

基于优化 DBSCAN 算法的玉米种子纯度识别*

刘双喜¹ 王盼¹ 张春庆² 王金星¹

(1. 山东农业大学山东省园艺机械与装备重点实验室, 泰安 271018; 2. 山东农业大学农学院, 泰安 271018)

【摘要】为快速有效地识别玉米种子纯度, 针对玉米种子图像特征, 对玉米种子的图像处理方法和聚类算法进行研究, 提出一种基于最远优先遍历的 DBSCAN 玉米种子纯度识别算法。该方法首先提取玉米种子冠部核心区域的 RGB、HIS 和 Lab 种颜色模型特征参数, 选取 H 、 S 、 B 作为识别向量; 其次通过最远优先遍历算法剔除密度差异特征向量边缘异常散点; 最后采用 DBSCAN 算法进行密度聚类。实验结果表明, 该方法玉米纯度识别正确率达 93.3%。

关键词: 玉米 种子纯度 识别 聚类 DBSCAN

中图分类号: TP391.4; S339.5+3 文献标识码: A 文章编号: 1000-1298(2012)04-0188-05

Maize Purity Identification Based on Improved DBSCAN Algorithm

Liu Shuangxi¹ Wang Pan¹ Zhang Chunqing² Wang Jinxing¹

(1. Shandong Provincial Key Laboratory of Horticultural Machineries and Equipments, Shandong Agricultural University, Taian 271018, China 2. College of Agriculture, Shandong Agricultural University, Taian 271018, China)

Abstract

In order to identify maize purity rapidly and efficiently, the image processing technology and clustering algorithm were studied according to the maize seed and characteristics of the seed images. An improved DBSCAN on the basis of farthest first traversal algorithm (FFT) adapting to maize seeds purity identification was proposed. The color features parameters of the RGB, HIS and Lab color models of maize crown core area were extracted. H , S and B were selected to be the effective characteristic vector. The abnormal points of different density characteristic vector points were separated by FFT. Then clustering results were combined after local density cluster by DBSCAN. Experiment results showed that the method played a great role in improving the accuracy of maize purity identification.

Key words Maize, Seed purity, Identification, Clustering, DBSCAN

引言

玉米是中国主要农作物之一, 在食品、饲料等行业中应用广泛, 玉米种子纯度直接关系到作物质量与产量^[1]。种子纯度是指本品种的种子数占供检样品种子数的百分比。生产上引起杂交玉米种子不纯的原因主要有: 母本田间去雄不彻底、不及时, 从而产生大量的自交籽粒, 即杂交种中掺杂母本; 亲本不纯且去杂不彻底; 杂交种加工时人为或机械混杂;

隔离条件不好, 造成串粉^[2]。近年来, 机器视觉技术以其客观、无损、快速的特点在农作物检测中被广泛应用, 研究玉米种子的机器视觉纯度检测技术代替手工检测意义重大。目前, 国内外应用机器视觉技术对玉米种子外观特征进行了相关研究。国外学者利用形状参数分析玉米种子轮廓特征进行破损识别与品质检测, 利用玉米种子色度鉴别玉米种子颜色^[3-6]。国内学者运用图像处理技术对玉米种子侧面形状特征^[7-10]、HIS 颜色特征^[9-12]、RGB 颜色特

收稿日期: 2011-09-24 修回日期: 2011-11-29

* 山东省博士后创新基金资助项目(200903031)

作者简介: 刘双喜, 讲师, 主要从事图像处理和模式识别研究, E-mail: lentree@sda.u.edu.cn

通讯作者: 王金星, 副教授, 主要从事精密工程与科学仪器研究, E-mail: jinxingw@163.com

征^[9~14]及玉米种子白色部分(胚部)与黄色部分(冠部)的面积^[10~11]进行了相关研究以更好地完成对玉米种子质量、品种及纯度的鉴别。

国内外对玉米种子形态特征的研究已经达到了较好的识别效果,但对玉米种子纯度的识别研究很少且图像识别所依据的主要是玉米种子侧面特征。本文针对 3 种常用的玉米种子进行研究时发现,冠部核心颜色参数对玉米种子纯度识别具有显著作用,利用最远优先遍历优化 DBSCAN 密度聚类算法进行纯度识别。

1 玉米种子冠部图像特征提取与分析

1.1 图像采集与图像处理

通过工业 CCD 相机及玉米种子摆放装置^[15]采集玉米种子冠部原始彩色图像,图像识别的基本依据是被识别图像中所包含的特征信息。根据采集的玉米种子冠部彩色图像,进行预处理,利用 RGB 颜色模型中 R 分量对图像进行背景分割,同时采用轮廓标记法将每粒种子以彩色图像显示提取出来。以单个玉米种子冠部区域形心为中心,10 个像素为半径,获得单粒玉米种子冠部核心区域,并同时编辑标号,便于纯度识别结果验证^[12]。以品种郑单 958 为例,玉米种子冠部原始图像如图 1 所示。

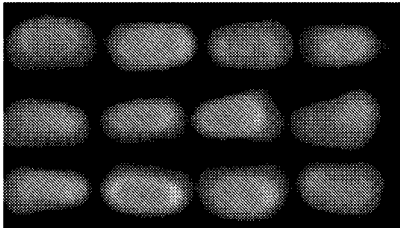


图 1 郑单 958 冠部原始图像
Fig. 1 Original image of Zhengdan 958

预处理结果图像如图 2 所示,序号标记图像与单粒玉米种子冠部核心图像分别如图 3、4 所示。

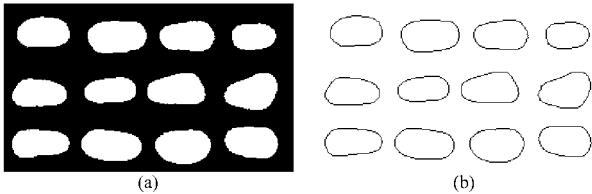


图 2 预处理图像
Fig. 2 Preprocessed image
(a) 背景分割图像 (b) 轮廓标记图像

1.2 颜色特征提取及特征向量选择

玉米种子图像的特征信息主要有形状特征和颜色特征。实际生产过程中,同一穗种子形状特征具有一定差异。该研究选取 RGB、HSI 及 Lab 3 个颜色特征模型,提取玉米种子冠部核心区域的 9 个颜

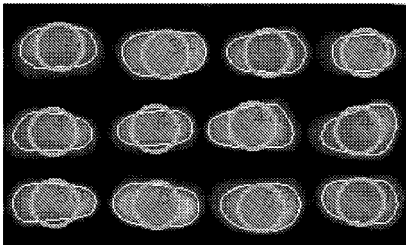


图 3 序号标记图像
Fig. 3 Number mark image

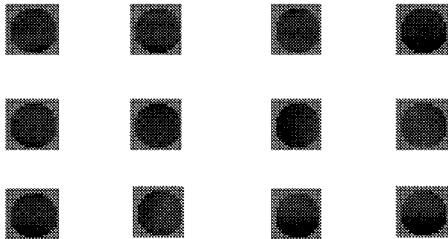


图 4 玉米冠部核心图像
Fig. 4 Maize crown core area image

色特征 R (红)、 G (绿)、 B (蓝)、 H (色调)、 S (饱和度)、 I (亮度)、 L (亮度)、 a (红绿色品)和 b (黄蓝色品)。特征值分别是每粒玉米种子冠部核心区域所有像素点的平均值。

为降低特征向量维数,提高识别速率,对上述 9 个特征进行分析,因 I 与 L 表征同样的特征参数,合并 I 与 L 。以郑单 958 和鲁单 981 样本为例,对样本统计计算得到各参数正态分布函数的均值、均方差和变异系数如表 1 所示。

表 1 玉米颜色特征参数统计数据
Tab. 1 Statistical data of color feature parameters of maize seeds

特征	郑单 958			鲁单 981		
	均值	均方差	变异系数	均值	均方差	变异系数
H	64.508 8	6.106 0	0.094 7	70.544 0	15.660 0	0.222 0
S	0.353 5	0.062 2	0.175 9	0.219 6	0.101 8	0.463 8
I	109.820 3	8.631 4	0.078 6	139.476 7	15.387 7	0.110 3
B	71.246 8	11.167 1	0.156 7	110.607 3	24.018 3	0.217 1
G	130.668 4	10.742 2	0.082 2	154.910 8	15.324 7	0.098 9
R	126.246 5	7.012 7	0.055 5	151.196 9	10.130 3	0.067 0
a	89.042 9	0.875 6	0.009 8	90.662 3	0.896 6	0.009 9
b	174.583 1	2.207 8	0.012 6	169.780 3	3.805 5	0.022 4

其中变异系数说明样本内均值差别程度,对于任一个数据集而言,边界点附近数据的密度分布相对离散,其具有较大的变异系数。标准偏差表征特征值与均值离散的程度,即杂粒与正常杂交粒的偏离程度^[16]。从表 1 可以看出, H 、 S 、 B 的变异系数比较大,即杂粒与正常杂交粒偏离程度比较大,对识别正确率的贡献大。为降低特征冗余,提高识别率,选

取 H 、 S 、 B 作为识别特征向量。

分别对样本郑单 958、农大 108、鲁单 981 的特征向量 (H, S, B) 作三维特征散点图 (S 取值范围为 $0 \sim 1$, 为体现其变化规律, 将其放大 100 倍后显示在图中^[12]), 以品种郑单 958 为例, 均呈现出如图 5 所示的特征参数分布特点。从图 5 可见, 整体上特征点分布呈条形流体状, 存在个别异常样本偏离, 且特征向量点的分布具有密度差异。

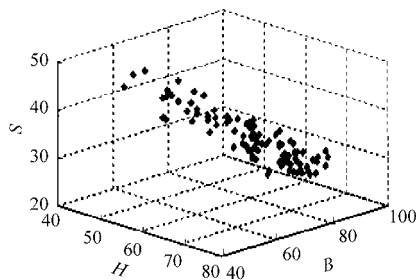


图 5 特征向量三维散点图

Fig. 5 3-D diagram of feature vector

2 基于最远优先遍历优化 DBSCAN 算法的玉米纯度识别

依据样本中特征向量点的三维密度分布特性, 采用 DBSCAN 密度聚类方法进行纯度识别。鉴于数据点密度不均匀, 利用最远优先遍历方法^[17]优化数据, 将整个数据空间中密度差异特征向量点划分为边缘异常散点及高密度 2 个区域, 剔除边缘异常散点, 对高密度区域进行 DBSCAN 密度聚类, 最后合并聚类结果。

优化 DBSCAN 算法分为 4 个基本步骤:

(1) 输入聚类数据集 P , 含有 n 个样本点。

(2) 采用最远优先遍历算法优化数据。根据得到的初始聚类中心 Z_1 和 Z_2 , 对应数据集划分为 2 个子集 $Z_1 \in G_1^{(t)}$ 和 $Z_2 \in G_2^{(t)}$, 剔除边缘异常散点集 $G_2^{(t)}$ 。

(3) 对高密度局部数据集进行 DBSCAN 聚类。

(4) 合并各个局部数据集的聚类结果, 得到最终的聚类结果。

2.1 样本数据优化

在识别过程中, 对于密度不均匀样本点 (H, S, B) , 采用全局统一变量, 聚类效果不佳。因此在执行 DBSCAN 聚类前, 首先要剔除数据中明显异常的数据。FFT (farthest first traversal) 是一种最远优先遍历方法, 它可以快速地集中数据, 集中相似样本, 剔除异常样本。主要参数为聚类个数和初始中心点, 纯度识别是 2 类问题, 即只需将样本分为正常杂交粒与杂粒。为了更精确地取得聚类的结果, 舍去初始中心点的随机性, 设定样本均值点作为初始聚类中心。具体算法如下:

(1) 令聚类数为 2, 计算样本均值点 (H, S, B) , 令其为初始聚类中心 Z_1 , $G_1^{(t)}$ 是以 $Z_1^{(t)}$ 为聚类中心的样本集。

(2) 将已有特征向量二维点 (H, S, B) 表示为 $(P_1, P_2, \dots, P_n)$, 计算各样本点到初始聚类中心的距离, 采用欧式距离

$$d(P_j, Z_1) = \sqrt{\sum_{i=1}^n (P_{ij} - Z_{1i})^2} \quad (i = 1, 2; j = 1, 2, \dots, n) \quad (1)$$

式中 P_{ij} ——第 j 点的第 i 维坐标

Z_{1i} ——初始聚类中心点的第 i 维坐标

(3) 取距离初始聚类中心最远的 P_j 点作为另一聚类中心 Z_2 , 令 $G_2^{(t)}$ 是以 $Z_2^{(t)}$ 为聚类中心的边缘异常散点集。

(4) 对 P_n 中剩余的每个对象 P_j , 分别计算其到 Z_1, Z_2 的距离 d_{j1}, d_{j2} 并比较, 经 t 次迭代, 将得到的所有样本集 P 中的点依次分配到与其最相似的聚类中, 直到聚类结束, 聚类结果如图 6 所示。

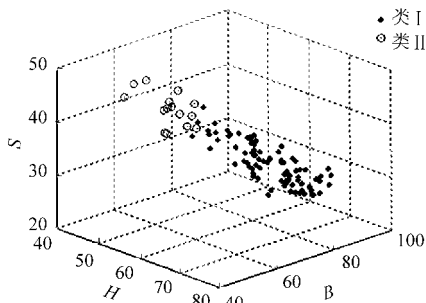


图 6 FFT 算法优化数据

Fig. 6 Data optimization of FFT algorithm

2.2 DBSCAN 密度聚类

2.2.1 DBSCAN 算法

DBSCAN 算法是一种颇具代表性的基于密度的聚类分析算法, 它具有可以发现任意形状的簇和有效屏蔽噪声数据干扰的优点^[18]。这类算法的核心思想是: 对于构成簇的每个对象, 其 Eps 邻域包含的对象个数必须不小于一个给定值 (D_{MinPts}), 即其邻域的密度必须不小于某个阈值。

DBSCAN 算法的聚类过程基于如下事实, 即一个聚类可以由其中的任何核心对象唯一确定, 等价地, 可以表述为:

(1) 给定任一满足核心对象条件的数据对象 P , 数据库 G_1 中所有从 P 密度可达到的数据对象 o 所组成的集合 $\{o | o > G_{1p}\}$ 构成一个完整的聚类 C , 且 $P \in C$ 。

(2) 给定一个聚类 C 和它的任一核心对象 P , C 等价于集合 $\{o | o > G_{1p}\}$ 。

数据对象集合 G_1 即为数据优化后局部高密度

特征向量集 (H,S,B) 。

2.2.2 参数确定

DBSCAN 要求用户指定 Eps 值(为减少计算量, D_{MinPts} 设定为 $4^{[19]}$)。为了确定 Eps 值,DBSCAN 计算所有数据对象与它的第 D_{Minpts} 个最邻近的对象之间的距离,并将结果按距离排序,产生 k-dist 图,用图形化的方法显示,然后确定此值。

k-dist 图中的横坐标表示为对应与某一 k-dist 距离的数据对象的个数,纵坐标表示数据对象 (H,S,B) 与它的第 4 个最邻近的对象间的距离;k-dist 图的作用就是确定最合适的 Eps 值,即下限密度。数据对象与它的第 k 个最近的对象间的距离采用降序排列,因此 k-dist 图也称为排序化 k-dist 图,如图 7 所示,观察图中曲线的平缓部分,将处在 A 位置的值设定为 Eps。

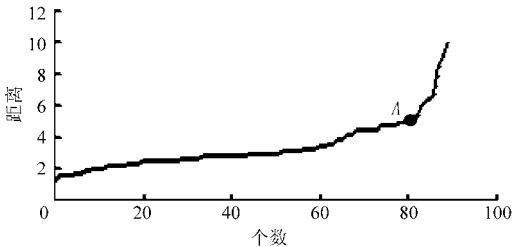


图 7 k-dist 图

Fig. 7 k-dist diagram

选取 D_{MinPts} 为 4,Eps 值为 5.2,DBSCAN 算法描述如下:

- (1) 选择数据集 $G_1^{(t)}$ 中任意一个不属于任何聚类且满足核条件的对象 P ,创建一个新的聚类。
- (2) 根据该聚类中的核对象,循环收集密度可达的核对象加入该聚类,直到没有新的核对象加入为止。
- (3) 若不存在不在任何聚类内的核对象则结束,否则执行步骤(1)。

在此过程中,记录那些与分区边界距离不大于 Eps 噪声点。它们可能是全局中某个聚类的边界点或者某一被分割的小聚类中的点。

2.3 局部聚类合并

根据聚类过程,将 FFT 数据优化中所剔除的边缘异常散点作为杂粒类,同时将局部聚类过程中边缘噪音点和小类点同时归到杂粒类。以郑单 958 为例,没有出现小类点,只存在噪声点。完成 DBSCAN 聚类后合并聚类结果如图 8 所示。

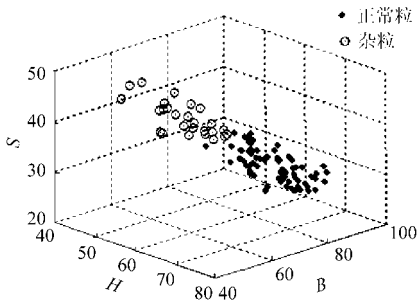


图 8 合并聚类结果

Fig. 8 Clustering results

2.4 实验结果与分析

选用郑单 958(金娃娃),农大 108(金色农华)及鲁单 981(山东农科院)各 400 粒对上述纯度识别算法进行实验验证并同时与 DBSCAN 和 K-均值 2 种传统算法作对比,结果如表 2 所示。从表中可以看出,优化后的 DBSCAN 算法识别率高于其他 2 种传统聚类方法,且纯度正确识别率达到 93.3%。基本满足种子流通现场玉米纯度识别的要求。

表 2 玉米纯度识别结果

Tab. 2 Maize purity identification results				%
玉米品种	正确识别率			
	DBSCAN	K-均值	优化 DBSCAN	
郑单 958	58.3	83.3	96.7	
农大 108	70.8	75.0	94.4	
鲁单 981	66.7	79.2	93.3	

由于冠部核心区域设定是人工多次实验经验设定,从某种程度上影响了玉米种子冠部核心特征的表达。另外因种子由人工摆放,冠部高低存在差距,从而增加了外部光源对颜色特征的影响。图像采集系统光源的均衡性也是影响识别结果的重要因素。

3 结论

(1) 建立以 H,S,B 作为特征向量的三维空间模型,针对密度分布差异,提出基于最远优先遍历优化的 DBSCAN 玉米纯度识别算法。该算法通过 FFT 优化数据,将边缘异常散点剔除,经 DBSCAN 对高密度区域进行局部聚类,最后合并聚类区域,得到纯度识别结果。

(2) 对郑单 958、农大 108、鲁单 981 共 3 个品种进行纯度识别实验,结果表明,优化 DBSCAN 算法优于其他 2 种传统聚类算法,正确识别率可达到 93.3%。

参 考 文 献

1 刘燕德,应义斌,成芳. 机器视觉技术在种子纯度检验中的应用[J]. 农业机械学报,2003,34(5):161~163.
Liu Yande, Ying Yibin, Cheng Fang. Research of machine vision in purity inspection of seed[J]. Transactions of the Chinese

- Society for Agricultural Machinery, 2003, 34(5): 161 ~ 163. (in Chinese)
- 2 李晓辉,李新海,李文华,等. SSR 标记技术在玉米杂交种种子纯度测定中的应用[J]. 作物学报,2003,29(1):63 ~ 68.
Li Xiaohui, Li Xinhai, Li Wenhua, et al. Application of SSR markers in hybrid seed purity test of maize [J]. Acta Agronomica Sinica, 2003, 29(1):63 ~ 68. (in Chinese)
- 3 Liao K, Paulsen M R, Reid J F, et al. Corn kernel breakage classification by machine vision using a neural network classifier [J]. Transactions of the ASAE, 1993, 36(6):1 949 ~ 1 953.
- 4 Ni B, Paulsen M R, Reid J F. Corn kernel crown shape identification using image processing [J]. Transactions of the ASAE, 1997, 40(3): 833 ~ 838.
- 5 Paliwal J, Shashidhar N S, Jayas D S. Grain kernel identification using kernel signature [J]. Transactions of the ASAE, 1999, 42(6):1 921 ~ 1 924.
- 6 Liu J, Paulsen M R. Corn whiteness measurement and classification using machine vision [J]. Transactions of the ASAE, 2000, 43(3):757 ~ 763.
- 7 宁纪锋,何东健,杨蜀秦. 玉米种子的尖端和胚部的计算机视觉识别[J]. 农业工程学报,2004,20(3):117 ~ 119.
Ning Jifeng, He Dongjian, Yang Shuqin. Identification of tip cap and germ surface of corn kernel using computer vision [J]. Transactions of the CSAE, 2004, 20(3):117 ~ 119. (in Chinese)
- 8 郝建平,杨锦忠,杜天庆,等. 基于图像处理的玉米品种的种子形态分析及其分类研究[J]. 中国农业科学,2008,41(4):994 ~ 1 002.
Hao Jianping, Yang Jinzhong, Du Tianqing, et al. A study on basic morphologic information and classification of maize cultivars based on seed image process[J]. Scientia Agricultura Sinica, 2008, 41(4): 994 ~ 1 002. (in Chinese)
- 9 程洪,史智兴,么伟,等. 基于支持向量机的玉米品种识别[J]. 农业机械学报,2009,40(3):180 ~ 183.
Cheng Hong, Shi Zhixing, Yao Wei, et al. Corn breed recognition based on support vector machine [J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(3): 180 ~ 183. (in Chinese)
- 10 王玉亮,刘贤喜,苏庆堂,等. 多对象特征提取和优化神经网络的玉米种子品种识别[J]. 农业工程学报, 2010, 26(6):199 ~ 204.
Wang Yuliang, Liu Xianxi, Su Qingtang, et al. Maize seeds varieties identification based on multi-object feature extraction and optimized neural network[J]. Transactions of the CSAE, 2010, 26(6): 199 ~ 204. (in Chinese)
- 11 史智兴,程洪,李江涛,等. 图像处理识别玉米品种的特征参数研究[J]. 农业工程学报,2008,24(6):193 ~ 195.
Shi Zhixing, Cheng Hong, Li Jiangtao, et al. Characteristic parameters to identify varieties of corn seeds by image processing [J]. Transactions of the CSAE, 2008, 24(6): 193 ~ 195. (in Chinese)
- 12 闫小梅,刘双喜,张春庆,等. 基于颜色特征的玉米种子纯度识别[J]. 农业工程学报,2010,26(增刊1):46 ~ 50.
Yan Xiaomei, Liu Shuangxi, Zhang Chunqing, et al. Purity identification of maize seed based on color characteristics [J]. Transactions of the CSAE, 2010, 26(Supp.1): 46 ~ 50. (in Chinese)
- 13 宋鹏,吴科斌,张俊雄,等. 基于计算机视觉的玉米单倍体自动分选系统 [J]. 农业机械学报, 2010, 41(增刊): 249 ~ 252.
Song Peng, Wu Kebin, Zhang Junxiong, et al. Sorting system of maize haploid kernels based on computer vision [J]. Transactions of the Chinese Society for Agricultural Machinery, 2010, 41(Supp.):249 ~ 252. (in Chinese)
- 14 权龙哲,祝荣欣,雷博,等. 基于 K-L 变换与 LS-SVM 的玉米品种识别方法[J]. 农业机械学报,2010,41(4):168 ~ 172.
Quan Longzhe, Zhu Rongxin, Lei Pu, et al. Recognition method of maize cultivars based on K-L transform and LS-SVM [J]. Transactions of the Chinese Society for Agricultural Machinery, 2010, 41(4): 168 ~ 172. (in Chinese)
- 15 山东农业大学. 一种孔板式玉米种子摆放装置:中国,ZL201020530776. 1[P]. 2011-05-11.
- 16 薛丽香,邱保志. 基于变异系数的边界点检测算法[J]. 模式识别与人工智能,2009,22(5): 799 ~ 802.
Xue Lixiang, Qiu Baozhi. Boundary points detection algorithm based on coefficient of variation [J]. PR & AI, 2009, 22(5):799 ~ 802. (in Chinese)
- 17 Gonzalez T F. Clustering to minimize the maximum intercluster distance [J]. Theoretical Computer Science, 1985, 38: 293 ~ 306.
- 18 冯少荣,肖文俊. DBSCAN 聚类算法的研究与改进[J]. 中国矿业大学学报, 2008,37(1):106 ~ 111.
Feng Shaorong, Xiao Wenjun. An improved DBSCAN clustering algorithm [J]. Journal of China University of Mining & Technology, 2008, 37(1): 106 ~ 111. (in Chinese)
- 19 Yue S H, Li P, Guo J D, et al. A statistical information based clustering approach in distance space [J]. Journal of Zhejiang University Science, 2005,6A(1):71 ~ 78.