统计信号处理大作业
烟草杂质识别问题

蒙治伸　无 47 班　2014011207
刘前　无 47 班　2014011216
王舒鹤　无 47 班　2014011221

1　引　言

烟草中识别杂质是一个较为复杂的问题。从人眼观察的角度来看，烟草与各类异物在颜色、形状、大小、材质等方面都存在差异。但是由于被检测的烟草的分布较不规律，同时杂质的种类较多，因此是一个对复杂的问题。

针对这一问题，我们提出的基本的解决思路如下：首先对图像进行预处理，将训练样本分为标准的图像块便于后序处理；然后从每个图像块上的 RGB 分布这一原始特征出发，提取出对样本与杂质而言有区分性的特征；接下来根据提取的正负样本的特征训练分类器，根据训练样本调整分类器的参数；最后将需要被检测的图像输入分类器中即可得到分类检测结果。下面我们将从理论与具体实现上对上述过程作详细介绍。

由于现有的模式识别与分类的理论较多，我们尝试了不同的方法对样本进行特征提取以及分类。经多次实验，我们发现 RGB 高斯拟合和高维特征空间核方法是两种识别效果相对较好的方法。以下我们分别介绍这两种方法的过程以及理论分析。

2　RGB 空间高斯拟合方法

2.1　问题分析

2.1.1　问题定性

给定一幅烟草图像，只需关注哪些部分是正常的烟草，哪些是杂质，因此，可以将这一问题视为简单的分类问题、但是这样对于这种情况的分类会出现一定的问题：理论上可以将烟草视为一类，杂质视为另外一类；但是对于杂质来说，包含塑料、铁丝、纸片等很多更细小的类别，这些小类之间的特征一般也是有很大的区别，在处理时无法简单地视作一类。而实际上对杂质进行再分类在杂质检测这一问题背景下是没有什么意义的。因而，该问题更接近于一个异常点检测 (Anomaly Detection) 问题，或者称为一类分类 (One-Class Classification) 问题。将烟草视为一类，给这一类确定一个标准，当不满足该标准时，就视为异常点，也就是杂质。

2.1.2　图像中的特征提取

一幅图像有很多特征，包括颜色、亮度、灰度、纹理特征等等。哪些特征对于烟草识别问题才是最有效的呢？

首先，从目前个人的知识能力，基本排除使用纹理特征。一方面，纹理特征在数学表达上比较难于操作，不仅需要找到能够表示纹理特征的量，还需要更高级的知识[1]；另一方面，在烟草杂质识别这一问题中，有些杂质往往与烟草具有相似的纹理特征，此时分辨效果较差，不是完全可行的。

之后纳入考虑范围的特征包括颜色和明暗特征，最常用的是 RGB 三维特征、HSI 模型、Ohta 空间等等。本次大作业尝试了 RGB 颜色模型和 HSI 模型，发现 RGB 模型下检测杂质的效果要明显优于 HSI 模型。本文将详细介绍基于 RGB 颜色模型的检测算法。

2.2　图像数据的降维

RGB 颜色模型下，图像的数据包含了每个像素点的 RGB 值；而对于一幅完整的烟草图像，这往往意味着数千万个像素点，每个像素点又包含了 RGB 三维的数据信息，因而仅一幅烟草图像就包含了数十 MB 的信息，更何况在实际应用中往往还需要对烟草实时在线进行检测。如果完全利用这些数据，不论从时间还是从 (存储) 空间都需要很大的代价，考虑到这一点，必须对图像数据进行降维。

数据的降维应当以尽量少地损失有效信息为前提，常见方法包含 PCA(主成分分析) 和 LDA(线性判别分析) 等。在烟草图像中，颜色分布是最重要的有效信息，因而降维时应该尽量不改变烟草图像中 R、G、B 三个通道的分布。本文采用一种简单而有效的方法，是将烟草图像进行分块之后，每小块取 R、G、B 三个通道各自的平均值作为该小块的数据特征，从而使得数据量大大减小。

2.3　算法描述

通过之前的分析和数据准备，可以设计得到一种基于 RGB 的检测算法。算法的流程主要包括：图像分块、RGB 求平均值、确定数据分布、选择距离度量、基于训练样本计算阈值、重叠式遍历检测图像以及在测试样本上评估算法。

2.3.1　图像分块

图像分块主要有两个目的：一是减小数据量，每个小块可以仅用 R、G、B 三通道的均值作为数据信息；二是便于对图像进行操作，利用“分而治之”的思想，将整幅图像的杂质检测问题转化为每一小块的检测问题，最后将所有小块检测的结果综合起来即可作为整幅图像的检测结果。

图像分块的主要问题是分块的大小。默认使用正方形的

[1] 包括灰度差分统计法、灰度共生矩阵法等等。

小块，这样执行起来明显比其他形状的小块更加方便。正方形小块的边长是一个重要的参数：如果正方形过小，不但削弱了减小数据量的优点，还可能因为关注的范围过小导致大量的虚警；但如果正方形过大，则很可能会漏检一些较小的杂质。通过尝试和比较，发现正方形大小在 50 至 70 之间都是比较合适的，在这个范围内，小块边长对最终结果的影响可以基本忽略。由于本次大作业提供的图像边长一般为 2048，4096 等数字，这些都是 64 的倍数，因而最终选择将图像分割为边长为 64 的正方形小块。

说明：后续对图像分块检测时，并不是简单地将整幅图直接分割，而是进行重叠式的分割，需要引入一个步长参数(详见**重叠式遍历检测**部分)。在确定了步长参数及块边长为 64 后，为方便遍历图像，需要对烟草图像的长和宽进行调整，确保遍历时块大小不变并且能够遍历图像的每个位置。对图像大小的预处理方法采用“补全”的思想，计算出水平和竖直方向上需要补全的数量，并使用最近的像素进行补全。

2.3.2 RGB 求平均值

对于每个 64×64 的小块，分别求 R、G、B 三个通道的平均值，作为该小块的数据特征。R、G、B 三个通道的取值范围均为 0 至 255。在对测试图像的每个小块进行检测时，同样以 R、G、B 三个均值作为判断的依据。

2.3.3 确定数据的分布

每个小块对应 R、G、B 三个数据，设测试样本包含了 N 个小块，则对应为 N 行 3 列的数据矩阵，每行代表一个小正方形，3 列分别对应 R、G、B。

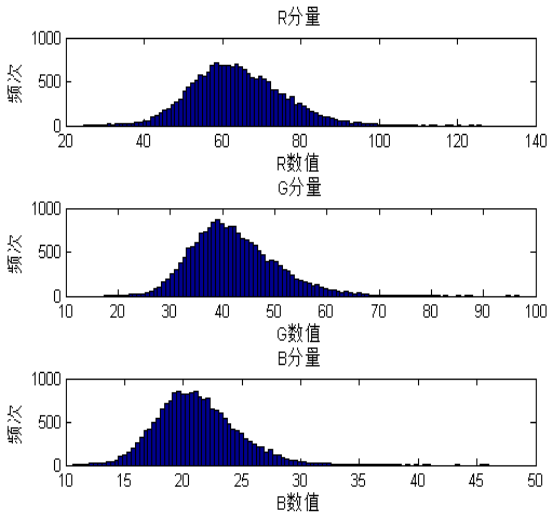

图 1: 烟草图像 R、G、B 三通道的分布

具备了这些数据，尝试通过作直方图发现其分布规律。分别对 R、G、B 作频数分布直方图，发现结果如上图所示，明显具有高斯分布的形状。事实上，这一发现在文献 1 中也得到了证实：自然图像中 RGB 的分布趋向于高斯分布，高斯分布是最适合 RGB 色彩空间模型的分布。因此，本次大作业将 R、G、B 三维数据视为三维联合高斯分布，这一结论对后续距离度量的选择和阈值的训练都有很大的帮助。

2.4 距离度量 (Metric) 的选择

对于异常点检测问题，需要对“正常”定义一个标准，不符合此标准的点即为异常点。由统计信号处理课程的相关知识，可以从另一个角度对该问题进行理解。

给定一个信号，判断该信号是否属于一个已知分布，统计信号处理中的方法往往视情况根据一些标准 (贝叶斯准则、极大极小风险准则等) 确定一个阈值，当信号处于阈值以内的范围时，即认为该信号属于已知分布；否则认为是异常点。

对于烟草检测问题，此时的信号相当于给定一个正方形小块的 R、G、B 均值，已知分布相当于通过烟草样本训练得到的 R、G、B 的联合高斯分布。

在确定阈值之前，首先需要选择一个衡量标准 (Mertic)，用什么数字特征来度量新获取的数据与已知分布之间的距离呢？

常用的有以下两种度量 (Metric)。设联合高斯分布的均值向量为 $\mu = [\mu_1, \mu_2, \mu_3]$，协方差矩阵为 Σ。

1. 欧氏距离 (Euclidean Distance)

假设测试数据中某一样本为 $X = (x_1, x_2, x_3)$，欧式距离可以表示为：

$$L_E = ||X - \mu||^2 = (X - \mu)^{\mathrm{T}}(X - \mu) = \sum_{i=1}^{3}(x_i - \mu_i)^2$$

2. 马氏距离 (Mahalanobis Distance)

同样假设测试数据中某一样本为 $X = (x_1, x_2, x_3)$，马氏距离可以表示为：

$$L_M = (X - \mu)^{\mathrm{T}}\Sigma^{-1}(X - \mu)$$

一开始使用的是欧氏距离，但是在后面寻找阈值时发现不同图像的阈值差距较大。于是又尝试了马氏距离，高斯分布下，马氏距离的度量性能比欧氏距离好很多，从而对算法进行了优化。事实上，马氏距离具有诸多优点：a) 不受量纲的影响，两点之间的马氏距离与原始数据的测量单位无关；b) 由标准化数据和中心化数据 (即原始数据与均值之差) 计算出的二点之间的马氏距离相同；3) 马氏距离还可以排除变量之间的相关性的干扰。因而本次大作业选择马氏距离作为度量标准。

2.4.1 基于训练样本计算阈值

确定使用马氏距离作为度量之后，即需要基于训练样本确定阈值，阈值的大小直接决定了虚警概率和漏检概率的大小。在烟草杂质检测问题中，更重要的指标是漏检概率，将漏检概率降低到较低水平，允许有一定的虚警，只要虚警概率不要太高就可以，因为虚警的烟草往往可以在之后的检测中得到验证。

训练样本分为烟草样本 (正样本) 和杂质样本 (负样本)，都是从测试图像中采集得到的。测试图像中框出了杂质，本人使用 MATLAB 抠取了其中的杂质部分，以便于杂质样本的采集。

获取训练样本之后，对标注好的烟草样本 (正样本) 和杂质样本 (负样本) 进行分类。尝试选择不同的马氏距离

阈值，检测其分类的性能。本次大作业选择的阈值是使得测试集的漏检概率接近虚警概率 +0.15 左右时的马氏距离，此时杂质的检测效果在测试集上表现不错，即可将该阈值作为训练结果，用于测试图像的杂质检测。

说明：在实际操作中，基于训练样本确定的阈值并不能保证对所有的测试样本都能达到较好的效果，可能需要一定的微调，在微调之后的分辨效果往往非常精确。但是这一操作在实际应用中不太实际，缺乏可行性。为了充分体现该算法在实际应用中的表现情况，本次大作业中采用统一的阈值，因而算法的性能实际上没有发挥到理论的最优水平。

2.4.2 重叠式遍历检测图像

根据“分而治之”的思想，本算法对每一小块的图像进行检测。起初采用的方法是直接将完整的图像分割成互不重叠的小块，然后遍历每个小块进行检测。不过后来发现这个算法有一些弊端：如果一个杂质相对于正方形小块较小，而且在分割时又恰好被分在了两个或两个以上的正方形小块中，这样就会导致每个小块都无法检测出杂质。为了解决这一问题，本算法在分割和遍历正方形小块时增加了一个步长参数，能够实现相邻的小块之间有重叠的部分。添加了这一参数后，经过训练选择一个合适的步长后，检测杂质的查全率明显比之前有所提高。

2.4.3 在测试图像上评估算法

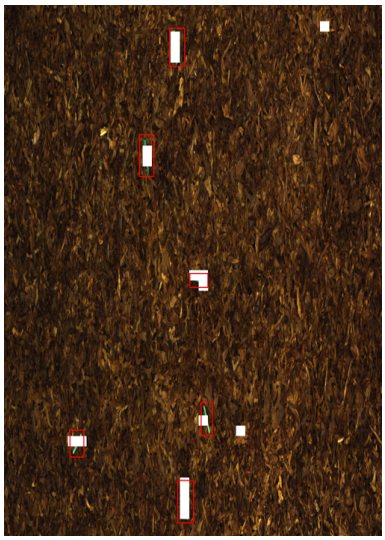

图 2: 算法在测试图像上的表现

上图展示了其中一个测试图像上的性能，其中白色部分是基于该算法检测出来的杂质，红色框出的是真实的杂质。可以看出，基于 RGB 颜色模型的检测算法基本检测出该图像中的全部杂质，并且虚警也只是极个别情况，能够满足实际应用的需要。因此，这一算法在解决烟草杂质问题上是很成功的。

表1展示了算法在给出的 12 个测试图像上的结果。

测试图像	实际杂质数	检测杂质数	虚警数
2.bmp	10	9	5
3.bmp	10	10	3
4.bmp	10	10	4
5.bmp	10	9	5
6.bmp	10	10	4
7.bmp	10	10	4
8.bmp	10	9	3
9.bmp	10	5	5
10.bmp	10	6	3
11.bmp	10	5	2
12.bmp	9	1	2
13.bmp	12	5	4

表 1: RGB 空间高斯拟合方法结果分析

2.5 总结与分析

本部分设计并优化了“基于 RGB 颜色模型的烟草杂质检测算法”，算法的主要流程包括：图像分块、RGB 求平均值、确定数据分布、选择距离度量、基于训练样本计算阈值。

在测试样本上对算法的泛化性能进行了测试，表现出了良好的性能，尤其是测试图像中的 2 至 8.bmp，杂质几乎全部被检测出来 (95% 以上)，而且虚警也控制在合理的范围内。同时，本算法的时间复杂度比较低，每幅图像的处理时间在 2s 左右，完全能够满足实际应用中在线检测杂质的需求。因而综合考虑算法的检测性能和耗费时间，本算法具有较高的可行性。

但是，这一算法仍然存在一些比较严重的问题和缺陷，比如：

1. 算法易受到亮度 (光照) 的影响，当图像中局部亮度过高时，会导致虚警的增加，可以预先对图像进行对比度和亮度方面的处理，使得在检测前图像的亮度基本一致；
2. 基于训练样本得到的阈值，无法保证对所有的测试样本都能达到预期的效果，鲁棒性可能较差；
3. 最大的问题是：对于一些颜色与烟草相近的杆状杂质，该算法几乎无法检测出来 (测试图像 9.bmp 至 12.bmp, 尤其是 12.bmp)；等等。

以上缺陷需要增加其他的机制，或者使用其他的处理手段进行更一步的优化。下一种算法在这些测试样例上的性能比基于 RGB 的识别算法要好一些，后文将有细致的分析。

3 高维特征空间核方法

3.1 分类与识别理论

对于线性可分且特征分布近似服从高斯分布的数据，采用高斯拟合的方法既简洁，耗时短且效果较好。但在实际处理时我们对正负样本的特征分别用直方图方法得到其分布，发现对于一部分特征的直方图，正负样本的重合度很高，因此如果采用线性的分类方法，并不能很好地将烟草与杂质分离开。因此可能需要更加抽象的，非线性的特征来描述数据。我们猜测如果将数据映射到高维特征空间中，可能会令原来区分度较低的数据提高区分度，即采用核方法增加传统的线性的分类器的学习能力。下面我们简单介绍核方法的基本原理，以及采用核方法进行烟草模式识别的流程。

3.1.1 核方法的基本原理

核方法的基本原理为：若数据非线性可分，则使用一个非线性变换 $\Phi(\cdot)$ 将输入模式空间中的数据映射到高维特征空间中, 在其中再构造新的线性的分类函数。在此过程中，不必明确非线性变换的具体表达式，而只需要知道在该空间中的内积在原模式空间中的形式——通常以核函数的形式给出，即

$$< \boldsymbol{x}, \boldsymbol{y} > \rightarrow K(\boldsymbol{x}, \boldsymbol{y}) =< \Phi(\boldsymbol{x}), \Phi(\boldsymbol{y}) >$$

在具体问题中，如果只需要用到特征的内积，则可以将复杂的非线性变换函数的计算转为简单的核函数的计算，从而极大地降低非线性变换的计算量。在这里，我们所使用的核函数均为高斯径向基核函数，其表达式为 $K(\boldsymbol{x}, \boldsymbol{y}) = exp(-||\boldsymbol{x} - \boldsymbol{y}||^2/\sigma^2)$。从理论上讲，高斯径向基核函数对于非线性支持向量机具有良好的泛化性能。

3.1.2 核主成分分析 (KPCA) 变换特征

由于正常样本的数量远多于异常样本的数量，我们采用了单类支持向量机 SVDD 方法对数据在高维空间构造一个“超球”决策平面以分割正负样本。为了方便构造分类器，我们先要在高维空间对特征进行变化，使得变换后的特征互相独立且方差彼此相近，即能使高维空间的特征基底两两垂直，且样本点在高维空间的分布基本上包裹在一个超球面内。(具体见下一小节)

KPCA 变换特征向量的原理如下：对 N 个输入模式空间中的 M 维特征 $x_i, i = 1, 2, ..., N$，其在高维特征空间中的协方差阵可表示为

$$C = \Phi(X)\Phi(X)^T = \sum_{j=1}^{N} \Phi(x_j)\Phi(x_j)^T$$

对于 C 的特征向量 v，设其满足方程：

$$Cv = \lambda v$$

代入 C 的表达式即有：

$$v = \frac{1}{\lambda} \sum_{j=1}^{N} \Phi(x_j) < \Phi(x_j), v >= \sum_{j=1}^{N} \alpha_j \Phi(x_j)$$

即特征向量 v 在 $\Phi(x_j)$ 张成的子空间内，同时考虑到

$$\lambda < \Phi(x_j), v >=< \Phi(x_j), Cv >$$

联立上面两式即有

$$\lambda K\alpha = K^2\alpha$$

也即

$$\lambda\alpha = K\alpha$$

其中 K 为 $N \times N$ 的矩阵，满足 $K_{ij} =< \Phi(x_i), \Phi(x_j) >$，$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_N]$。

由于数据投影到高维空间后，需要其在每一个特征方向上的方差相等，这样才能使数据分布在高维空间中呈“球状”。而数据在第 k 个特征向量的方向上的方差为

$$\frac{1}{N} \sum_{j=1}^{N} (\hat{x}_j)^2 = \frac{1}{N} \sum_{j=1}^{N} (\sum_{i=1}^{N} \alpha_i^k K(x_i, x_j))^2 = \frac{1}{N} (\boldsymbol{\alpha}^k)^T K^2 \boldsymbol{\alpha}^k$$

要令其对所有特征方向都相同，则必有

$$\lambda_k^2 (\boldsymbol{\alpha}^k \cdot \boldsymbol{\alpha}^k) = 1$$

故在实际处理时，只需要得到 K，并对 K 作特征值分解，然后根据特征值的大小调整对应的特征向量的长度即可。

3.1.3 采用单类支持向量机的 SVDD 方法分类

SVDD(support vector data description) 的方法是在高维特征空间中，找到一个体积尽量少且包含尽量多训练样本的超球面，在判决时将分布在超球面外的数据判定为杂质。如下图所示。这种分类方法的好处在于训练时只需要使用正常样本，而通过判别正常和异常样本来调整参数达到最优。

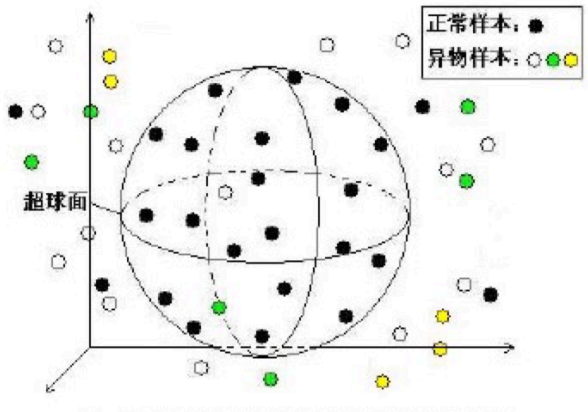

图 3: SVDD 超球面判决杂质示意图

设 a 和 R 为超球面的球心和半径，则分类问题可等效为求解一个二次优化问题：

$$minF(R, a, \xi_i) = R^2 + C \sum_{i=1}^{N} \xi_i,$$

$$s.t. ||\Phi(x_i) - a||^2 \leqslant R^2 + \xi_i, \xi_i \geqslant 0, i = 1, 2, ..., N$$

其中参数 C 为正则化系数，用于控制被分错的样本的比例，从而实现超球面半径与包含样本数的折衷。ξ_i 为错分松弛因子。由拉格朗日乘子法最终可得上述问题的对偶形式为

$$minF(\boldsymbol{\alpha}) = min \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i K(x_i, x_j),$$

$$s.t. \sum_{i=1}^{N} \alpha_i = 1, 0 \leqslant \alpha_i \leqslant C, i = 1, 2, ..., N$$

其中 $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ 为核函数。这样原问题就转化为求解 $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_N]$ 的二次优化问题。由于我们选择的核函数为高斯径向基核函数，因此 $K(x_i, x_i) \equiv 1$，则原问题即等价于求解以下标准的二次规划问题

$$minF(\boldsymbol{\alpha}) = min \boldsymbol{\alpha}^T K \boldsymbol{\alpha}$$

通常情况下，大部分 α_i 为零，不为 0 的 α_i 对应的样本即为支持向量。可知对于 $0 < \alpha_i < C$ 的样本满足

$$\begin{cases} \xi_i = 0 \\ R^2 - K(x_i, x_i) + 2\sum_{j=1}^{N} \alpha_j K(x_i, x_j) - a^2 = 0 \end{cases}$$

因此若令 SV 为支持向量的集合，定义 $K_{ij} = K(x_i, x_j), E_i = \sum_{j=1}^{N} \alpha_j K_{ij}$，则对任意 $x_i \in SV$，应有

$$a^2 - R^2 = 2E_i - K_{ii} = 2E_i - 1$$

对所有支持向量取平均，设支持向量的个数为 N_{sv}，可定义离心系数：

$$\omega = a^2 - R^2 = \frac{1}{N_{sv}} \sum_{x \in SV} 2E_i - 1$$

其中 ω 体现了超球球心与半径之间的空间关系。

综上，对于任意测试样本 z，判别函数即为

$$f(z) = sgn((\Phi(z) - a)^T(\Phi(z) - a) - R^2)$$

可知 $f(z) = 1$ 则 z 为异常样本，否则为正常样本。代入 ω 即有

$$f(z) = sgn(1 - 2\sum_{i=1}^{N} \alpha_i K(z, x_i) + \omega)$$

3.2 识别算法实现

具体的算法实现是根据上面的理论分析完成的，但由于实际问题的复杂性，因此在实现时还有一些修改与折衷。而核方法与 RGB 高斯拟合的方法在图像预处理阶段核方法与基本相同，我们主要讨论接下来的特征提取与分类识别部分。

3.2.1 特征提取

样本图像的特征主要包括色度特征，与纹理和均匀性特征。起初我们采用图像块的 RGB 直方图向量作为图像的色度特征，共 $256 \times 3 = 768$ 维；而对图像块的 RGB 层分别作 DCT 变换，并将相同的频率的分量的平方相加，最终对所有频率作归一化处理，作为均匀性的特征；由于是边长为 64 的图像块，因此每一层的频率分量有 128 个，即共有 $128 \times 3 = 384$ 维。故原始特征共有 1152 维。但后续实验效果表明这种取法并不够理想，同时由于样本的特征维数过高，使得训练样本与测试的时间都过长，空间复杂度也相当高。因此我们认为需要选取较少的，同时能更加精炼地概括图像特点的特征。

最终我们选取了 11 个维度的特征作为核方法的样本初始特征，其中色度特征有 8 个，分别是图像块的 R,G,B 的平均值，RGB 作非线性变换到 HSV 空间中之后 H,S,V 的平均值，RGB 作非线性变化到 Lab 空间中之后 a,b 的平均值；均匀性特征有 3 个，分别是对 RGB 三层作灰度差分图像，再取各层灰度差分的平均值。

HSV 是根据颜色的直观特性创建的颜色空间，其参数分别为色调 (H), 饱和度 (S)，明度 (V)。而 Lab 则基于人对颜色的感觉，L 表示明度，a 表示从洋红色至绿色的范围，b 表示从黄色至蓝色的范围。由于 RGB 到这两个空间的转换均为非线性转换，因此可以在一定程度上区分一些在原来 RGB 空间不可分的正负样本。而在 matlab 中，由 RGB 到这两个颜色空间的转换可直接用 matlab 自带的函数得到。

而对于纹理与均匀性，一种简单的方法就是利用灰度变化的灰度差分来进行纹理分析。如果假设像素点 P_1 和 P_2 的灰度分别为 $g(P_1), g(P_2)$，则这两点的灰度差分即为

$$\triangle g = |g(P_1) - g(P_2)|$$

实际统计图像灰度差分特性时，可考察图像中各像素点与其相邻各像素点的灰度差。为简单起见，我们采用在两两相邻的 4 个像素中，用两组对角线像素的灰度差分的绝对值之和作为某一像素点的灰度差分值，即

$$\triangle g(i,j) = |g(i,j) - g(i+1,j+1)| + |g(i,j+1) - g(i+1,j)|$$

然后再对整个图像块取平均值即可。

3.2.2 训练分类器

对于核方法，若训练样本的数量为 N，则其空间复杂度为 $O(N)$，而时间复杂度为 $O(N^2)$，因此用于训练的样本数量不能太高。在对用于训练的图像进行处理后，我们一共得到了 20000 个左右的 64×64 的烟草图像块与 700 个左右的杂质图像块。我们选择随机抽取烟草的八分之一用于训练，而将剩余的烟草以及杂质用于交叉验证分类器效果以及调整参数上。

首先需要先对训练样本用 KPCA 规范化，得到用于规范的矩阵 U。由于对于任一其他样本 z，其在高维空间的规范需要用到 z 与所有训练样本的内积，因此还需要保存所有训练样本的特征。

由于我们选择的核函数为高斯径向基核函数，因此分类器的参数共有两个：σ 和 C。其中 σ 决定了非线性映射的程度，C 控制支持向量占训练样本的比例。最终我们使用 $\sigma = 2.5, C = 0.5$。在训练过程中要解决的二次优化问题可采用 matlab 自带的二次优化函数 quadprog 得到最优解。

训练完成后可以得到支持向量及其对应的比例 α_i，离心系数 ω。将这些以及所有训练样本的特征，规范矩阵 U 和核方法的参数 σ 存入一个文件中，在之后的识别过程中只需要提取这些即可完成识别。

3.2.3 识别图像

将要检测的图像输入我们的识别程序后，首先对其分块扫描处理，然后对所有图像块在由样本特征张成的高维空间中进行规范化，再输入到训练好参数与支持向量的分类器中即可得到结果。

但在实际操作中，我们发现上述算法对于新的测试图像的识别效果并不好，虽然漏检概率很小，能够基本上检测所有杂质，但虚警概率非常高 (5%-10%)。我们认为这是由于训练样本取自多个图像，其分布区域是多个图像平均的结果；而对于某一单个的图像，其数据特征在高位空间的分布的中心可能并未与训练样本重合。我们采用了对检测图像进行再一次自适应的 svdd 训练并检测的方法来弥补。具体做法为：对被第一次分类分为烟草的图像块，再随机选取其中的十二分之一，训练一个新的 svdd 分类器，然后将在第一次分类中被分为杂质的图像块输入新的分类器得到结果。经实验表明，这种方法大大降低了虚警率，将在第一次中被判为杂质的正常区域重新判回为烟草。

但由于新的 svdd 分类器的训练样本的特征分布中心实

际上仍然不是整个图像中烟草的特征分布中心，仍然存在不少小的零散的被误检的图像块。我们的处理方法为遍历检测结果中所有的杂质的连通域，将所有像素个数小于 10 个的连通域删除。最终的效果比较令人满意。

3.3 实验结果分析

我们对所给的测试样例进行检测，并统计其检测效果。需要特殊说明的是我们的算法中没有考虑地面与杂质的区别 (地面的种类也很多，且引入地面后问题的复杂性超出我们考虑的范围)；通常情况下，由于地面与烟草的差异较大，会将地面识别为杂质。因此我们在虚警与误检中均未考虑对地面的检测结果。以下表格为核方法对 13 个测试图像的检测效果：

测试图像	实际杂质数	检测杂质数	虚警数
1.bmp	0	0	7
2.bmp	10	10	9
3.bmp	10	10	7
4.bmp	10	10	2
5.bmp	10	10	5
6.bmp	10	10	10
7.bmp	10	10	5
8.bmp	10	9	11
9.bmp	10	10	8
10.bmp	10	10	10
11.bmp	10	7	9
12.bmp	9	9	7
13.bmp	12	10	10

表 2: 高维特征空间核方法结果分析

可以看到，核方法对于杂质的识别效果较好，漏检的概率较低；虚警的概率较小，基本上保持在可接受的范围内。由于大作业的实际要求为，在满足一定的虚警概率下保证漏检率尽可能低，因此核方法的效果很好；同时对于不同的图像，杂质的漏检概率与虚警概率也基本上保持稳定，因此具有较高的鲁棒性。

但由于这一方法仍然存在不少问题，例如在特征的选取上不够好，在 KPCA 处理时为节约存储空间没有对数据进行中心化处理，SVDD 的参数选择不一定是最优参数，SVDD 的训练过程没有优化等等，因此还存在一些问题，包括：虚警的数量较多，在色度与杂质接近的图中，虚警的数量在 7 ∼ 10 个左右，这说明分类器的泛化能力仍然不够强；对于色度与杂质接近的图，漏检的概率也较高，如 11.bmp，算法没有最有区分性的特征；时间复杂度过高，一般而言，分类器的训练时间在 2 ∼ 3 分钟左右，而检测一张 9000×2048 的图像的用时约为 4 ∼ 5 分钟，因此不能用于在线检测；空间复杂度较高，主要是需要存储训练样本（约 2500 个）的特征 (11 维)，以及同样大小的转换矩阵 U。我们也会在之后对这些局限之处继续思考，寻找解决方案。

4 总 结

在烟草杂质识别的问题中，我们最终得到了两种效果较好的方法——RGB 空间高斯拟合方法和高维特征空间核方法，并用测试集对两种方法进行了验证，对于杂质与烟草区别较大的图像，前者的虚警数量更少，且运行时间很短，空间复杂度低，优于后者的性能；但是当杂质与烟草的相似度提高时，后一种方法表现出了较高的稳定性，虽然运行时间较长，但基本上能够检测出杂质同时控制虚警的数量。在当前的结果之上，要继续提高时间上与识别效率上的性能，还需要我们继续寻找更具有代表性的特征与合适的模型，我们也将继续我们的思考之路。

参考文献

[1] 褚江, 陈强. 自然图像颜色空间统计规律性研究 [J]. 计算机科学, 2014, 41(11):309-312.

[2] 冯爱民, 陈松灿. 基于核的单类分类器研究 [J]. 南京师范大学学报 (工程技术版), 2008, 8(4):1-6.

[3] DAVID M. J. TAX, PIOTR JUSZCZAK. Kernel Whitening for One-Class Classification[C]// International Workshop on Pattern Recognition with Support Vector Machines. Springer-Verlag, 2002:40-52.

[4] 房小兆. 超球结构支持向量机的研究与应用 [D]. 广东工业大学, 2011.

[5] 刘笑嶂. 核方法的若干关键问题研究及其在人脸图像分析中的应用 [D]. 中山大学, 2010.

[6] 田昊. 基于图像处理的机采棉杂质检测技术研究 [D]. 石河子大学, 2014.

[7] 朱晓芳. 基于支持向量机的田间杂草识别方法研究 [D]. 江苏大学, 2010.

[8] 谈蓉蓉. 利用颜色和形状特征的杂草识别方法研究 [D]. 江苏大学, 2009.

[9] 陈文涛. 烟草异物在线高速模式识别与剔除技术研究 [D]. 重庆大学, 2003.

[10] 刘军. 烟草在线异物实时识别与自动剔除系统研究 [D]. 重庆大学, 2003.

[11] 陈斌. 异常检测方法及其关键技术研究 [D]. 南京航空航天大学, 2013.

[12] 焦智. 异性纤维检测算法的研究和实现 [D]. 北京工业大学, 2006.

[13] 姚富光. 智能高速在线异物识别分拣关键技术研究 [D]. 重庆大学, 2009.

附　代码和文件目录

1　基于 RGB 模型的高斯拟合方法

- tabacoo.mat　烟草的训练样本数据
- impurity.mat　杂质的训练样本数据
- rgb_dist.m　作图得到训练图像中烟草的 RGB 分布
- extraction.m　抠取训练图像中标记出的杂质
- savepic.m　分别保存抠取的杂质及剩下的烟草图像
- getpic.m　调用前两个函数分离训练图像的杂质和烟草
- rgb_model_train.m　基于训练图像得到烟草 RGB 相关参数
- rgb_model.mat　存储训练得到的烟草 RGB 相关参数
- rgb_detect.m　实现对烟草图像的杂质检测
- rgb_comp.m　将算法检测的结果与真实杂质结果对比
- rgb_test.m　将所有算法应用于所有测试图像并对比

2　高维特征空间核方法

- main.m　主函数，检测待测图像中的杂质并生成给杂质加框的图像
- classifer.m　分类器，根据输入的图像块的特征判断是否为杂质
- drawline.m　根据杂质的集合在原图上用蓝色方框框出杂质
- extFeature.m　根据输入的 64×64 的图像块提取 11 维特征
- FI.m　高斯径向基核函数，输入为两个 $N \times M$ 的矩阵 A, B，输出为长为 N 的向量，其中第 i 个元素为 A, B 的第 i 行的高维空间“内积”
- findsetarea.m　用递归的方法找到二值图中的连通域
- kpca.m　核主成分分析，输入为 N 个特征，输出为规范化后的特征与规范矩阵
- kwhiten.m　对待检测图像的特征进行核白化，输入为训练样本的特征与规范矩阵，输出为规范化后或白化后的待测图像特征
- near.m　判断杂质是否是孤立的像素点或处于一个像素数量很小的连通域中
- svdd.m　用 SVDD 训练样本得到分类器
- uniwrong.m　得到整张图中所有的杂质的连通分支
- sv.mat　保存了已经训练好的分类器的特征和参数，可直接用于第一次分类检测