图文信息技术

基于 SVM 的高通量 dPCR 基因芯片荧光图像分类研究

刘丽,孙刘杰,王文举

(上海理工大学,上海 200093)

摘要:目的 为了实现高通量 dPCR 基因芯片荧光图像的亮点分类与计数,提出一种基于支持向量机 (SVM)的荧光图像分类与计数方法。方法 首先对荧光图像进行去嗓、对比度增强等图像预处理,对 预处理后荧光图像进行亮点区域提取标注,去除背景与暗点的冗余信息,利用方向梯度直方图 (Histogram of Oriented Gradient, HOG)提取鉴别特征,计算合并所有样本的亮点特征得到 HOG 特征向 量,根据已得到的 HOG 特征向量创建一个线性 SVM 分类器,利用训练好的 SVM 分类器对荧光图像亮 点进行分类与计数。结果 对比传统算法,文中算法具有较高的分类识别精度,平均准确率高达 98%以 上,可以很好地实现荧光图像亮点分类与计数。结论 在有限的小样本标注数据下,文中算法具有良好 的分类性能,能够有效识别荧光图像中的亮点,对其他荧光图像分类研究也具有一定参考价值。 关键词:dPCR;支持向量机;方向梯度直方图;荧光图像;亮点分类 中图分类号:TP751 文献标识码:A 文章编号:1001-3563(2020)19-0223-07 DOI:10.19554/j.enki.1001-3563.2020.19.032

Classification of Fluorescent Images in High-throughput dPCR Gene Chips Based on SVM

LIU Li, SUN Liu-jie, WANG Wen-ju

(University of Shanghai for Science and Technology, Shanghai 200093, China)

ABSTRACT: The work aims to propose a fluorescence image classification and counting method based on support vector machine (SVM) to achieve the classification and counting of bright spots in high-throughput dPCR gene chip fluorescence image. Firstly, image preprocessing such as denoising and contrast enhancement was performed on the fluorescent image, and bright spot region was extracted annotated on the preprocessed fluorescent image to remove redundant information of background and dark points. A histogram of orientation gradient (HOG) was used to extract discriminative features, and the bright spot features of all samples were combined to obtain the HOG feature vector. A linear SVM classifier was created based on the obtained HOG feature vectors. The trained SVM classifier was used to classify and count the bright spots of the fluorescent image. Compared with traditional algorithms, the proposed algorithm had higher classification and counting of bright spots in fluorescent images. With limited small sample annotation data, the algorithm in this paper has good classification performance, can effectively identify bright spots in fluorescent images, and has certain reference value for other fluorescent image classification studies.

收稿日期: 2019-12-16

基金项目:上海市科学技术委员会科研计划(18060502500)

作者简介:刘丽(1996-),女,上海理工大学硕士生,主攻数字图像处理。

通信作者:孙刘杰(1965—),男,博士,上海理工大学教授,主要研究方向为光信息处理技术、印刷机测量与控制技术、 数字印刷防伪技术。

KEY WORDS: dPCR; SVM;HOG; fluorescence image; bright spot classification

数字 PCR (digital Polymerase Chain Reaction, dPCR)在1999年被K.Mullis发现并提出,是可以在 体外模拟体内核酸合成过程的一种核酸绝对定量检 测技术^[1],包括 PCR 扩增和荧光信号分析,也是继实 时荧光定量 PCR 后的一种新型基因芯片。dPCR 基因 芯片凭借其灵敏度高、精确度高、效率高等优点,广 泛应用于病原微生物检测等生物医学临床领域^[2],也 是早期的肿瘤诊断^[3]、胎儿的产前诊断^[4]、基因拷贝 数变异^[5]以及细胞的基因分析表达^[6]等检测的重要手 段之一。在高通量数字 PCR 中检测到的荧光图像呈 密集分布,且具有低亮度、低对比度等特点。对高通 量 dPCR 基因芯片的研究仍在不断探索中,对荧光图 像中微滴点的分析方法还有许多问题仍待解决。

数字 PCR 技术通过分析基因芯片荧光图像阳性 微滴点(亮点)与阴性微滴点的成像信息,达到医学 判断的目的。对于荧光图像的分析方法也在不断涌 现,现今主流的荧光图像分析方法有荧光图像预处 理、荧光图像定位、荧光图像分割、荧光图像微滴数 目统计及阴阳微滴点信息分析等^[7]。

高通量 dPCR 基因芯片通过窄带 LED 作为激发 光源,一次性得到的基因芯片面积较大,但是,在成 像的过程中,荧光图像受外部噪音和光源光照的干扰 较大,最终成像的荧光图像中微滴与背景的亮度明显 不均匀化,对后续的目标分类与计数会造成一定的影 响。目前对于 PCR 荧光微滴图像的分类识别方法有: 基于广义帕累托分布的荧光微滴分类方法[8];通过对 分水岭分割算法改进对荧光微滴识别分类^[9];采用改 进分水岭方法结合半监督最小误差重构分类器对荧 光微球图像进行分割及分类[10];根据测试样本与 SVM 分离超平面的相对距离和特征空间中的训练聚 类中心的相对距离,定义测试样本的相似度的荧光显 微图像分类^[11]等。现有的 PCR 荧光图像亮点分类计 数方法主要根据亮点阈值范围和亮点形状尺寸对荧 光微滴进行分割,对分割后的微滴点进行特征提取以 及分类,当阳性点与阴性点阈值差别较小时,对于亮 点的识别精度与适应性较差,但是对于高通量 dPCR 基因芯片荧光图像的分类识别研究较少。

1 荧光图像亮点分类实现过程

针对上述问题,文中提出了一种基于 SVM 的高 通量 dPCR 基因芯片荧光图像分类方法,对预处理后 的荧光图像,结合 SVM 与 HOG 特征实现荧光图像 的分类与计数。文中研究的荧光图像亮点分类实现首 先对荧光图像进行去噪、对比度增强等图像预处理, 然后将亮点区域进行标注,将标注区域内的特征信息 进行归一化创建 HOG 描述器,通过 HOG 描述器对 已标注的荧光图像亮点区域进行高维特征描述,利用 提取的 HOG 特征进行训练,创建一个线性 SVM 分 类器进行分类识别,测试结果标记为 light 的区域即 分类出的荧光图像亮点。荧光图像亮点分类实现流程 见图 1。



图 1 荧光图像亮点分类流程 Fig.1 Flowchart classification of bright spots in fluorescence image

其中,标准化方法采用 L2 标准化,在对图像进行 HOG 特征提取时,首先将图像划分为尺寸为 3×3 的块(block),将块划分成尺寸为 2×2 的胞元(cell),然后计算统计块与胞元的 9 个梯度方向直方图,并将 块内各像素点梯度大小的投影量作为权值进行累加统计,累加结果即为图像的 HOG 特征向量。最后将 HOG 特征向量输入 SVM 分类器进行训练,在特征向量提取过程中应保持一定的长度,若扭曲了原始候选 区域的目标信息,不利于 SVM 的识别。

选择合理的样本可以提升分类器的分类精度,同时也能大幅提高样本的学习速率。在使用荧光图像亮 点提取 HOG 特征向量训练 SVM 分类器时,对于 SVM 分类器分类精度的影响因素主要有训练样本的选取、 核函数 *K* 的选择以及惩罚权重 *C* 的确定,因此选取 荧光图像训练 SVM 分类器时应注意以下问题:避免 不平衡样本集和荧光图像亮点样本具有代表性。

文中方法具有以下优势:在图像预处理时,通过 改进的二维 Gamma 函数对图像进行动态校正,结合 引导滤波与自适应直方图均衡化达到荧光图像增强 效果;高通量 dPCR 基因芯片荧光图像的亮点部分梯 度信息明显易识别,HOG 特征具有几何和光学转化 不变性,易提取,有利于表现荧光图像亮点的原始特 征;只提取标注的荧光图像的亮点区域进行特征提取 鉴别,减少了暗点与背景等冗余信息,降低了计算复 杂度;对分类的亮点进行自动计数,在一定程度上可 以实现帮助诊断判别;使用 SVM 分类器,使用有限 的少量的标记数据就能实现对高通量 dPCR 荧光图像 亮点分类,分类效果较佳,分类时间较短。

2 相关理论与关键算法

2.1 荧光图像预处理

原始的高通量 dPCR 荧光图像存在对比度低、噪声含量多以及光照不均等问题,直接进行特征提取效 果较差,导致阳性点(即亮点)与阴性点错分机率增 大。在尽量保持不丢失图像隐含细节与图像不失真的 条件下,首先采用中值滤波去除荧光图像中大部分噪 声,减小对比度增强处理效果的影响。然后通过优化 的二维 Gamma 函数对光照分量进行校正,光照分量 值取荧光图像光照均值见式(1),光照分量提取由 引导滤波实现。在光照均匀化的同时会导致图像对比 度降低,因此使用自适应直方图均衡化对荧光图像进 行对比度增强处理。

$$\gamma = \varepsilon \left(\frac{n - L(x, y)}{n} \right) \tag{1}$$

高通量 dPCR 荧光图像属于灰度图像, 不适用于 颜色特征提取算法,由于阳性点(即亮点)与阴性点 在不同通道下的阈值区别较小,使用阈值算法时会导 致大面积错分情况,因此并不适用高通量 dPCR 荧光 图像。在荧光图像中,使用 Sobel 算子对预处理后的 荧光图像边缘梯度进行计算,见图2。阳性点与阴性 点的梯度区域存在明显差异,高通量 dPCR 荧光图像 的局部目标的表象和形状能够被梯度或边缘的方向 密度分布很好地描述,即梯度主要存在于边缘地区, 阴性点与阳性点边缘处之间的梯度跃变明显。由于 HOG 特征是在图像的局部单元上操作,因此,高通 量dPCR荧光图像的几何和光学的形变都能保持很好 的不变性。再通过空域抽样与方向抽样和局部的光学 归一化, 使得高通量 dPCR 荧光图像中残存的噪音被 忽略,对于检测结果影响效果较小。可见 HOG 特征 能够突出表现高通量 dPCR 荧光图像阳性点与阴性点 之间的区别,阳性点特征突出明显。



图 2 梯度图 Fig.2 Gradient graph

2.2 方向梯度直方图特征

方向梯度直方图(Histogram of Oriented Gradient, HOG)特征^[12]是通过特征描述符提取图像相关信息, 丢弃无关信息来简化图像的表示方法,被广泛应用于 图像识别^[13]中。HOG 特征提取原理见图 3。



图 3 HOG 特征提取原理 Fig.3 HOG feature extraction principle

1)图像归一化。高通量 dPCR 荧光图像与普通 图像不同,在多数情况下暗部的光照分量较小,图像 纹理表现性强,通过进行伽马(Gamma)矫正,光照 分量值取荧光图像光照均值,调节图像对比度,减少 光照不均匀化、局部细节阴影对图像的影响,使得在 进行 HOG 特征提取时对光照的抗鲁棒性增强。

$$G = 0.3R + 0.59R + 0.11B \tag{2}$$

$$I(x, y) = I(x, y)^{c_{Gamma}}$$
(3)

2) 计算梯度图。首先使用内核为 1 的 Sobel 算 子计算水平梯度 g_x 和垂直梯度 g_y , 然后再计算 x 和 y方向梯度的合梯度,包括幅值 $\nabla g(x,y)$ 和方向 θ 。求 梯度操作滤除高通量 dPCR 荧光图像中无关的、变化 剧烈的数据,弱化对荧光图像亮点的影响。此操作不 仅能够捕获高通量 dPCR 荧光图像中阴性点与阳性点 的轮廓和纹理信息,还能进一步弱化光照的影响。梯 度方向取绝对值,得到的角度范围为[0,180°]。

$$g_{x} = I(x+1, y) - I(x-1, y)$$
(4)

$$g_{y} = I(x, y+1) - I(x, y-1)$$
(5)

$$\Delta g(x, y) = \sqrt{g_x^2 + g_y^2}$$
(6)

$$\theta = \arctan\left(\frac{g_y}{g_x}\right) \tag{7}$$

3) 计算梯度直方图。对 9个直方图通道中的各像

素点梯度大小作为投影,依次将各像素点梯度大小的 投影量作为权值累加到9个梯度方向直方图,见图4。



图 4 梯度方向 Fig.4 Gradient direction

4)块内归一化梯度直方图。由于局部光照与背景-阴性点-阳性点的变化,使得高通量 dPCR 荧光图像梯度强度变化范围较大,整体光照非常敏感,因此需将块内梯度强度做"归一化"处理,进一步弱化对光照、阴影和边缘对高通量 dPCR 荧光图像亮点的影响。



图 5 块与胞元图 Fig.5 Block and cell graph



图 6 块内归一化 Fig.6 Intra-block normalization

5) HOG 特征向量整合。把提取的高通量 dPCR 荧光图像亮点区域的 HOG 特征输入 SVM 分类 器中,寻找一个最优超平面作为决策函数,得到分类 模型。

2.3 支持向量机

支持向量机(SVM)^[14]是以最优化算法为基础

的机器学习方法,也是核函数方法。由于 SVM 具有 精度高和处理高维数据的能力,从而被广泛应用于分 类问题。荧光图像中只含有阳性点(即亮点)与阴性 点 2 类,是一个较好的二值分类样本。将使用 HOG 特征提取的亮点区域特征向量输入 SVM 分类训练 器,得到 SVM 训练模型。再输入荧光图像测试图片, SVM 分类模型在荧光图像中寻找符合特征的亮点输 出并标记。

为解决 SVM 算法误分或漏分问题, CORTES 和 VAPNIK 等^[15]提出了软边缘最优超平面。由于荧光图 像的阳性点(即亮点)与阴性点在不同通道下的亮度 差异对比度相差较小,为了防止预处理后图像仍出现 错分,引入松弛变量 *ξ*、惩罚权重 *C* 和核函数^[16]。在 对亮点区域的训练样本中,最优超平面见图 7。



图 7 支持向量与距离 Fig.7 Support vector and distance

3 实验结果及分析

为了验证文中提出的高通量 dPCR 基因芯片荧光 图像分类方法的有效性,通过由 CCD 相机拍摄获取 的高通量 dPCR 基因芯片荧光图像, 仿真实验在 Python 2.7 平台上实现, SVM 分类器的核函数是用来 解决数据线性不可分而提出的,把数据从源空间映射 到目标空间(线性可分空间),对数据中存在的噪声 有着较好的抗干扰能力,因此文中选择线性核函数, 惩罚权重为 C=5, 对于荧光图像亮点分类的结果计数 功能主要通过对函数 cv2.rectangle()的循环统计实 现。由于不同通道的荧光图像亮点特征信息量不同, 选取 ROX, CY5 以及 FAM 3 个通道的荧光图像进行 验证,为了进行实验结果对比,选取基于分水岭识别 分类算法一同进行测试。分别为 ROX 通道的荧光图 像、CY5 通道的荧光图像以及 FAM 通道的荧光图像, 3 幅图像尺寸均为 521×521。ROX 通道的荧光图像 亮点分类结果图像见图 8, CY5 通道的荧光图像亮点 分类结果图像见图 9, FAM 通道的荧光图像亮点分类 结果图像见图 10。



a 原图

b SVM 算法亮点分类

c 分水岭算法分类





a 原图

bSVM 算法亮点分类

c 分水岭算法分类

图 9 CY5 通道的荧光图像亮点分类结果 Fig.9 Classification results of bright spots in fluorescence images of CY5 channels



a 原图

b SVM 算法亮点分类

c 分水岭算法分类

图 10 FAM 通道的荧光图像亮点分类结果 Fig.10 Classification results of bright spots in fluorescence images of FAM channels

将 ROX 通道的原图与 SVM 算法亮点分类图结 果进行对比发现, SVM 算法基本识别分类荧光图像 亮点,统计亮点数目为 37,与原图数目保持一致; 分水岭算法处理的荧光图像亮点识别结果与原图对 比结果表明,分水岭算法分类在图 8c 中圆圈圈出位 置,出现漏检情况,还存在过分割问题,引起此问题 的原因为局部存在许多极值点,分割效果较差。

将 CY5 通道的原图与 SVM 算法亮点分类图结果 进行对比发现, SVM 算法基本识别分类荧光图像亮 点,统计亮点数目为 6,与原图数目保持一致;分水 岭算法处理荧光图像分类结果图中对于荧光图像亮 暗点没有识别分类出来,效果较差。 将 FAM 通道的原图与 SVM 算法亮点分类图结 果进行对比发现, SVM 算法基本识别分类荧光图像 亮点,统计亮点数目为 41,与原图数目保持一致; 分水岭算法处理的荧光图像亮点识别结果与原图对 比结果表明,分水岭算法分类没有识别分类出图 10c 中圆圈圈出位置的亮点,出现漏检情况,且分割后荧 光图像亮点由单目标变为多目标,像素间发生断裂。

与 SVM 方法相比, 分水岭算法是其区域增长的 种子像素根据图像梯度值进行选择, 选择梯度极小值 点作为增长起点, 不同的梯度极小值点的增长终点相 同, 当增长终点相同时, 用边界线进行隔断, 形成分 水岭线。在分水岭分割算法当中, 像素梯度大小是分 割的主要依赖特征。对比图 8、图 9 与图 10 可发现, 检测的亮点在图 8 和图 10 中偏少, 而在图 9 中检测 到的亮点偏多。主要是因为 CY5 通道荧光图像阴性 点、阳性点以及背景的像素梯度值较为接近, 在分水 岭分割过程中将阴性点也作为分割对象进行分割, 导 致图 9 中检测亮点偏多。

通过上述 3 个通道的高通量 dPCR 基因芯片荧光 图像亮点分类图像结果表明, 文中算法能够适用于荧 光图像亮点分类, 且较基于分水岭识别分类算法, 分 类结果更为理想。由于荧光图像亮点为多目标共存, 不同的通道的 HOG 特征也存在差异性, 为了评估算 法的识别性能, 采用识别概率 *P*_d、虚警概率 *F*_d进行 度量。为达到评估的有效性, 对 CY5 通道选取 240 幅荧光图像进行测试。

$$P_{\rm d} = N_{\rm R} / N \tag{8}$$

$$F_{\rm d} = N_{\rm F} / \left(N_{\rm R} + N_{\rm F} \right) \tag{9}$$

式中: N_R 为荧光图像中正确识别的亮点数量; N 为荧光图像中的实际亮点数量; N_F 为荧光图像中错 误识别的非亮点数量。识别概率与虚警概率的统计见 表 1。

表 1 荧光图像 CY5 通道亮点分类结果 Tab.1 Classification results of bright spots in fluorescence images of CY5 channel

$N_{ m R}$	$N_{\rm F}$	Ν	$P_{\rm d}$	$F_{\rm d}$
11228	98	11396	0.9852	0.0086

由表 1 可以看出,在 CY5 通道的识别率和虚警 率中,对于 CY5 通道的高通量 dPCR 基因芯片荧光 图像亮点的分类精确度较好,总体荧光图像亮点分类 精度较为平均,分类识别准确率高达 98%以上,表明 文中方法在荧光图像中仍能保持较高的识别率与较 低的虚警率,无明显起伏变化,具有较好的分类性能。

4 结语

为了实现高通量 dPCR 基因芯片荧光图像的亮点 分类与计数,提出了一种基于支持向量机(SVM)的 荧光图像分类方法。通过对荧光图像样本数据库进行 学习训练,实现了基于 SVM 分类的荧光图像亮点分 类,算法实验结果表明,该算法对在不同通道下的荧 光图像中仍能保持较高的识别率与较低的虚警率,具 有良好的分类性能,对其他荧光图像分类研究也具有 一定参考价值。

参考文献:

- VOGELSTEIN B, KINZLER K W. Digital PCR[J]. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(16): 9236-9241.
- [2] FRAISSE A, COUDRAY-MEUNIER C, MARTIN-LATIL S, et al. Digital RT-PCR Method for Hepatitis a Virus and Norovirus Quantification in Soft Berries[J]. International Journal of Food Microbiology, 2016, 243: 36–45.
- [3] SEFRIOUI D, BEAUSSIRE L, CLATOT F, et al. Heparinase Enables Reliable Quantification of Circulating Tumor DNA from Heparinized Plasma Samples by Droplet Digital PCR[J]. Clinica Chimica Acta, International Journal of Clinical Chemistry, 2017, 472: 75–79.
- [4] RAGNI M V. Prenatal Diagnosis by Droplet Digital PCR[J]. Blood, 2017, 130(3): 240—241.
- [5] BELRRADER P, TANNER S C, REGAN J F, et al. Droplet Digital PCR Measurement of HER2 Copy Number Alteration in Formalin-fixed Paraffin-embedded Breast Carcinoma Tissue[J]. Clinical Chemistry, 2013, 59(6): 991—994.
- [6] KUMARESAN P, YANG C J, CRONIER S A, et al. High-throughput Single Copy DNA Amplification and Cell Analysis in Engineered Nanoliter Droplets[J]. Analytical Chemistry, 2008, 80(10): 3522–3529.
- [7] 周淑芳, 苟彤, 方伟波, 等. 基于微流控芯片的数字 PCR 荧光图像分析方法[J]. 生命科学仪器, 2018, 16(1): 27—31.
 ZHOU Shu-fang, GOU Tong, FANG Wei-bo, et al. Digital PCR Fluorescence Image Analysis Method Based on Microfluidic Chip[J]. Life Science, 2018, 16(1): 27—31.
- [8] 刘聪,董文飞,张涛,等. 微滴式数字 PCR 中低浓度 荧光微滴分类[J]. 光学精密工程,2018,26(3):647—653.
 LIU Chong, DONG Wen-fei, ZHNAG Tao, et al. Classification of Low-concentration Fluorescent Droplets In Droplet Digital PCR[J]. Optical Precision Engineering, 2018, 26(3): 647—653.
- [9] 刘聪,董文飞,蒋克明,等.基于改进分水岭分割算法的致密荧光微滴识别[J].中国光学,2019,12(4):783—790.

LIU Chong, DONG Wen-fei, JIANG Ke-ming, et al.

Dense Fluorescent Droplet Recognition Based on Improved Watershed Segmentation Algorithm[J]. China Optics, 2019, 12(4): 783-790.

- [10] 黄鸿,金莹莹,李政英,等.基于分水岭及半监督最 小误差重构的荧光微球分割及分类方法[J].中国激 光,2018,45(3):136—142.
 HUANG Hong, JIN Ying-ying, LI Zheng-ying, et al. Segmentation and Classification of Fluorescent Microspheres Based on Watershed and Semi-supervised Minimum Error Reconstruction[J]. China Laser, 2018, 45(3):136—142.
- [11] LIN D, LIN Z, SOTHIHARAN S, et al. An SVM Based Scoring Evaluation System for Fluorescence Microscopic Image Classification[C]// 2015 IEEE International Conference on Digital Signal Processing(DSP), IEEE, 2015: 534—547.
- [12] DALAL N, TRIGGS B. Histograms of Oriented Gradients for Human Detection[C]// 2005 IEEE Computer

Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE, 2005.

- [13] GAO Y, SHI M, TAO D, et al. Database Saliency for Fast Image Retrieval[J]. IEEE Transactions on Multimedia, 2015, 17(3): 359–369.
- [14] PHAN A V, NGUYEN M L, BUI L T. Feature Weighting and SVM Parameters Optimization Based on Genetic Algorithms for Classification Problems[J]. Applied Intelligence, 2016, 46(2): 455–469.
- [15] CORTES C, VAPNIK V. Support Vector Networks[J]. Machine Learning, 1995, 20(3): 273–297.
- [16] 张迪飞,张金锁,姚克明,等.基于 SVM 分类的红 外舰船目标识别[J]. 红外与激光工程, 2016, 45(1): 1—6.
 ZHANG Di-fei, ZHANG Jin-suo, YAO Ke-ming, et al.
 - Infrared Ship Target Recognition Based on SVM Classification[J]. Infrared and Laser Engineering, 2016, 45(1): 1—6.