

# 基于支持向量机的颅骨性别识别

杨 稳, 刘晓宁, 朱 菲

(西北大学 信息科学与技术学院 陕西 西安 710127)

**摘 要:** 颅骨性别识别在法医学和颅骨面貌复原等领域具有重要研究意义和应用价值。文中以新疆吐鲁番地区 117 例维吾尔族成人三维颅骨数字模型为研究对象, 首先, 对颅骨模型利用自主开发的系统标定 78 个特征点, 其中 12 个位于颅骨正中矢状面、66 个对称分布于颅骨两侧; 然后, 提取可测量特征和非可测量特征, 对可测量特征直接测量, 对非可测量特征进行量化表示; 最后, 利用支持向量机方法对提取的特征向量进行降维并设计分类器, 实现对颅骨的性别分类。实验结果表明, 将测量特征和非可测量特征结合包含更多的性别识别信息, 支持向量机的方法可以很好地实现性别分类, 并能提高性别识别精度, 利用留一交叉验证进行测试, 其中男性识别正确率达 90.0%, 女性识别正确率达 94.7%, 平均识别正确率达 92.4%。

**关键词:** 性别识别; 可测量特征; 非可测量特征; 支持向量机; 留一交叉验证

中图分类号: TP181

文献标识码: A

文章编号: 1673-629X(2019)02-0043-05

doi: 10.3969/j.issn.1673-629X.2019.02.009

## Sex Determination of Skull Based on Support Vector Machine

YANG Wen, LIU Xiao-ning, ZHU Fei

(School of Information Science and Technology, Northwest University, Xi'an 710127, China)

**Abstract:** Sex determination of skull has important research significance and application value in forensic medicine and skull rejuvenation. In this paper, we take the 117 cases of Uygur adult three-dimensional skull digital model in Turpan area of Xinjiang as the research object. First, 78 features points of the skull model are calibrated by the independent development system, of which 12 are located in the median sagittal plane of the skull and 66 are symmetrically distributed on both sides of the skull. Then, the measurable features and non-measurable features are extracted, and the former is measured directly and the latter is quantized. Finally, the support vector machine method is used to reduce the dimensions of the extracted eigenvectors and design the classifier to complete the classification of the skull. The experiment shows that combining the measured and non-measurable features with more sex determination information, SVM can achieve gender classification well and improve the accuracy of gender identification. Using leave-out cross validation test, the recognition accuracy of males is 90%, that of females is 94.7%, and that of the average 92.4%.

**Key words:** sex determination; measurable features; non-measurable features; support vector machine; leave-out cross validation

### 1 概 述

骨骼遗骸的性别决定是法医人类学鉴定过程中的重要一步。人类学家研究表明, 在人体骨骼中, 颅骨是最能体现性别差异的骨骼之一<sup>[1]</sup>。在传统方法<sup>[2-5]</sup>中, 应用最普遍的是线性判别分析方法。在对颅骨进行实体测量的基础上, Ramamoorthy 等针对南印度 70 个成人颅骨样本, 测量了 26 项特征指标, 利用 SPSS 建立判别函数进行分析, 单变量、逐步和多变量判别函数的准确率分别为 77.1%、85.7% 和 72.9%; 李明等对国

内西南地区 67 个性别明确的成人颅骨测量了颅长、颅宽等 16 项指标, 建立单变量及多变量性别判定方程, 得到的男性判别准确率为 89.2%, 女性判别准确率为 90.0%。随着计算机技术的飞速发展, 研究者开始借助计算机对颅骨特征指标进行测量, 因此计算机辅助测量成为趋势。

Tanya 等对 50 名成人颅骨数字侧位 X 线片使用 Sidexis XG 软件测量上颌窦, 计算上颌窦指数进行判别函数分析, 并推导判别性别的判别式, 得出的判别函

收稿日期: 2018-03-10

修回日期: 2018-07-12

网络出版时间: 2018-11-15

基金项目: 国家自然科学基金(61363065); 陕西省自然科学基金(2014JM8358); 研究生自主创新项目(YZZ17181)

作者简介: 杨 稳(1993-), 男, 硕士研究生, CCF 会员(86535G), 研究方向为机器学习与模式识别; 刘晓宁, 副教授, 硕士生导师, CCF 会员(20819M), 研究方向为图像处理与可视化技术。

网络出版地址: <http://kns.cnki.net/kcms/detail/61.1450.tp.20181115.1046.018.html>

数分析能够区分性别,其敏感性为 68%,特异性为 76%。线性判别分析方法虽然简单,但是该方法需要严格的假设前提,且不同地区、不同人种的颅骨的特征指标具有明显的差异,而且测量工作繁琐,准确率也不够高。

目前有很多学者<sup>[6-8]</sup>在研究中发现,选取合适的分类器在性别识别过程中具有重要作用。Afrianty 等对 91 例人类骶骨测量了 6 项特征指标,将其作为反向传播网络的输入,分别用两种网络架构进行实验,识别准确率达到 99.03%,并与传统的判别函数分析方法进行了对比,反向传播神经网络的性别识别率明显高于判别函数分析方法,但该方法对数据集的要求高,如何选取合适的样本实例作为训练集是个难题;随着三维数字化技术的发展,Luo 等提出一种基于稀疏主成分分析将颅骨的局部形态特征与性别分类相关联的自动方法,对 208 例中国成人颅骨进行实验,结果显示 SP-CA 对颅骨性别识别非常有效,识别率达 95% 以上,但该方法对颅骨样本完整性要求高,颅骨必须具有局部特征,局部信息影响识别结果。

基于上述分析,文中提出一种基于支持向量机的颅骨性别识别方法。该方法结合法医人类学和颅骨解剖学知识,标定自定义颅骨特征点集;利用 Fourier 变换对额骨和鼻根形态进行量化表示,用自主开发的计算机测量系统完成对颅骨可测量特征的测量,将非可测量特征和可测量特征融合;对上述步骤中得到的特征向量进行降维,采用支持向量机(support vector machine, SVM)设计分类器进行颅骨性别鉴定。算法流程如图 1 所示。

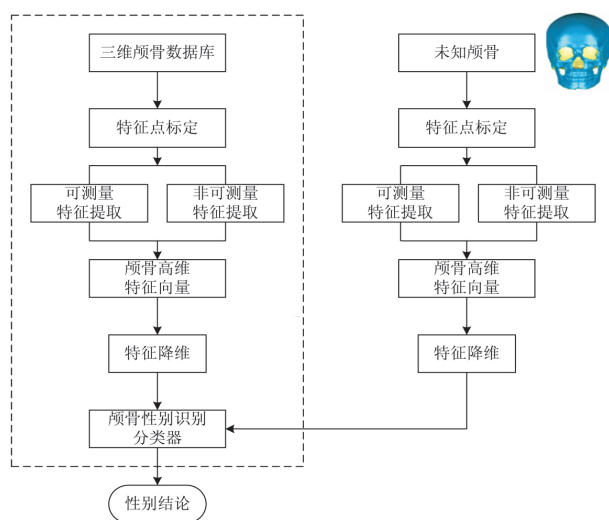


图 1 算法流程

## 2 颅骨特征提取

### 2.1 特征点定义与标定

以西门子多排螺旋 CT 机采集的新疆吐鲁番地区

267 例维族成人颅骨数据为研究对象,随机选取性别明确、无颅病理的 117 例完整颅骨数据  $L = \{L_1, L_2, \dots, L_n\}$ ,  $n = 117$  作为实验对象,其中  $L$  数据中男性 60 例、女性 57 例,男性和女性的年龄均值分别为 46.95 和 47.7,标准差分别为 6.58 和 4.39。

利用项目组自主开发的系统对颅骨 CT 数据进行重构,得到三维数字化颅骨模型,将模型转换到法兰克福坐标系下并进行归一化处理,然后进行颅骨特征点的标定。很多学者对颅骨特征点的标定进行了研究,文中根据文献[9]中颅骨特征点的定义,完成了颅骨性别鉴定问题的颅骨特征点标定。对颅骨定义了 78 个颅骨特征点,其中正中矢状面 12 个,对称地分布于颅骨两侧的 66 个。

从 117 例颅骨中选取一套外观完整的模型作为标准模型,对标准模型用项目组自主开发的标定系统手动标定定义的 78 个颅骨特征点。其余颅骨模型利用 ICP(iterative closest point)配准算法<sup>[10]</sup>使其与标准模型对齐,自动实现特征点标定。

### 2.2 可测量特征提取

根据法医学和颅骨解剖学知识和定义的 78 个特征点,并考虑计算机软件自动测量过程的要求,文中定义了 27 项可测量指标,其中 22 项几何测量指标,5 项角度测量指标,通过欧几里德和测地线距离以及角度测量软件计算特征指标。

### 2.3 非可测量特征量化

通过阅读颅骨形态特征相关文献<sup>[11]</sup>,颅骨额骨和鼻根形态等为非可测量的形态。可利用数字几何和曲线拟合方法实现形态量化,将其转化为可测量并可进行统计的三维颅骨特征。

非可测量特征额骨和鼻根是颅骨性别差异的重要区域,这里应用傅里叶变换对这两个非测量特征进行数据量化表示。首先,在额骨与鼻根区域范围内分别标定 18 个点,运用 Matlab 自带的 cftool 曲线拟合工具箱拟合出三维颅骨的额骨线和鼻根点凹陷曲线;其次,利用 LM(Levenberg Marquardt)算法对空间曲线进行优化;最后,将三维空间曲线向二维平面  $XY$  进行投影,对投影曲线  $S$  做傅里叶变换。以曲线拟合额骨线为例,男女额骨拟合曲线向  $XY$  平面进行投影,获取投影后的额骨线如图 2 所示。

使用 cftool 曲线拟合工具箱对男女的额骨线进行曲线拟合,拟合后的男女曲线方程分别为:

$$y_1 = -8.6685 - 1.4380x - 2.3916x^2 - 3.9866x^3 + 1.0611x^4 - 4.0987x^5 - 3.2631x^6$$

$$y_2 = -16.3093 - 5.0773x - 7.2894x^2 + 0.1779x^3 - 0.0003x^4 + 2.0852x^5 -$$

$$4.652\ 5x^6$$

利用文献[12]中的傅里叶变换也可对男女额骨线的形态进行量化表示,将二维曲线 $S$ 的 $X$ 轴划分为32份,求曲线上对应的 $Y$ 值,最后计算出合成振幅作为性别鉴定的测量指标。对额骨和鼻根形态均利用Fourier变换,共获取了32个性别鉴定的测量指标。

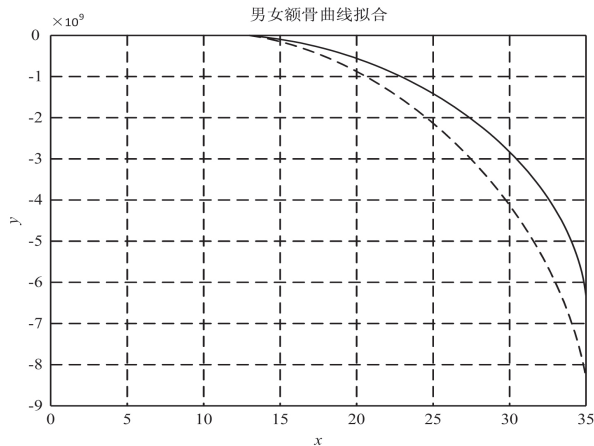


图2 额骨线

### 3 支持向量机

支持向量机是一种模式分类和回归的学习算法。支持向量机的基本训练原理是寻找最优线性超平面,使未知的测试样本的预期分类误差最小化,即良好的泛化性能。根据结构风险最小化原则,一种精确分类训练数据属于一组具有最低VC维度的函数将其优化,而不管输入空间的维数如何。基于这个原则,线性SVM使用系统的方法来找到具有最低VC维的线性函数。对于线性不可分数据,SVM可以将输入映射到线性超平面的高维特征空间中。由于SVM具有很好的学习能力且能够解决小样本、非线性及高维度分类等问题<sup>[13]</sup>,因此,SVM成为处理性别鉴定问题的首选分类器。另外,SVM中核函数的选取是模式识别领域的重要研究内容,分类器参数的设置是建立颅骨识别模型的关键。

#### 3.1 支持向量分类机

给定一个标记的 $M$ 个训练样本集 $(x_i, y_i)$ ,其中 $x_i \in R^N$ 和 $y_i \in R^N$  ( $y_i \in \{-1, 1\}$ )是相关联的。SVM分类器找到正确最大分离超平面数据点的一小部分,同时最大化任意一个类到超平面的距离。Vapnik<sup>[14]</sup>表明边距最大化等价于在构建最优超平面时最小化VC维。计算最好的超平面是一个约束优化问题,并使用二次规划技术解决。判别式超平面由水平集定义,如式1所示:

$$f(x) = \sum_{i=1}^M y_i \alpha_i \cdot k(x, x_i) + b \quad (1)$$

其中, $k(\cdot)$ 是核函数; $f(x)$ 的符号决定了 $x$ 的隶

属度。构造一个最优超平面就相当于找到所有的非零值 $\alpha_i$ 。对应于非零 $\alpha_i$ 的任何向量 $x_i$ 是最优超平面的支持向量。支持向量机的理想特征是保留为支持向量的训练点的数量通常很小,因此提供了一个紧凑的分类器。

对于线性SVM,核函数只是输入空间中的简单点积,而非线性SVM中的核函数通过非线性映射函数有效地将样本投影到更高(可能无限)维度的特征空间: $\Phi: R^N \rightarrow F^N, M \gg N$ 。然后在 $F$ 中构造一个超平面。这种映射背后的动机是它更有可能在高维特征空间中找到线性超平面。使用Mercer定理,将样本投影到高维特征空间中所需的昂贵计算可以用满足条件的更简单的核函数来代替,如式2所示:

$$k(x, x_i) = \Phi(x) \cdot \Phi(x_i) \quad (2)$$

其中, $\Phi(x)$ 是低维向高维空间投影的映射函数; $\cdot$ 表示两个函数做内积运算。

为使不同类之间的分离超平面间隔 $2/\|w\|$ 达到最大,而训练样本之间的误差 $\sum_{i=1}^l \xi_i$ 尽量小,引入惩罚参数 $C$ 。凸二次规划问题可表示为:

$$\min_{\gamma, b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \quad \xi_i \geq 0, i = 1, 2, \dots, m \quad (3)$$

$$y^{(i)} (\omega^T x^{(i)} + b) \geq 1 - \xi_i, i = 1, 2, \dots, m$$

其中, $C$ 是一个常量,当 $C(>0)$ 越大表示对性别判定错误的惩罚越大,越小则对性别判定错误的惩罚越小。

为了获取二次规划问题中的最佳分隔超平面,通过构建一个拉格朗日算子来实现,得到式4:

$$L(\omega, b, \xi, \alpha, r) = \frac{1}{2} \omega^T \omega + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y^{(i)} (\omega^T x^{(i)} + b) - 1 + \xi_i] - \sum_{i=1}^m r_i \xi_i \quad (4)$$

其中, $\alpha_i$ 和 $r_i$ 是拉格朗日乘子。

对式4将其看作是变量 $\omega$ 和 $b$ 的函数,分别对其求偏导,得到 $\omega$ 和 $b$ 的表达式。然后代入式4,求其极大值,最后得到:

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle > 0 \leq \alpha_i \leq C, i = 1, 2, \dots, m$$

$$\sum_{i=1}^m \alpha_i y^{(i)} = 0 \quad (5)$$

其中, $\alpha_1, \alpha_2, \dots, \alpha_m$ 需满足半正定和非负约束的条件。

#### 3.2 核函数及最优参数选择

性别判别的准确率受到核函数选取的直接影响,

文中对 SVM 的各种核函数比较分析后选用径向基函数(radial basis function, RBF)作为颅骨特征映射的核函数。RBF 能够尽可能准确地拟合颅骨数据集上的连续函数<sup>[15]</sup>。数学表达式为:

$$k(\|x - x_i\|) = \exp\left\{-\frac{(\|x - x_i\|)^2}{\delta^2}\right\} \quad (6)$$

其中,  $x_i$  为核函数中心;  $\delta$  为核函数宽度参数, 控制核函数的径向作用范围。

在性别判定的训练阶段, 参数  $C$  和  $\delta$  对性别鉴定的效果影响最大。参数  $C$  的改变能将分类正确的样本和分类错误的样本显著分开。  $C$  越大时分类错误率较小, 但是间隔也较小,  $C$  越小时间隔较大, 但是分类错误率也较大。参数  $\delta$  的改变直接影响核函数的计算能力, 进一步影响性别判定效果。  $\delta$  越大时, 可能会出现误判情况, 即将训练样本或测试样本都划分到同一类别;  $\delta$  越小时, 容易出现过拟合现象, 即能够将训练颅骨样本性别正确分类, 但对测试颅骨样本的分类准确率不高, 泛化能力差。因此, 选取合适的参数  $C$  和  $\delta$  对性别判定效果非常重要。

优化参数  $C$  和  $\delta$  的常用方法有网格搜索法、遗传算法及混沌优化算法等。文中利用文献[16]中的算法来确定合适的参数  $C$  和  $\delta$ 。设定参数  $C$  和  $\delta$  的范围, 即  $2^{-5} \leq C \leq 2^{15}$ ,  $2^{-15} \leq \delta \leq 2^5$ , 步长设为 0.5, 进而获得  $M$  个  $C$  值及  $N$  个  $\delta$  值。利用构造的 SVM 模型对颅骨样本进行分类, 获取性别识别率。根据性别识别率确定最优参数  $C$  和  $\delta$ 。文中利用留一交叉验证法进行测试, 将全部的颅骨样本均分成  $N$  份, 1 份作为测试集, 其余  $N - 1$  份为训练集, 循环  $N$  次进行测试。求解得到所有颅骨样本分类结果的均值, 即对应于  $C$  和  $\delta$  的精确度。重复以上步骤, 最后, 最优参数就是平均识别率最高时所对应的参数值。若最优分类结果仍没达到预想效果, 根据分类准确率变化的趋势, 重新设定  $C$  和  $\delta$  的范围和步长, 直到得到平均识别率最高所对应的参数组合为止。

#### 4 实验结果与分析

实验从 117 个颅骨模型中选择 78 个颅骨(40 男, 38 女)作为训练样本, 采用径向核 SVM 方法建立分类模型, 并用其余的 39 个(20 男, 19 女)颅骨模型作为测试样本, 并进行回代检验。径向核 SVM 的分类步骤如下:

(1) 对样本数据进行归一化处理, 将数据归一化到  $[0, 1]$  之间;

(2) 利用网格搜索和交叉验证方法寻求最优的参数对  $C$  和  $\delta$ , 设定网格搜索的参数为  $2^{-5} \leq C \leq 2^{15}$ ,  $2^{-15} \leq \delta \leq 2^5$ , 搜索步长为 0.5, 可得到 78 个训练样本

下的最优参数  $C = 1.414 \ 2$ ,  $\delta = 0.5$ ;

(3) 对 78 个颅骨训练样本应用 SVM-RFE 算法, 根据特征指标的权重大小对 27 项颅骨特征指标进行排序, 选取前  $n$  个特征为特征集合, 训练 SVM 模型, 分别可得到前  $n$  维特征集合相对应的分类精度, 如图 3 所示。

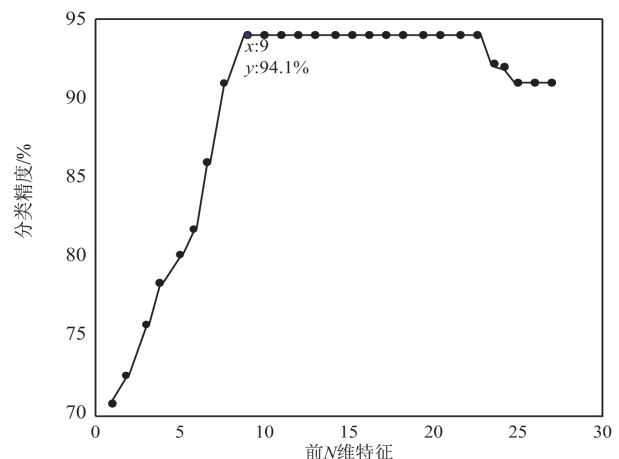


图3 特征子集数目与分类精度的关系

从图3可以看出, 在训练集合上, 分类精度最高可达到 94.1%, 当特征子集数目逐渐从 1 增加到 9 时, 其训练的分类器的分类精度也从 70.9% 逐渐增加到 94.1%; 当特征子集数目逐渐从 9 增加到 22 时, 分类精度在 94.1% 保持稳定; 但是当特征子集数目从 23 再逐渐增加到 27 的过程中, 分类精度开始下降, 由于引入了冗余特征; 最后随着特征子集数目的增加, 分类器分类精度保持在 91.5%。因此, 选取分类精度最高且特征数目最小的前 9 维特征( $X_{25}$ 、 $X_{21}$ 、 $X_5$ 、 $X_{19}$ 、 $X_{23}$ 、 $X_{24}$ 、 $X_{13}$ 、 $X_2$ ) 作为颅骨的最优特征子集。

(4) 根据 9 维最优特征子集训练 SVM 模型, 对 39 例测试颅骨模型进行性别预测, 其预测结果如图 4 所示, 回代检验结果如表 1 所示。

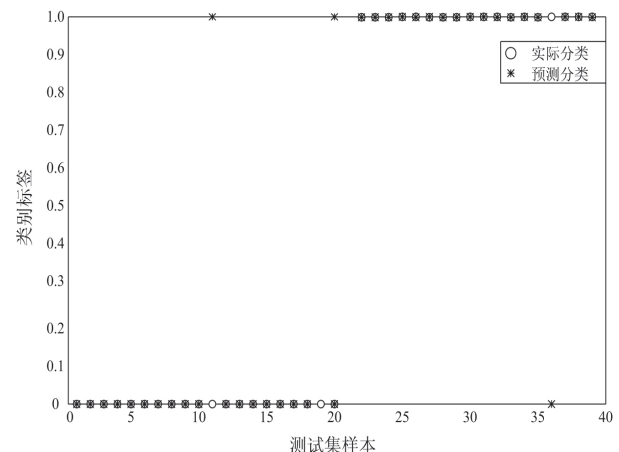


图4 SVM 测试样本预测结果

从图4可以看出, 预测分类效果与实际分类效果存在误差。测试集中的第 11、19 和 36 个样本出现误

判,分类产生了错误。

表 1 SVM 回代检验结果

		男	女	合计
计数	男	18	2	20
	女	1	18	19
准确率	男	90.0	10.0	100.0
	女	5.3	94.7	100.0

从表中可以看出,在 20 个男性颅骨中有 2 个被误判,18 个被正确分类,判定率为 90.0%;在 19 个女性颅骨中有 1 个被误判,18 个被正确分类,判定率为 94.7%。男女判定的平均准确率为 92.4%。

## 5 结束语

针对传统性别识别过程中需要专家参与且依赖于人的主观经验导致分类精度低的问题,提出了一种基于支持向量机的颅骨性别识别方法。根据先验知识和自主开发的系统半自动实现颅骨特征点的定义标定;提取颅骨的可测量特征和非可测量特征,将非可测量特征量化,利用计算机软件测量特征指标;利用 SVM 对特征向量降维并设计分类器,通过网格搜索算法优化参数,得到最佳分类器,实现对目标样本的有效分类。实验结果表明,该方法能够取得较高的分类正确率。由于是首次利用颅骨对象完成维吾尔族颅骨性别识别的研究,所以样本较少,但是方法客观不依赖主观经验,可以为实际应用提供参考依据。下一步将继续对维吾尔族颅骨性别识别进行研究,增加颅骨样本并进一步提高分类精度,为法医人类学、刑侦等领域的实际应用提供更为可靠的参考。

## 参考文献:

- [1] UBELAKER D H, VOLK C G. A test of the phenice method for the estimation of sex [J]. *Journal of Forensic Sciences*, 2002, 47(1): 19-24.
- [2] RAMAMOORTHY B, PAI M M, PRABHU L V, et al. Assessment of craniometric traits in South Indian dry skulls for sex determination [J]. *Journal of Forensic and Legal Medicine*, 2016, 37: 8-14.
- [3] 李明, 范英南, 喻永敏, 等. 西南地区成人面颅骨的性别判定 [J]. *中国法医学杂志*, 2012, 27(2): 132-134.
- [4] KHAITAN T, KABIRAJ A, GINJUPALLY U, et al. Cephalometric analysis for gender determination using maxillary sinus index: a novel dimension in personal identification [J]. *International Journal of Dentistry*, 2017, 2017: 7026796.
- [5] 刘玉勇. 华北地区汉族成人面颅骨 X 线片性别判定的研究 [J]. *中国司法鉴定*, 2016(1): 26-31.
- [6] AFRIANTY I, NASIEN D, KADIR M R A, et al. Back-propagation neural network for gender determination in forensic anthropology [J]. *Studies in Computational Intelligence*, 2015, 575: 255-281.
- [7] LUO Li, CHANG Liang, LIU Rong, et al. Morphological investigations of skulls for sex determination based on sparse principal component analysis [C]//Chinese conference on biometric recognition. [s.l.]: Springer, 2013: 449-456.
- [8] MARANI R, RENÒ V, NITTI M, et al. A modified iterative closest point algorithm for 3D point cloud registration [J]. *Computer-Aided Civil and Infrastructure Engineering*, 2016, 31(7): 515-534.
- [9] 税午阳, 周明全, 武仲科, 等. 数据配准的颅骨面貌复原方法 [J]. *计算机辅助设计与图形学学报*, 2011, 23(4): 607-614.
- [10] MAVRIDIS P, ANDREADIS A, PAPAIOANNOU G. Efficient sparse ICP [J]. *Computer Aided Geometric Design*, 2015, 35-36: 16-26.
- [11] LIN Chenghe, JIAO Benzhen, LIU Shanshan, et al. Sex determination from the mandibular ramus flexure of Koreans by discrimination function analysis using three-dimensional mandible models [J]. *Forensic Science International*, 2014, 236: 191.e1-191.e6.
- [12] 李春彪, 孙尔玉. 应用 Fourier 变换对东北地区成人颅骨性别差异的研究 [J]. *人类学学报*, 1992, 11(4): 312-318.
- [13] 业巧林. 若干 SVM 算法的改进与设计 [D]. 南京: 南京林业大学, 2009.
- [14] VAPNIK V N. The nature of statistical learning theory [M]. Berlin: Springer, 2000: 988-999.
- [15] 田中金. 基于分区变形颅面复原算法的研究与实现 [D]. 西安: 西北大学, 2011.
- [16] 王兴玲, 李占斌. 基于网格搜索的支持向量机核函数参数的确定 [J]. *中国海洋大学学报: 自然科学版*, 2005, 35(5): 859-862.