则保持在默认的1即可。如果需要更加好的参数，可能需要自动化的参数寻优方法。
   4. Level与Pyramid
      1. 因为基本的性能并未达标，且暂未找到与论文性能存在差异的原因，为减少其他变量以便测试，再加上以原论文中所述，要想提升性能，增加特征量是最有效的，所以暂未实现这两项功能。