第 59 卷 第 4 期/2022 年 2 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于拓扑与网格双特征的铭文图形识别方法

刘文腾[1]，王慧琴[1*]，王可[1]，王展[2]

[1]西安建筑科技大学信息与控制工程学院，陕西 西安 710055；

[2]陕西省文物保护研究院，陕西 西安 710075

摘要 青铜器铭文图像有效特征的提取是进行铭文识别的关键步骤，针对以图像为信息载体的铭文特征提取方法由于特征维度高、特征向量复杂而识别准确度低的问题，提出了一种基于拓扑与网格双特征的铭文图形集成学习识别方法。以图形为铭文特征的表征，所提方法提取拓扑特征和7维文字结构图形特征，有效描述了铭文文字的结构信息。在此基础上，所提方法利用降维后铭文全局结构信息和局部结构信息的8维4方向弹性网格特征，解决了提取铭文图像特征导致的特征向量维度高的问题。最后，以拓扑特征和弹性网格特征作为集成学习样本的特征向量，利用Bagging方法对特征向量敏感程度不同的机器学习分类器进行集成，提升模型训练效率、提高识别精度。实验结果表明，与图像特征提取方法相比，所提方法对铭文识别准确率提高了15.54个百分点，并且铭文特征向量维度及运行时间大幅度降低。

关键词 图像处理；青铜器铭文；拓扑特征；网格特征；机器学习

中图分类号 TP391.1 **文献标志码** A **doi**：10.3788/LOP202259.0410018

Recognition Method of Inscription Graphics Based on Dual Features of Topology and Mesh

Liu Wenteng[1], Wang Huiqin[1*], Wang Ke[1], Wang Zhan[2]

[1]*Collge of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, Shaanxi* 710055, *China*;

[2]*Shaanxi Institute for the Preservation of Cultural Heritage, Xi'an, Shaanxi* 710075, *China*

Abstract Extracting the effective features of bronze inscription image is the key step of inscription recognition. Aiming at the problem of low recognition accuracy of inscription feature extraction method with image as information carrier due to high feature dimension and complex feature vector, an integrated learning and recognition method of inscription graphics based on the dual features of topology and mesh is proposed. Taking graphics as the representation of inscriptions, the proposed method extracts topological features and 7-dimensional text structure graphic features, which effectively describes the structure information of inscription text. On this basis, the proposed method uses the 8-dimensional and 4-direction elastic mesh features of the global and local structure information of the inscription after dimensionality reduction to solve the problem of high dimension of the feature vector caused by the extraction of the image features of the inscription. Finally, taking topological features and elastic mesh features as the feature vectors of integrated learning samples, Bagging method is used to integrate machine learning classifiers with different sensitivity of feature vectors, so as to improve the model training efficiency and recognition accuracy.

收稿日期：2021-03-08；**修回日期**：2021-04-01；**录用日期**：2021-04-14

基金项目：教育部归国留学人员科研扶持项目(K05055)、陕西省自然科学基金(2021JM-377)、陕西省科技厅国际科技合作计划(2020KW-012)、陕西省教育厅重点项目高端智库(18JT006)、西安市科技局项目(GXYD10.1)

通信作者：*hqwang@xauat.edu.cn

The experimental results show that compared with the image feature extraction method, the proposed method improves the accuracy of inscription recognition by 15.54 percent, and the dimension of inscription feature vector and running time are greatly reduced.

Key words image processing; bronze inscription; topological feature; mesh feature; machine learning

1 引 言

青铜器铭文是商周时期经济、文化、社会活动历史的见证。通过对青铜器铭文的研究,可考见当时社会的制度,纠正古籍中的伪谬。青铜器铭文的识别对探究古代历史和上古语言文字具有重要的意义。

从20世纪末开始已有学者借助计算机来实现古文字的识别[1],从此计算机替代人工来完成纷繁复杂的古文字研究工作已成为可能。最早于1996年,周新论等[2-3]提出的“甲骨文计算机识别方法”“甲骨文自动识别的图论方法”以拓扑结构来提取甲骨文特征并进行识别,有效地提取甲骨文中的拓扑特征,但缺乏关于特征提取操作的理论依据。酆格斐[4]的“基于数学形态学的甲骨文拓片字形特征提取方法”通过构造甲骨文的12项指标来表示其特征,通过计算特征向量之间的欧氏距离进行分类,但该方法关于模型分类的可解释性不强,在样本不平衡的情况下对特定类别的预测率低。吕肖庆等[5]的“一种基于图形识别的甲骨文分类方法”通过一种曲率直方图的傅里叶描述子对甲骨文进行分类,该方法具有平移、旋转、尺度不变性等特点,分类准确性高,但是实验数据量较小,仅有8个文字类别,不具有普适性。李文英等[6]的“一种基于深度学习的青铜器铭文识别方法”是基于改进的restnet18网络结构实现铭文识别的,该方法解决了铭文文字分类问题,但是识别效果不够理想,Top-1的情况下识别精度仅为58.3%。现有的隶定铭文数据集中,大部分铭文字头下存在的铭文变体数量较少,属于小样本数据集,对铭文图像特征的提取将增加小样本集数据的特征维度,容易造成识别分类器过学习的问题。

针对上述问题,本文提出了一种基于拓扑与网格(T-MF)双特征的铭文图形集成学习识别方法。拓扑特征将铭文图像抽象成图论中平面无向图进行处理,通过拓扑等价性质对变体之间相同的拓扑特征进行聚合,有效区分不同类别之间通过拓扑特征表现出的差异性。所提方法利用网格特征来描述铭文局部特征与全局特征,以4方向弹性网格模糊特征来表征铭文的文字结构特征,通过叠加子网格4方向分量来实现网格特征的降维。通过图形提取角度的拓扑与网格双特征,所提方法在降低特征维度的同时提高了特征提取效率。

2 基本原理

2.1 拓扑等价

拓扑学(Topology)是从图论演变而来的,是几何学的分支[7]。拓扑学是关于几何图形或空间在连续改变形态后还能保持不变的一些性质的研究,主要将实体抽象地解释成物体间的位置关系,而不考虑大小形状,将实体内部关系抽象成线来研究点与线之间的关系。

在拓扑学中,当一个对象通过弯曲、延展等变换变为另一个对象时,这两个对象为同胚(Homeomorphism)或拓扑等价[8]。圆形经过拉伸延展变换可以变为一个矩形,但无法通过变换成为圆环,因此圆形与矩形拓扑等价,不与圆环拓扑等价,如图1所示。设X和Y是拓扑空间,f为拓扑空间的映射关系,如果$f: X \to Y$是一一映射的,并且f及其逆g都是连续的,则f是同胚映射或拓扑变化。连续映射在由全部拓扑空间所构成的范畴中表示为$\to$,如果此时X与Y作为范畴中的对象同构,则X与Y拓扑等价。

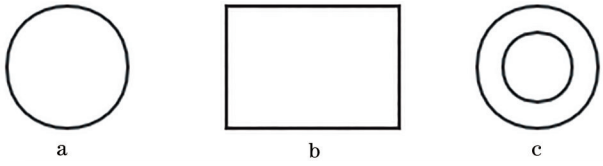

图1 拓扑等价示意图

Fig. 1 Topological equivalence diagram

对于铭文文字来说,同一种变体之间虽然会表现出不同的形态,但是拓扑特征是非常相似的。图2(a)、(c)同为字头“元”的变体,但是字形差异较大,图2(b)、(d)为图2(a)、(c)的骨架,可将图2(a)、(c)抽象为图2(b)、(d)。图2(b)中边e_1的两个顶点为c、a,图2(b)、(d)中顶点与边的对应关系相同,存在$f_b \to f_d$的映射关系,同时保持点(V)、线(E)、块(H)的数量不变,即$V_b = V_d = 8$,$E_b = E_d = 6$,

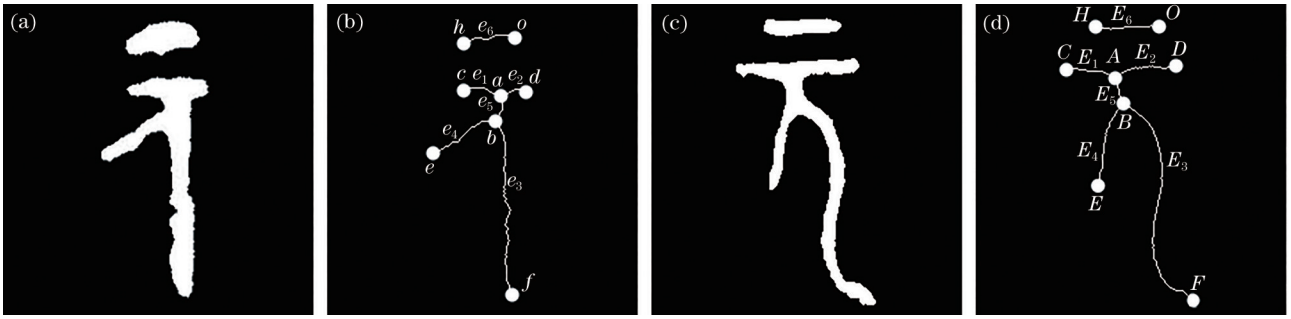

图 2　铭文图形拓扑等价。(a)(b)字头“元”及其骨架;(b)(d)字头“元”变体及其骨架

Fig. 2　Inscription graph topological equivalence. (a) (b) Prefix “Yuan” and its skeleton; (b) (d) prefix “Yuan” variant and its skeleton

$H_b = H_d = 2$,因此可以判定图 2(b)与图 2(d)拓扑等价。通过铭文抽象这一过程,从图形拓扑角度来分析,形态差异看似较大的两个铭文可视为同一种拓扑结构。

2.2　网格特征

网格特征是文字识别中常用的特征之一,是通过一组假想的线条对铭文图像进行区域的划分,进而通过对铭文图像的子网格进行特征提取而得到的[9-10]。网格特征大致上可以分为均匀网格和弹性网格两种[11],如图 3 所示。图 3(a)均匀网格,水平方向和垂直方向分别用线条进行划分,形成 4×4 的区域,每个区域称为子网格,每个子网格大小相等;图 3(b)为弹性网格,是根据铭文图像的笔画分布,用非均匀的网格线划分铭文得到的网格,又称非均匀网格。

根据铭文图像网格线条的方向,将弹性网格分成纵横弹性网格和对角弹性网格,从而构成 4 方向网格特征。目前 4 方向网格特征已经被证实是一种比较好的文字特征。4 方向网格特征可以理解为汉字中的“横”“竖”“撇”“捺”4 种笔画的方向特征[12]。4 方向弹性网格将子网格的大小作为权值来统计 4 个方向分量的大小。

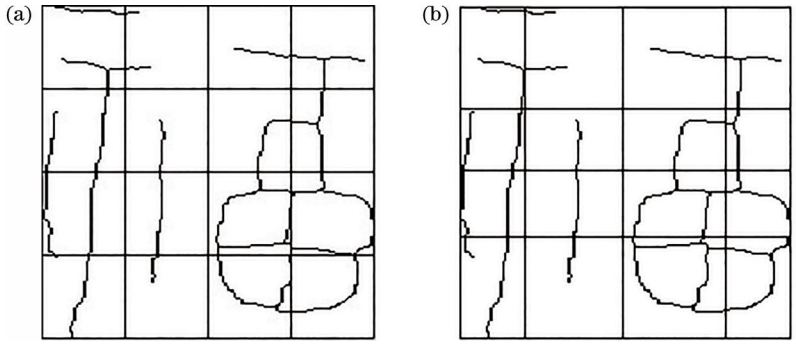

图 3　铭文网格划分示意图。(a)均匀网格;(b)弹性网格

Fig. 3　Schematic diagram of inscription mesh. (a) Uniform mesh; (b) Elastic mesh

3　铭文拓扑与网格双特征提取

拓扑特征将铭文图像抽象成无向图,利用同胚或拓扑等价的性质构建拓扑特征向量;同时为了更加全面地表征铭文的特征信息引入网格特征作为补充,利用降维后的 4 方向模糊弹性网格特征来表征铭文图像横、竖、撇、捺的 4 个分量。基于图形的特征可以有效减少铭文图像的高维特征信息,降低冗余特征对分类器的影响,提高铭文识别的准确率。

3.1　铭文拓扑结构特征

由于铸刻方式、时间跨度、地域文化的不同,青铜器铭文字形变化复杂。铭文文字虽然不具备现代汉字书写的线条化和规范性,但象形程度较高,展现更多的是线条图的特征。铭文中笔画的构成有 3 种:点、线、块,其中点代表笔画的起落点或者笔画的相交点;线表示相邻点之间的笔画线段;块表示不连通的笔画。基于图论的图概念,把青铜器铭文抽象成平面无向图处理。

结合铭文字形的笔画关系,可以将铭文拓扑特征归纳为 7 种:连通区域、亏格、端点、三叉点、四叉点、顶点相邻、平均度。

假设有 2 个像素点 A、B,如果像素点 A 与像素

点 B 邻接，则称 A、B 连通；假设有 3 个像素点 A、B、C，如果像素点 A 与像素点 B 邻接，B 与 C 邻接，则称 A、C 连通，如图 4 所示。视觉上彼此连通的像素点的集合构成了一个连通区域；具有偏旁部首分布特征的铭文，例如上下结构或左右结构的铭文，可视为具有多个连通区域特征，记为 T_C。

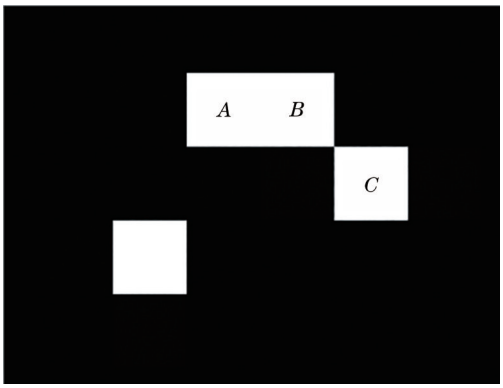

图 4　像素点连通区域

Fig. 4　Pixel connected area

若曲面中最多可画出 n 条闭合曲线同时不将曲面分开，则称该曲面的亏格为 n。以闭合曲面为例，亏格 n 就是曲面上孔洞的个数。铭文图像中许多字形皆包含一个或多个亏格，并且不同字头之间亏格数量差别较大，同类字形亏格数量差别较小。将一个图像区域中的孔数 H 和连接部分数 C 的差值定义为欧拉数 E[13]。

$$E = C - H, \tag{1}$$

式中：连接部分数 C 等价于 T_C；孔数 H 等价于亏格数 T_G。

$$T_G = C - E。 \tag{2}$$

无向图中度数等于 1 的点为端点。铭文中都有起笔和落笔，不存在无始无终的字形。由于铭文字形复杂程度不同，端点数也随之变化。简单字形仅有几个端点，而复杂铭文的端点数可以达到二十多个。因此对于铭文类别的区分，端点数是一个重要参数。直接提取铭文图像的端点数比较困难，因此需要对铭文图像进行细化处理，降低端点提取的复杂程度。设定端点特征值为 T_V。

无向图中度数为 3 的点为三叉点，度数为 4 的点为四叉点。在铭文图像中，两条笔画的相交之处为交叉点，可以分成 4 种：三叉点、四叉点、五叉点、六叉点。四叉点由 2 个三叉点组成，以八连通域为模板，三叉点可表示为 $\sum N_8(P)=3$，当 2 个三叉点之间的欧氏距离 $D \leqslant \sqrt{72}$ 时，这两个三叉点即为一个四叉点。以 T_3、T_4 分别表示三叉点与四叉点。

假设两个顶点被一条边相连，则称这两个顶点互为邻居顶点。在图 2 中，图 2(b)是图 2(a)细化后的图像，有 6 对相邻的顶点。设定顶点相邻特征值为 T_A。

无向图中一个顶点的度是指与该顶点相连边的个数，表示为 Degree(V_i)；平均度是指无向图中度数与顶点的平均数，表示为

$$T_{AD} = \frac{1}{N}\text{Degree}(V_i)。 \tag{3}$$

铭文拓扑结构特征提取算法步骤如下。

1) 输入预处理铭文二值图像，调用 Matlab 连通域标记函数 $[\boldsymbol{L}, N]$=bwlabel(image, 8)，其中 8 表示按照 8 邻域来划分，$\boldsymbol{L}$ 为 image 标记矩阵，N 表示连通域数，即 T_C。

2) 由(2)式可知，输入图像的亏格数 T_G 的值为 $T_C - E$，调用欧拉函数 bweuler 可求得输入图像的欧拉数，输出亏格数 T_G。

3) 对输入的二值图像进行 Zhang-Suen 细化处理，结合文献[14]中的改进算法进行处理，将 $\sum N_8(P)=1$ 的像素点标记，统计所有标记的端点，即 $T_V = \sum_{i=1}^{n}\sum N_8(P)=1$。

4) 三叉点和四叉点的特征值获取方法比较相似，在步骤 3) 细化处理后的图像中，标记 $\sum N_8(P)=3$ 的像素点，并统计所有标记点，即 $T_3 = \sum_{i=1}^{n}\sum N_8(P)=3$。对标记后的三叉点进行欧氏距离的计算，统计结果中欧氏距离 $D \leqslant \sqrt{72}$ 的标记点则为四叉点，记为 T_4。

5) 由步骤 3)和步骤 4)可获取端点与三叉点标记信息，将三叉点的标记点像素置为 0，利用步骤 1) 函数获取相关顶点邻特征值 T_A。

6) 无向图中的度 Degree(V_i)为顶点相邻边的个数，步骤 3)和步骤 4)中得出的端点数与三叉点数之和 $\sum_{i=1}^{N}\text{Degree}(V_i) = T_V + 3T_3$，无向图中的顶点数 $N = T_V + T_3$，因此平均度 $T_{AD} = \dfrac{T_V + 3T_3}{T_V + T_3}$。

3.2　四方向弹性网格模糊特征

设铭文图像 $I(i,j)$ 尺寸为 $m \times n$，子网格 B^i 周围的八邻域子网格分别为 B_1^i, B_2^i, B_3^i, B_4^i, B_5^i, B_6^i, B_7^i, B_8^i，如图 5 所示。方向特征提取相当于提取

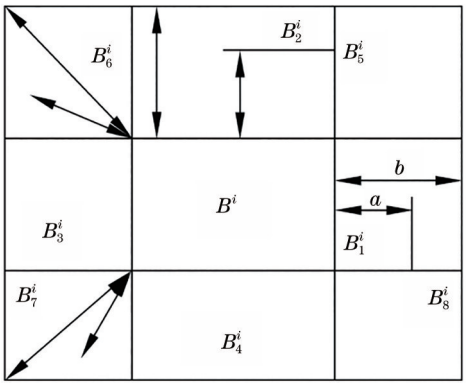

图 5　四方向关联图

Fig. 5　Four-direction correlation graph

“横”“竖”“撇”“捺”4 方向特征分量。以八邻域右子网格 B_1^i 和左子网格 B_3^i 的水平方向特征矢量作为子网格 B^i 的横方向弹性网格模糊特征；以八邻域上子网格 B_2^i 和下子网格 B_4^i 的竖直方向特征矢量作为子网格 B^i 的竖方向弹性网格模糊特征；以八邻域右上子网格 B_5^i 和左下子网格 B_7^i 的右倾方向特征矢量作为子网格 B^i 的撇方向弹性网格模糊特征；以八邻域左上子网格 B_6^i 和右下子网格 B_8^i 的左倾方向特征矢量作为子网格 B^i 的捺方向弹性网格模糊特征。

模糊特征主要表示子网格 B^i 的八邻域关系内的铭文笔画与子网格 B^i 之间的距离远近程度，越近则表示相关性越大，越远则表示相关性越小。采用动态模糊隶属度函数来反映子网格间的相关性，模糊隶属度函数表达式为

$$\mu(a)=\exp\left[-\frac{(3a/b)^2}{2}\right], \tag{4}$$

式中：子网格 B^i 到八邻域铭文笔画像素点的距离 $a\geqslant 0$；b 表示子网格 B^i 到八邻域外围边框的距离。横方向子网格 B^i 模糊特征的表达式为

$$T_{\mathrm{h}}^i=\frac{\sum H^{B^i}(p,q)}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}+\frac{\mu(a_{\mathrm{h}})\left[\sum H^{B_1^i}(p,q)+\sum H^{B_3^i}(p,q)\right]}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}, \tag{5}$$

式中：a_{h} 表示子网格 B^i 到 B_1^i 或 B_3^i 邻域网格铭文笔画像素点的距离；$H^{B_1^i}(p,q)$ 表示水平方向子网格 B_1^i 内的铭文笔画像素和；$H^{B_3^i}(p,q)$ 表示水平方向 B_3^i 内铭文笔画像素和。竖方向子网格 B^i 模糊特征的表达式为

$$T_{\mathrm{v}}^i=\frac{\sum V^{B^i}(p,q)}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}+\frac{\mu(a_{\mathrm{v}})\left[\sum V^{B_2^i}(p,q)+\sum V^{B_4^i}(p,q)\right]}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}, \tag{6}$$

式中：a_{v} 表示子网格 B^i 到 B_2^i 或 B_4^i 邻域网格铭文笔画像素点的距离；$V^{B_2^i}(p,q)$ 表示竖直方向子网格 B_2^i 内的铭文笔画像素和；$V^{B_4^i}(p,q)$ 表示竖直方向 B_4^i 内铭文笔画像素和。撇方向的子网格 B^i 模糊特征的表达式为

$$T_{\mathrm{p}}^i=\frac{\sum L^{B^i}(p,q)}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}+\frac{\mu(a_{\mathrm{p}})\left[\sum L^{B_5^i}(p,q)+\sum L^{B_7^i}(p,q)\right]}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}, \tag{7}$$

式中：a_{p} 表示子网格 B^i 到 B_5^i 或 B_7^i 内铭文笔画像素点的距离；$L^{B_5^i}(p,q)$ 表示撇方向子网格 B_5^i 内的铭文笔画像素和；$L^{B_7^i}(p,q)$ 表示撇方向 B_7^i 内铭文笔画像素和。同样，捺方向子网格 B^i 的模糊特征的表达式为

$$T_{\mathrm{w}}^i=\frac{\sum W^{B^i}(p,q)}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}+\frac{\mu(a_{\mathrm{w}})\left[\sum W^{B_6^i}(p,q)+\sum W^{B_8^i}(p,q)\right]}{\sum_{p=1}^{m}\sum_{q=1}^{n}I(p,q)}, \tag{8}$$

式中：a_{w} 表示子网格 B^i 到 B_6^i 或 B_8^i 内铭文笔画像素点的距离；$W^{B_6^i}(p,q)$ 表示捺方向子网格 B_6^i 内的铭文笔画像素和；$W^{B_8^i}(p,q)$ 表示捺方向 B_8^i 内铭文笔画像素和。

铭文图像 $I(i,j)$ 横、竖、撇、捺方向特征为 T_{h}、T_{v}、T_{p}、T_{w}。8×8 弹性网格的特征维度为 $8\times 8\times 4=256$，很大程度上存在特征冗余的情况。在拓扑特征缺少“横”“竖”“撇”“捺”4 方向细节特征的情况下，将 8×8 弹性网格特征降维，仅表现 4 方向分量结构模糊特征。

4 方向弹性网格模糊特征降维方法步骤如下。

1）以 8×8 弹性网格将铭文图像 $I(i,j)$ 分成

$8\times8=64$个子网格。

2）计算子网格B^i大小，并将64个子网格升序排列形成序列$A=\left\{n_{B^1},n_{B^2},\cdots,n_{B^{64}}\right\}$，降序排列形成序列$B=\left\{n_{B^{64}},n_{B^{63}},\cdots,n_{B^1}\right\}$。

3）将A序列中的子网格占比与4个方向特征相乘，求取子网格B^i 4方向弹性网格模糊特征$T_{dA}^i=\frac{n_{B^i}}{\sum_{i=1}^{64}n_{B^i}}T_d^i$，其中$d=(\mathrm{h},\mathrm{v},\mathrm{p},\mathrm{w})$。

4）将B序列中的子网格占比与4个方向特征相乘，求取子网格B^i 4方向弹性网格模糊特征$T_{dB}^i=\frac{n_{B^{64-i+1}}}{\sum_{i=1}^{64}n_{B^i}}T_d^i$，其中$d=(\mathrm{h},\mathrm{v},\mathrm{p},\mathrm{w})$。

5）对由A、B序列求得的子网格4方向特征进行叠加，得到表征局部细节特征的4方向弹性网格模糊特征T_{dA}和表征铭文整体特征的4方向弹性网格模糊特征T_{dB}。

通过对铭文图像子网格的4方向弹性网格模糊特征进行叠加融合，将弹性网格的$8\times8\times4=256$维特征降为$4\times2=8$维特征，同时保留了4方向的铭文局部细节特征和4方向的铭文全局特征。

4 铭文图形双特征的集成学习识别

青铜器铭文识别属于监督学习范畴，所用数据集是一种小样本多分类的特殊数据集。由于铭文数据样本过小并且铭文变体之间图像特征差异较大，深度学习无法提取同类别铭文中的共有特征，从而导致模型泛化误差变大。传统机器学习分类算法对于铭文图形特征处理具有时间复杂度低、异常数据敏感度低、处理数值型数据效率高、小样本数据泛化能力强等特点[15]。支持向量机(SVM)[16]、K最近邻(KNN)、CART决策树(decision tree)算法都是特别经典且高效的机器学习分类算法。这3种算法对于铭文的识别精度差别较小，但是每一种分类算法都有对于铭文数据处理的特殊性。由于铭文图形特征的数据结构，不同机器学习分类算法的识别效果有所不同。SVM分类算法对于样本点的敏感度比较高，以至于分类结果比较极端。对于铭文特征向量在样本数据中特征属性差别较大的一类，SVM分类算法识别正确率可以达到100%，相反对个别类甚至可以降至0%，在样本的特征空间中特征值差别越大，分类效果越明显。因为少数支持向量决定了分类的结果，所以将SVM作为基分类器不仅使算法模型复杂度降低而且鲁棒性较好[17]；KNN分类算法具有良好的适应能力，通过计算两个特征向量之间的欧氏距离来表示相似程度[18]，对于特征值相同或近似的铭文识别精度非常高，但由于需要对每个训练样本求解一次最近距离，数据量过于庞大的数据集会导致KNN算法时间复杂度升高。铭文数据集属于小样本数据，并且特征向量只有15维，所以KNN对铭文数据集具有高效准确的识别效果；CART算法是通过计算不同特征之间的GINI系数来选取最优指标，递归生成决策树的。CART算法是多分类中最经典的一种分类方法，对铭文数据中的离群样本点表现不敏感、计算复杂度低、构造简单[19]。

上述3种机器学习分类器对铭文数据特征较为敏感，若采用单一分类器进行铭文识别，分类效果相对较差，因此将多个弱学习器采用结合策略集成为强学习器，通过集成学习器的训练模型进行铭文数据预测。Bagging集成学习算法[20]通过有放回采样生成了多个子数据集，降低了奇异点对训练结果的影响，并且融合多个弱分类器，使强分类器具有更好的泛化能力，有效解决了弱分类器对铭文数据表现敏感的问题。

Bagging集成学习器算法实现步骤如图6所示。

Input：训练集D，测试集T，基分类器数目K，基学习器算法SVM、KNN、CART，结合策略为加权投票法。

将基分类器集合记为$E=\left\{e_k\right\}$，$k=1,2,\cdots,k$，输入样本x，基分类器e_k的输出类别记为j_k，即$e_k(x)=j_k$，$j_k\in\{1,2,\cdots,M\}$。用C_i表示第i个类别，$i\in\Lambda=\{1,2,\cdots,M\}$，有

$$T_k\left(x\in C_i\right)=\begin{cases}1, e_k(x)=i\\0, e_k(x)\neq i\end{cases},\tag{9}$$

$$T_E\left(x\in C_i\right)=\omega_j\sum_{k=1}^{k}T_k\left(x\in C_i\right),\ i\in\Lambda=\{1,2,\cdots,M\},\tag{10}$$

式中：$T_E\left(x\in C_i\right)$表示第i个类别的投票；$\sum_{k=1}^{k}T_k$表示基分类器e_k输出正确的数量和；ω_j表示分类器SVM、KNN、CART的权值，$\omega_1+\omega_2+\omega_3=1$并根据3种基分类器在铭文数据集上预测正确率赋予不同的权值ω_j，$\omega_1=0.312$、$\omega_2=0.338$、$\omega_3=0.350$。

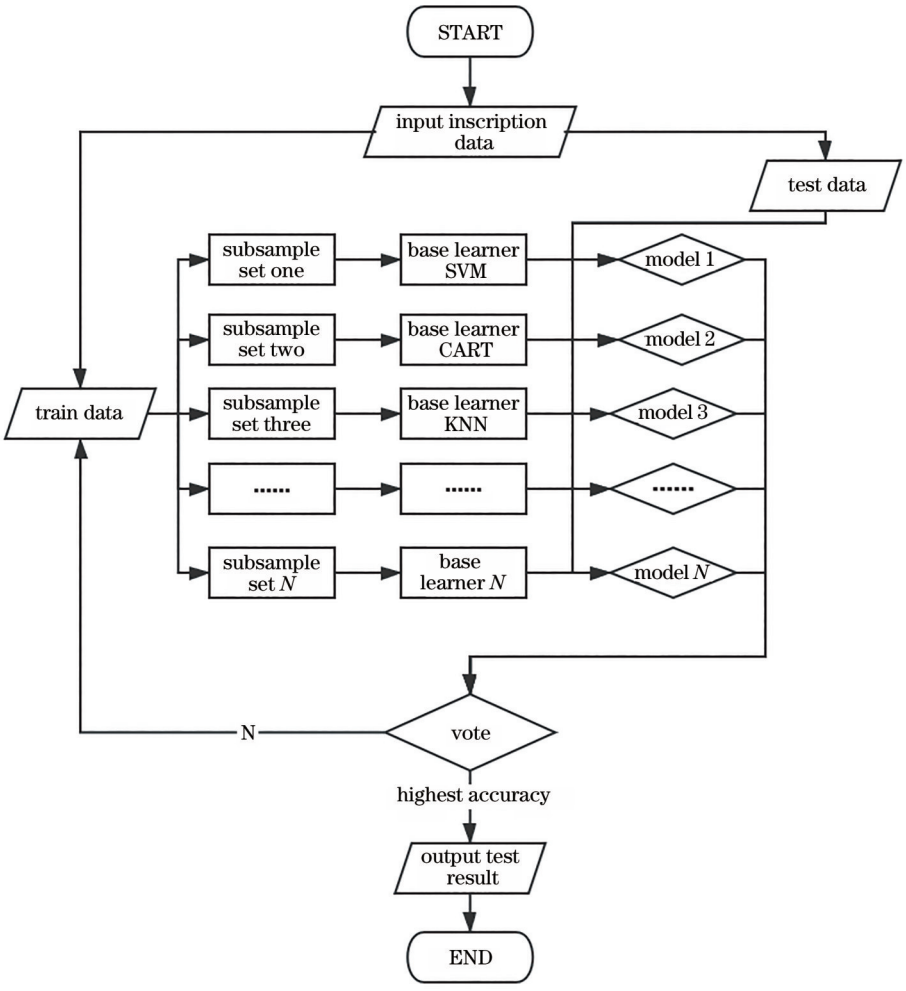

图 6　Bagging 模型训练技术路线图

Fig. 6　Training technology roadmap of Bagging model

则投票法的表决函数可以表示为

$$E(x)=\begin{cases} j, T_E(x\in C_j)=\max T_E(x\in C_i) \\ i=j, T_E(x\in C_j)=T_E(x\in C_i) \end{cases}, \tag{11}$$

式中：$E(x)$表示基分类器集合E对输入样本x的预测结果。

Output：集成学习模型的分类结果。

1）训练集D采取又放回随机抽样，组成子训练集$D_i, i=1, 2, \cdots, k$；

2）使基学习器在子训练集D_i上进行学习，生成子模型$M_i, i=1, 2, \cdots, k$；

3）在测试集T上采用投票表决法对各个子模型M_i的预测结果进行组合并返回投票分类的结果。

5　实验结果分析

目前的铭文识别特征提取方法使用图像形状特征与高阶特征较多。图像形状特征中尺度不变特征变换（SIFT）和方向梯度直方图（HOG）特征[21]的使用最为广泛，文献[7]中使用的RestNet18则属于图像高阶特征。为了说明所提算法的有效性，将所提算法与SIFT、HOG、RestNet18三种图像特征提取算法进行了对比。结合实验数据对铭文图像算法与铭文图形算法进行分析。

5.1　铭文特征提取的评价标准

为了验证算法的性能，以特征维度、运行时间、准确率3个方面作为铭文特征提取算法的评价标准。

1）特征维度D。特征提取算法的特征维度直接影响了分类模型的识别效率。对于图像特征来说，特征维度为特征提取算法提取到图像中特征点的复杂度w与图像中特征点数T'的乘积。

$$D=\sum_{t=1}^{T'} w_t。 \tag{12}$$

2）运行时间T。对于相同的分类器，不同特征

提取算法的运行时间也是一个非常重要的评价指标。对于客观事物的识别，实时性可以体现一个识别模型的合理性。为了更好地体现这一评价标准，将铭文数据集分为相同的训练集与测试集并进行训练和测试。为了保证准确性，进行多次实验求平均运行时间。

$$T=\frac{1}{N}\sum_{i=1}^{N}t/n。\tag{13}$$

3）准确率。准确率是一个特征提取算法相对重要的指标，准确率表示模型预测结果正确数量在测试集中的比例。

$$A=\frac{C}{T_c}\times 100\%,\tag{14}$$

式中：C 表示预测结果正确数量；T_c 表示测试集数量。

5.2　实验分析

实验使用青铜器铭文信息数据库中的铭文数据进行实验，为了减少铭文单一性带来的误差，选取铭文变体较多的类别作为铭文数据集。该数据集包括铭文字头 32 个，铭文原始图像 1120 张，平均每个铭文字头下变体数量大约为 30 个。由于铭文识别属于多分类问题，并且样本数据量较小，为了保证实验数据的可靠性，对 32 个铭文字头下的铭文变体进行数据增强，通过小幅度旋转和水平、垂直平移操作实现铭文数据扩增。图 7 为部分铭文数据。

图 7　部分铭文数据

Fig. 7　Partial inscription data

为了直观地对比不同算法的特征维度，随机抽取了不同类别中 5 张铭文图像进行分析。以特征维度为对象，研究图像与图形两种不同的特征提取方向。表 1 为不同算法的特征维度。

表 1　不同特征提取算法的特征维度

Table 1　Feature dimension of different feature extraction algorithms

Class	Propoesd algorithm	SIFT	HOG	RestNet18
1	15	11136	8100	25088
2	15	25344	8100	25088
3	15	38912	8100	25088
4	15	66176	8100	25088
5	15	12672	8100	25088
Feature structure	$7+4\times 2$	$128\times n$	$36\times 15\times 15$	$7\times 7\times 512$

由表 1 可以看出，图像特征的维度远远高于图形特征。原因在于图像特征提取算法需要研究图像中具有特殊性的像素点或者像素区域。以 SIFT 为例，每个关键点的维度为 128，铭文图像中的关键点多达上百个。对于铭文而言，铭文数据是一种小样本的数据集，样本属性值具有唯一性，以至于图像特征提取算法得到的铭文特征差异较大。虽然所提算法只有 15 个特征属性，但是每一种特征属性都从图形角度解析铭文特征，并对铭文的结构进行表征，每一种特征属性都具有较好的区分能力。

集成学习基分类器数量是较为关键的参数，决定了学习器的准确率及模型的效率。在分类器数-泛化误差曲线达到极小值时，集成学习器分类结果最优，如图 8 所示。从图 8 中可以看出，当分类器数量

逐渐变多时，泛化误差明显降低，考虑到分类器学习的效率，当基分类器数量为450时达到最佳效果。

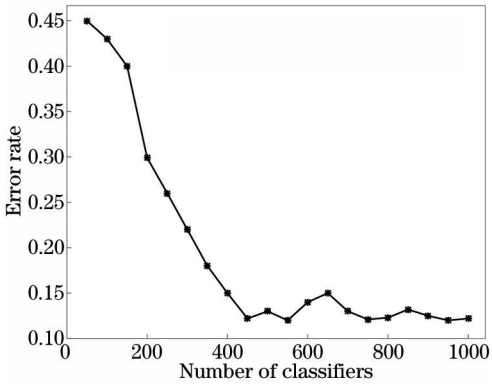

图8　分类器数-泛化误差曲线

Fig. 8　Number of classifiers-generalization error curve

基分类器的组合关系同样会影响到集成分类器的准确率，通过对不同分类器赋予权值来研究对集成分类器准确率的影响。设SVM、KNN、CART三种基分类器的权值参数为x、y、z并且满足条件$x+y+z=1$，其中z可以表示为$1-(x+y)$。通过调整变量x、y可以得到相应的基分类器组合关系。从图9基分类器组合关系曲面拟合图可知，$x=y=0.33$时，集成分类器的效果达到最优。

表2　特征提取算法模型效率

Table 2　Feature extraction algorithm model efficiency

Classification model	Algorithm	Average running time /s
Bagging	Propoesd algorithm	0.02
	SIFT	1.53
	HOG	1.31
RestNet18	RestNet18	0.68

从表3中可以看出，由于铭文小样本的特殊性，卷积神经网络模型不能全面地提取铭文数据中同类别之间的共性，导致分类正确率较低。SIFT特征提取算法对相同的铭文图像具有较高的辨识度，但是不同变体之间铭文形态的差异性导致识别率降低。所提双特征提取算法相比SIFT特征提取算法的平均正确率提高了15.54个百分点。

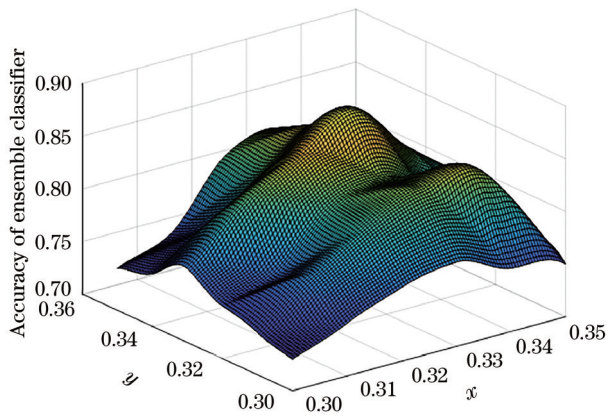

图9　基分类器组合关系曲面拟合图

Fig. 9　Combination relationship surface fitting diagram of base classifier

所提算法的特征维度非常低，很大程度降低了分类模型的时间复杂度。表2为不同特征提取算法在分类模型上的运行效率。很显然，相同的分类器模型下，SIFT和HOG特征提取算法由于特征复杂度比较高，分类模型的运行时间也随之增大。所提算法的15维特征很大程度上降低了分类模型的时间复杂度。

为了更全面地评估所提双特征提取算法，对铭文数据进行5折交叉验证，5次实验结果如表3所示。

通过对特征提取算法的特征维度、时间复杂度、识别正确率3个方面进行实验分析，相比于SIFT、HOG及RestNet18，所提算法特征维度仅有$7+4\times2=15$维，极大地降低了分类模型的时间复杂度的同时，识别识别准确率最高，为实现铭文多种类、实时性识别提供了可靠的特征提取方案。

表3　四种特征提取算法的正确率

Table 3　Accuracy of four feature extraction algorithms　　unit: %

Five-fold cross validation	Bagging			RestNet18
	T-MF	SIFT	HOG	
1	89.29	75.45	66.52	56.25
2	86.16	68.75	64.73	54.46
3	87.50	73.66	68.75	61.60
4	91.07	76.34	67.86	59.82
5	85.71	67.86	63.39	52.23
Average	87.95	72.41	66.25	56.87

6 总　　结

提出了一种基于拓扑和结构特征的铭文图形机器学习识别方法。通过研究铭文的拓扑结构将青铜器铭文抽象为无向图，将拓扑等价性质作为铭文拓扑特征相似度的表征。研究了网格特征降维方法，构建了8×8弹性网格4方向模糊特征，提取了子网格4方向模糊特征，通过不同权值的影响形成局部的4方向弹性网格模糊特征与全局的4方向弹性网格模糊特征。实验结果表明，利用拓扑与网格特征提取算法提取的特征向量具有特征维度低、特征复杂度小、冗余特征少的优点，有效提高了分类器模型的实时性和准确性。

参 考 文 献

[1] Tian Y. The script recognition technology of the bamboo slips in the warring states period based on deep metric learning[D]. Kaifeng: Henan University, 2020.
田园. 基于深度度量学习的战国简文字识别技术[D]. 开封: 河南大学, 2020.

[2] Zhou X L, Li F, Hua X C, et al. A method of Jia Gu Wen recognition based on a two-level classification [J]. Journal of Fudan University(Natural Science), 1996, 35(5): 481-486.
周新伦, 李锋, 华星城, 等. 甲骨文计算机识别方法研究[J]. 复旦学报(自然科学版), 1996, 35(5): 481-486.

[3] Li F, Zhou X L. Recognition of Jia Gu Wen based on graph theory[J]. Journal of Electronics & Information Technology, 1996, 18(S1): 41-47.
李锋, 周新伦. 甲骨文自动识别的图论方法[J]. 电子科学学刊, 1996, 18(S1): 41-47.

[4] Feng G F, Gu S T, Yang Y M. Feature extraction method of oracle bone inscriptions based on mathematical morphology[J]. Journal of Chinese Information Processing, 2013, 27(2): 79-85.
酆格斐, 顾绍通, 杨亦鸣. 基于数学形态学的甲骨拓片字形特征提取方法[J]. 中文信息学报, 2013, 27(2): 79-85.

[5] Lü X Q, Li M N, Cai K W, et al. A graphic-based method for Chinese oracle-bone classification[J]. Journal of Beijing Information Science & Technology University, 2010, 25(S2): 92-96.
吕肖庆, 李沫楠, 蔡凯伟, 等. 一种基于图形识别的甲骨文分类方法[J]. 北京信息科技大学学报(自然科学版), 2010, 25(S2): 92-96.

[6] Li W Y, Cao B, Cao C S, et al. A deep learning based method for bronze inscription recognition[J]. Acta Automatica Sinica, 2018, 44(11): 2023-2030.
李文英, 曹斌, 曹春水, 等. 一种基于深度学习的青铜器铭文识别方法[J]. 自动化学报, 2018, 44(11): 2023-2030.

[7] Gu S T. Identification of oracle-bone script fonts based on topological registration[J]. Computer & Digital Engineering, 2016, 44(10): 2001-2006.
顾绍通. 基于拓扑配准的甲骨文字形识别方法[J]. 计算机与数字工程, 2016, 44(10): 2001-2006.

[8] Dahmen R, Lukács G. Long colimits of topological groups I: continuous maps and homeomorphisms[J]. Topology and Its Applications, 2020, 270: 106938.

[9] Qi Y M, Tian X D, Zuo L N. Segmentation method of ancient Chinese character images based on hesitant fuzzy sets[J]. Science Technology and Engineering, 2019, 19(30): 232-240.
齐艳媚, 田学东, 左丽娜. 基于犹豫模糊集的古籍汉字图像切分方法[J]. 科学技术与工程, 2019, 19(30): 232-240.

[10] Tian X D, Chai Y L, Wang H B. Retrieval method of ancient Chinese character images based on hesitant fuzzy features[J]. Computer Engineering, 2019, 45(3): 217-224.
田学东, 柴彦立, 王海彬. 基于犹豫模糊特征的古籍汉字图像检索方法[J]. 计算机工程, 2019, 45(3): 217-224.

[11] He H Z, Zhu N B, Liu W. Handwritten Chinese character recognition based on Hough transformation and elastic mesh[J]. Computer Simulation, 2008, 25(1): 240-243.
何浩智, 朱宁波, 刘伟. 基于霍夫变换和弹性网格的手写汉字识别方法[J]. 计算机仿真, 2008, 25(1): 240-243.

[12] Wang J P, Wang G X, Li W T, et al. An off-line handwritten Chinese character cognitive model based on simulated feedback mechanism[J]. Control Engineering of China, 2019, 26(3): 476-483.
王建平, 王光新, 李帷韬, 等. 基于仿反馈机制的脱机手写体汉字认知模型[J]. 控制工程, 2019, 26(3): 476-483.

[13] Ni J H, Zhou X G. The general description of topological relations based on node degree and Euler-number[J]. Remote Sensing Technology and Application, 2011, 26(4): 527-532.
倪建华, 周晓光. 基于结点度和欧拉数的拓扑关系一般化描述[J]. 遥感技术与应用, 2011, 26(4):

527-532.

[14] Chang Q H, Wu M H, Luo L M. Handwritten Chinese character skeleton extraction based on improved ZS thinning algorithm[J]. Computer Applications and Software, 2020, 37(7): 107-113, 164.
常庆贺, 吴敏华, 骆力明. 基于改进ZS细化算法的手写体汉字骨架提取[J]. 计算机应用与软件, 2020, 37(7): 107-113, 164.

[15] Zhang R, Wang Y B. Research on machine learning with algorithm and development[J]. Journal of Communication University of China (Science and Technology), 2016, 23(2): 10-18, 24.
张润, 王永滨. 机器学习及其算法和发展研究[J]. 中国传媒大学学报(自然科学版), 2016, 23(2): 10-18, 24.

[16] Juliet Selwyn E, Velayutham S S, George J F D. Improved compound image segmentation using automatic pixel block classification with SVM[J]. IET Image Processing, 2020, 14(8): 1605-1613.

[17] Liu J K, Li C Y, Lu H, et al. Classification and recognition of disposable masks based on Raman spectroscopy and mechine learning[J/OL]. Laser & Optoelectronics Progress: 1-16[2021-03-30]. http://kns.cnki.net/kcms/detail/31.1690.tn.20201015.0915.004.html.
刘金坤, 李春宇, 吕航, 等. 基于拉曼光谱和机器学习方法的一次性口罩分类识别[J/OL]. 激光与光电子学进展: 1-16[2021-03-30]. http://kns.cnki.net/kcms/detail/31.1690.tn.20201015.0915.004.html.

[18] Yu T, Yang J. Point cloud model recognition and classification based on K-nearest neighbor convolutional neural network[J]. Laser & Optoelectronics Progress, 2020, 57(10): 101510.
于挺, 杨军. 基于K近邻卷积神经网络的点云模型识别与分类[J]. 激光与光电子学进展, 2020, 57(10): 101510.

[19] Winster S G, Kumar M N. Automatic classification of emotions in news articles through ensemble decision tree classification techniques[J]. Journal of Ambient Intelligence and Humanized Computing, 2021, 12(5): 5709-5720.

[20] Liu Y X, Lü H, Hu T, et al. Research on character recognition based on Bagging ensemble learning[J]. Computer Engineering and Applications, 2012, 48(33): 194-196, 211.
刘余霞, 吕虹, 胡涛, 等. 基于Bagging集成学习的字符识别方法[J]. 计算机工程与应用, 2012, 48(33): 194-196, 211.

[21] Zhao R Q, Wang H Q, Wang K, et al. Recognition of bronze inscriptions image based on mixed features of histogram of oriented gradient and gray level co-occurrence matrix[J]. Laser & Optoelectronics Progress, 2020, 57(12): 121003.
赵若晴, 王慧琴, 王可, 等. 基于方向梯度直方图和灰度共生矩阵混合特征的金文图像识别[J]. 激光与光电子学进展, 2020, 57(12): 121003.