多中心动态聚类算法及对癌症与非癌症彩色手掌图像的分类*)

王宽全[1] 刘 莉[1] 邹向前[1] 张大鹏[1,2]

(哈尔滨工业大学计算机学院 哈尔滨150001)[1] (香港理工大学计算学系 香港九龙)[2]

Multi-central Dynamic Clustering Algorithm Classifying Color Palm Images for Cancer Diagnosis

WANG Kuan-Quan[1] LIU Li[1] WU Xiang-Qian[1] ZHANG Da-Peng[1,2]

(Computer College, Harbin Institute of Technology, Harbin 150001)[1]

(Department of Computation Theory HongKong University of Science and Technology, HongKong)[2]

Abstract Observing palm is one of diagnosis methods in Traditional Chinese Medicine and Holographic Medicine. Generally, the shape, color, ridge and line features of palm are all important for palm diagnosis. As the first attempt for automated palm diagnosis, the color is used and a new statistical feature of color, moment feature, is defined in this paper. Multi-central dynamic clustering algorithm based on our new feature is proposed to recognize cancerous palm images. Applying our approach to the images in the palm database including all kinds of pathological and healthy palm images, the experimental results indicate that it is effective to recognize cancerous palm images and superior over the K-mean algorithm.

Keywords Cancer and noncancer, Clustering algorithm, Automated palm diagnosis, Statistical feature

1 引言

近年来生物特征识别技术(Biometrics)的研究在身份鉴别这一应用领域取得了丰硕的成果,许多实用系统相继在市场出[1,2,4]。但生物特征识别技术应用于诊断人类疾病则是一个完全崭新的应用领域。

掌诊在传统的中医以及新兴的全息医学中都是一种重要的诊断疾病的方法。手掌是人体的一部分,作为一个重要的全息元,它的变化反映了人体各器官的盛衰,是整体的缩影。传统的中医理论对色泽及其变化比较重视,对手掌的颜色及光泽进行观察是掌诊的第一个步骤。就我们黄种人而言,正色可概括为"红黄隐隐,明润含蓄"。一般情况下,手掌呈有光泽的浅红色,中央色稍浅于四周。通过对掌色的观察可以及早发现疾病并对其性质、程度有大致的了解[5,6]。

癌症,即恶性肿瘤,是人类健康的大敌,它的治愈率低,死亡率高。我国人口的主要死亡原因中癌症占第二位(男性)或第三位(女性患癌)。九大重点癌症按其死亡率高低的顺序为:胃癌、食管癌、肝癌、宫颈癌、肺癌、肠癌、白血病、鼻咽癌、乳腺癌。其中,胃癌占全部癌症死亡总数的23%;胃癌、食管癌、肝癌,三者合计为癌症死亡总数的60.45%,近期肺癌有上升的趋势。

根据中医理论,患癌病人的手掌颜色多呈青暗、枯黄,有咖啡色斑点,或白色、暗红色斑点,与正常人(健康人和大部分非癌症病人)的掌色有较大的区别。因此,本文将以癌症与非癌症手掌图像的识别作为通过手掌颜色特征进行疾病诊断的首次尝试。

本文首先定义了特征矩阵;然后给出了癌症与非癌症手掌彩色图像的分类算法,即多中心动态聚类算法;最后是实验结果及结论。

2 特征矩阵

特征选取在模式识别中有着重要的作用[3]。本文选取手掌颜色在RGB空间中三个分量的直方图所对应的3阶原点矩,在HSV空间中色度分量直方图的3阶中心矩作为分类特征。原点矩和中心矩都是描述随机变量分布的数字特征。

2.1 基于RGB空间的特征矩阵

图像 I 在RGB空间中的基于原点矩的特征矩阵如下:

$$F^I_{RGB}=\begin{bmatrix} v^1_R & v^1_G & v^1_B \\ v^2_R & v^2_G & v^2_B \\ v^3_R & v^3_G & v^3_B \end{bmatrix}_I \tag{1}$$

$$v^i_X=EX^i=\sum_{k=0}^{255}k^i p_{Xk} \quad p_{Xk}=\frac{N_{Xk}}{N_I} \tag{2}$$

其中 $i=1,2,3$, $X=R,G,B$, $k=0,1,\cdots,255$, N_{Xk} 是图像 I 中 X 分量值为 k 的像素个数,N_I 是图像 I 的像素总个数。这里 v^i_R、v^i_G、v^i_B 分别是图像 I 的R、G、B三基色直方图的 i 阶原点矩。事实上,特征矩阵 F^I_{RGB} 不仅刻画了图像 I 色彩相对于原点的分布情况,而且还保留了三个分量之间的相互关系。

2.2 基于HSV空间的特征向量

在HSV空间中定义特征向量如下:

$$F^I_H=[\mu^1_H \quad \mu^2_H \quad \mu^3_H]^T_I \tag{3}$$

$$\mu^i_H=E(H-EH)^i=\sum_k(k-EH)^i p_{Hk} \quad p_{Hk}=\frac{N_{Hk}}{N_I} \tag{4}$$

其中 i、k、N_I 如式(2),N_{Hk} 是图像 I 在HSV空间中H分量值为 k 的像素个数,EH 是H分量的数学期望。μ_H 是HSV空间中色度分量直方图的i阶中心矩。一般而言,离散变量的一阶中心矩接近于零,但由于本文实验中的手掌图像分辨率较高,一阶中心矩已偏离了零点,在分类过程中可以起到微调的作用。二阶中心矩是方差,三阶中心矩刻画的是随机变量相对于中心的偏斜度。特征向量 F_H 反映了图像的颜色特征及其关

*)本文的研究工作得到国家863计划项目(863-306-ZD13-06-1)和哈工大交叉学科基金(HIT.MD2001.36)的资助。王宽全 教授,主要研究领域,生物识别技术,计算机网络。刘 莉 硕士生,主要研究领域:生物识别技术。邹向前 博士生,主要研究领域:生物识别技术。张大鹏 教授,博士生导师,主要研究领域:生物识别技术。

于中心的分布情况。

3 癌症与非癌症的手掌图像分类

在众多待分类手掌图像中，有一些手掌因为疾病及病情的原因，颜色特征比较明显。我们可以通过阈值法将其预先识别出来，以减少后续工作量。

3.1 基于 HSV 空间的图像预分类

根据中医理论，癌症病人的手掌发暗，没有光泽，掌色多呈青暗、枯黄，有咖啡色斑点，或白色、暗红色斑点。也就是说，癌症病人手掌颜色的色度、饱和度与正常人有较大的区别，这使得我们可以仅仅通过 HSV 空间中的色度 H、饱和度 S 两个分量即可将部分掌色偏差较大的图像预先识别出来。

经过对大量图像的分析研究，我们发现若用下列两组阈值设定，则可达到较好的效果：

$$I\in\begin{cases}\{cancerous\} & MS\leqslant sc\ and\ MaxH\geqslant hc\\ \{noncancerous\} & MS\geqslant sn\ and\ MaxH\leqslant hn\\ \{uncertained\} & otherwise\end{cases}\qquad(5)$$

其中 I 是待分类的图像，MS 是饱和度的均值，$MaxH$ 是色度的最大值。本文通过对大量图像应用试探法获得的阈值为 $sc=49$，$hc=17$ 和 $sn=60$，$hn=13$。

式(5)中的{*uncertained*}是在预分类过程中无法确定为癌症或非癌症的手掌图像的集合。我们将在下面的算法中对这些手掌图像作进一步的分类。

3.2 用于手掌图像分类的多中心动态聚类算法

在以均值误差平方和最小为准则的分类算法中，K-均值聚类算法是其中较为著名的一种动态聚类算法。但 K-均值算法的一个缺点是只采用均值作为一个类的代表点，这只有当类的自然分布为球状或接近于球状时，才可能有较好的效果[7]。然而样本数据在特征空间的结构是多样的，且真实数据中很少有紧致、界线清晰、比例均匀的类群。

为了提高分类的效果，我们提出了多中心动态聚类算法。该算法与 K-均值算法最大的不同是：一个类的代表不止一个点，即每一类都有一个中心集合，其中包括多个该类的中心。我们用样本与类中心的欧氏距离来度量样本与各类中心集合之间的相似性。这里，我们将通过式(1)、(3)得到的特征矩阵和特征向量分别作为样本特征，将与式(5)中的两组阈值邻近的图像样本作为各类的中心集合。用于手掌分类的多中心聚类算法描述如下：

(1)初始化。

a. 类别数 $C=2$，分别是患癌手掌类 C_A 和非癌手掌类 C_N。

b. 确定各类中心集合。在预处理的结果图像中分别取与两组临界值 $(MS_A, MaxH_A)=(49,17)$，$(MS_N, MaxH_N)=(60,13)$ 相邻的图像，作为各类的中心集合 M_A 和 M_N。两类中心的确定如下：

$I\in M_A$ iff $44\leqslant MS_I\leqslant MS_A\ and\ MaxH_A\leqslant MaxH_I\leqslant 19$

$I\in M_N$ iff $MS_N\leqslant MS_I\leqslant 65\ and\ 11\leqslant MaxH_I\leqslant MaxH_N$

其中，I 是待确定的中心，MS_I、$MaxH_I$ 是 I 的饱和度均值及色度的最大值。两类的中心数目分别记为 N_A 和 N_N。

c. 根据式(1)、(3)分别计算各中心的特征矩阵 M_i 和特征向量 H_i，i 是中心集合 M_A 或 M_N 中的中心样本。

(2)样本的双重最小距离。

a. 选择一个备选样本 y。

b. 计算 y 到各类中心集合的距离 ρ_{M_LY}。

$$\rho_{M_LY}=\min_{k=1\cdots N_L}\rho_{M_{Lk}Y}\qquad(6)$$

其中 $L=A,N$。RGB 空间中

$$\rho_{M_{Lk}Y}=\sqrt{\sum_{i=1}^{3}(Y'_R-M'_{LkR})^2+\sum_{i=1}^{3}(Y'_G-M'_{LkG})^2+\sum_{i=1}^{3}(Y'_B-M'_{LkB})^2}\qquad(7)$$

这里 $Y=F^Y_{RGB}$。M'_{LkX} 是中心集合 M_L 中第 k 个中心样本在 RGB 空间中的 X 分量的 i 阶原点矩，X、i 如式(2)。HSV 空间中

$$\rho_{M_{Lk}Y}=\sqrt{\sum_{i=1}^{3}(Y'_H-H'_{LkH})^2}\qquad(8)$$

这里 $Y=F^Y_H$。H'_{LkH} 是中心集合 M_L 中第 k 个中心样本在 HSV 空间的 H 分量的 i 阶中心矩。

c. 样本分类。$y\in C_i$ iff $\forall\ j(j\neq i)\rho^2_{M_iY}<\rho^2_{M_jY}$

d. 修改第 L 类的中心。

设第 L 类中心集中有 $N_L(L=A,N)$ 个中心，我们对这些中心样本进行编号，定义中心 k 在中心集 L 中的距离：$\forall\ j(j\neq k)\rho_{M^k_LM^j_L}=\max_j\|K-J\|$。这里 $k\in L$。其中 $K=F^k_S$，$J=F^j_S$，$S=RGB,H$。可知中心 j 与中心 k 的距离最远。

设样本 y 与第 L 类中心集中的第 k 个中心距离最短(即 $\|Y-M^k_L\|=\rho_{M_LY}$)，与第 l 个中心距离 $\rho_{M^l_LY}$ 最大。如果 $\rho^2_{M^l_LY}<\rho^2_{M^k_LM^l_L}$，则第 L 类中心集变为 $M_L=\{1,\cdots,N_L,y\}$，否则，第 L 类中心集不变。

(3)如果分类结果不再变化，则结束算法；否则 goto(2)。

显然，上述算法对于聚类集合 $C_L(L=A,N)$ 当中心集合被修正后，准则函数 $J=\sum_{L=A,N}\sum_{y\in C_L}\rho_{M_LY}$ 变小，即多中心动态聚类算法是收敛的。

4 实验结果与结论

我们从手掌数据库中选出了215幅手掌图像，其中包括42幅癌症手掌图像和173幅非癌症手掌图像。用于采集手掌图像的设备是一个自制的暗箱以及数码相机。为了统一光照模型，我们要求被采集人将手伸进暗箱中以隔绝自然光，然后，利用数码相机的闪光灯进行采光。

资料显示：恶性肿瘤死亡率占全死因的百分率，已由70年代的12.57%上升到17.94%。为了与之相应，我们所选图像中的癌症与非癌症比例是1:4。其中癌症病例包括胃癌、肝癌、肺癌、血癌(即白血病)、直肠癌、结肠癌和乳癌。

我们分别用 K-均值方法和多中心动态聚类算法对上述图像进行分类实验。实验中的特征分别为基于 RGB 空间的特征矩阵、基于 HSV 空间的特征向量以及结合前两种特征的3×4综合特征矩阵：$F=\begin{bmatrix}v^1_R & v^1_G & v^1_B & \mu^1_H\\ v^2_R & v^2_G & v^2_B & \mu^2_H\\ v^3_R & v^3_G & v^3_B & \mu^3_H\end{bmatrix}$。实验结果如表1。

表1 识别结果

方法 / 特征	K-均值算法		多中心动态聚类算法	
	识别数	识别率	识别数	识别率
RGB	21	50%	23	54.7%
H	25	59.5%	26	61.9%
RGB&H	28	66.7%	30	71.43%

(下转第95页)

Conf. Document Analysis and Recognition, Vol. 1, 1997. 263～267
13 Munich M E, Perona P. Visual Signature Verification using Affine Arc-length. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 1999
14 Munich M E, Perona P. Camera-based ID Verification by Signature Tracking. In: Proc. of the 5th European Conf. on Computer Vision, 1998. 782～796
15 Wu Quen-Zong, et al. On-line signature verification using LPC cepstrum and neural networks. IEEE Transactions on Systems, Man and Cybernetics, Part B, Vol 27, No 1, pp148～153
16 Lee L L, et al. Reliable online human signature verification systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 18, No 6, pp643～647
17 Parizeau M, Plamondon R. A comparative analysis of regional correlation, dynamic time warping, and skeletal tree matching for signature verification. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, No 7, pp710～717
18 Brault J-J, Plamondon R. Segmenting Handwritten Signatures at their Perceptually Important Points. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, No 9, pp953～957
19 李永镐，金太均，安居院猛，中屿正之. A Study on the English Signature Verification Using Tree Matching. 电子情报通信学会论文志(日), D-II Vol. J75-D-II, No 1, pp31～38
20 Rabiner L. A tutorial on Hidden Markov Models and selected applications in speech recognition, Proc. IEEE, 1989, 77(2): 257～286
21 Bengio Y. Markovian Models for Sequential Data, Neural Computing Surveys 2, pp129--162, 1999.
22 边肇祺，张学工等著. 模式识别. 机械工业出版社，2000. 1
23 黄德双著. 神经网络模式识别系统理论. 机械工业出版社，1996. 5
24 Rabiner L, Juang B-H. Fundamentals of speech recognition. 清华大学出版社 & Prentice-Hall International, Inc. ，1999. 6

(上接第91页)

在临床医学检验中，误检率大致为1/3。实验结果表明：本文提出的多中心动态聚类算法能有效地对患癌手掌进行识别。它比K-均值方法有更好的分类效果。图1是几种癌症与非癌症手掌图像的几个典型示例。

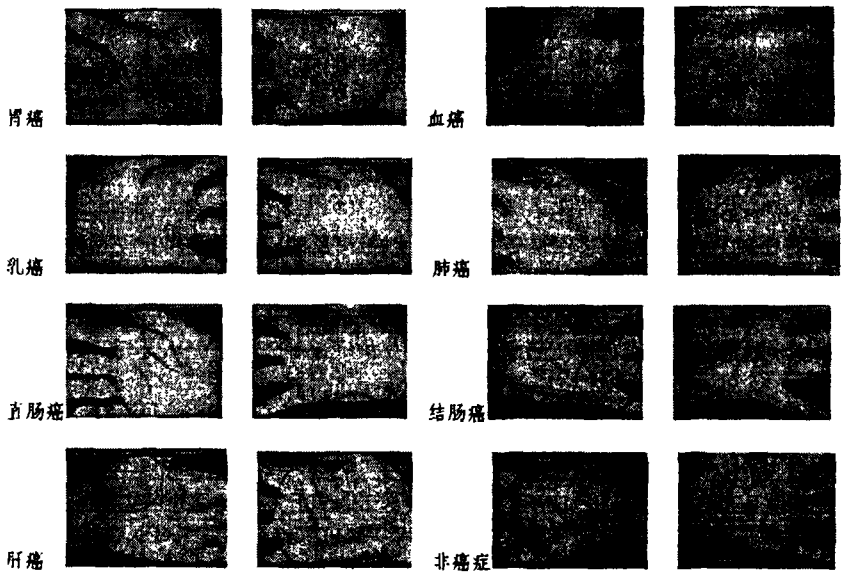

图1 几种癌症与非癌症的手掌图像

从图1可以看出，胃癌1(即胃癌示例中的第一幅图像，下同)，血癌1、2，肺癌1，直肠癌1、2，结肠癌1、2和肝癌2与非癌症手掌图像掌色差异明显或比较明显，可以通过掌色判定是患癌手掌图像。但实验不能根据手掌图像的全手掌颜色特征判断其余几幅手掌图像是否为癌症手掌。根据中医理论，与患病器官相对应的手掌区域内的颜色与正常人的掌色会有差别。胃癌2的圆圈限定的区域(以下简称区域)内有青黄色斑点，乳癌1的区域内有枯叶色的斑点，肺癌2的区域内青筋可见并伴有青黑色斑块，肝癌1的区域内有暗青色结节，这些都是相应器官癌变的表现。因此我们可以通过对手掌各个分区颜色特征的分析来进一步判断病情，提高识别的准确率。

结束语 生物特征识别技术应用于诊断人类疾病是一个全新的应用领域。本文对如何根据手掌的颜色特征进行自动手掌诊病进行了初步的探讨。实验表明，本文提出的多中心动态聚类算法可以较好地识别患癌手掌图像，识别率达到了临床应用的要求。这说明依据掌色特征进行自动手诊是可行的，并且能够通过生物特征识别技术取得较好的诊断效果。但是，有的手掌图像只是在与患病器官相应的手掌区域内的颜色特征有异而整体颜色与非癌症的手掌图像差异很小，这使得我们无法通过对全手掌颜色特征的分析来判断病人是否患癌。因此，我们的后续研究工作主要是如何根据手掌的不同分区内的颜色特征来诊断与之相应的器官的疾病。

参考文献

1 Zhang D D. Automated Biometrics Technologies and Systems. Kluwer Academic Publishers, 2000. 111～134
2 Wang Kuanquan, Zhang David, Pang Bo, Li Yanlai, Wu Xiangqian. Biometrics Based on Tongue Diagnosis of TCM. ICIG, 2000. 2008～2011
3 Kenneth R. Castleman. Digital Image Processing. Prentice Hall, NY, 1996
4 Shu W. Studies on Automatic Palmprint Identification. Dissertation, Tsinghua Univ. 1999
5 王晨霞. 现代掌纹诊病. 甘肃民族出版社，1992. 68～85
6 李莱田，田道正，焦春荣. 全息医学大全. 中国医药科技出版社，1999. 8
7 边肇祺，张学工，等. 模式识别. 清华大学出版社，2000，1：239～241