第33卷第6期 北京交通大学学报 Vol.33 No.6
2009年12月 JOURNAL OF BEIJING JIAOTONG UNIVERSITY Dec. 2009

文章编号:1673-0291(2009)06-0132-05

程序树层次化结构统计模型及MOSES改进算法

闻凌云[a],刘贵全[a],赵英海[b]

(中国科学技术大学 a.计算机科学技术系;b.电子工程与信息科学系,合肥 230027)

摘 要:为提高MOSES效率,提出了一种新的程序树层次化结构统计模型.该模型通过统计分析同类群,自动发现子树特征来指导优化.该模型不需要hBOA算法那样对变量集合进行建模,也不需要像MRTS算法那样遍历小规模的种群来发现潜在的有指导意义的子树.通过解决人工蚂蚁问题对算法进行了测试,结果表明改进后的MOSES算法更加高效.

关键词:自主程序演化;MOSES(语义进化搜索优化);子树;人工蚂蚁问题

中图分类号:TP181 **文献标志码**:A

Hierarchical Statistical Structure Model of Program Trees and MOSES Algorithm Improvement

WEN Lingyun[a], *LIU Guiquan*[a], *ZHAO Yinghai*[b]

(a. Department of Computer Science and Technology; b. Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230027, China)

Abstract: To improve the efficiency of MOSES algorithm, this paper proposes a new hierarchical statistical model of program trees. This model conducts hierarchical statistical analysis on program trees and can generate potential subtrees automatically to guide algorithm optimization. This model leaves out the operations of creating models for the variables set like the previous hBOA algorithm; and also doesn't need the tedious operations to traversal small population to find certain superior individuals as subtrees like the MRTS method. Experimental results on solving artificial ant problem indicate that our proposed algorithm is more effective and efficient than the previous hBOA-based MOSES.

Key words: competent programming evolution; meta-optimizing semantic evolutionary search (MOESES); subtree; artificial ant problem

2006年Moshe Looks在他的博士论文competent programming evolution[1]中提出了MOSES算法(Meta-Optimizing Semantic Evolutionary Search, MOSES).MOSES算法是遗传规划或遗传程序设计(Genetic Programming, GP)[2]的一个分支,运用了GP中关于程序树,种群等概念.MOSES算法通过程序树标准化操作来去除程序树种群的语义冗余,通过建立框架程序和同类群,并结合hBOA进行优化.MOSES已成功应用于计算生物,情绪评估和agent控制等[3-4]多个领域.自主程序演化(competent programming evolution)及MOSES算法,已成为当前国际研究的热点.

相比其他的程序演化系统,MOSES更加精确,而且要求的人工控制参数更少.但是MOSES借助层次贝叶斯优化算法(hierarchical Bayesian Optimization Algorithm, hBOA)[5]对程序树进行优化,由于hBOA算法计算量大,且容易陷入局部极值,导致MOSES的整体效率有待提高.本文作者提出了程

收稿日期:2009-03-18

作者简介:闻凌云(1981—),女,山东日照人,硕士生.**email**:wlingyun@mail.ustc.edu.cn.
刘贵全(1970—),男,四川彭由人,副教授,博士.

序树层次化结构统计模型并对 MOSES 算法进行了改进.该模型通过统计分析,自动发现子树特征用于指导优化.实验表明,基于程序树层次化结构统计模型的 MOSES 改进算法更加高效.

1　基于 hBOA 的 MOSES 框架简介

1.1　MOSES 算法流程

为方便数学表述,我们首先对 MOSES 算法的相关概念表示如下:程序树字符集为 K,K' 代表程序树扩展字符集;T_i 代表框架程序的一个节点值,T' 为框架程序树(knob,简称为框架程序);同类群(deme) D 为具有一系列共同的框架程序 T' 的程序树的集合,如图 1 中例子所示.

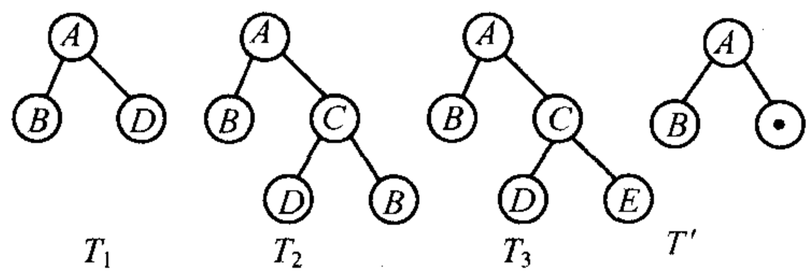

图 1　同类群及框架程序示例

Fig.1　An example of deme and its knob

MOSES 算法通过面向问题的领域专家知识生成语义限制规则.然后利用语义规则生成关于问题解决的初始程序,并用它来生成随机程序样本,形成初始种群.如果在优化过程中生成了新的问题解决模式,也将其抽取成一个新的框架程序,并生成新的随机程序树.

MOSES 算法具体流程如下.

1)构建初始的框架程序集 S,随机实例化框架程序中的通配符来产生初始的同类群,并将其加入种群集合 G.

2)从 G 中任意选择一个同类群 D 进行优化:①从 D 中随机选择一些程序构成样本集 Y;②对 Y 中的程序使用 hBOA 优化算法产生适应度更高的程序,取代 D 中较差的程序.

3)若上一步得到的同类群可以抽取出新的框架程序,并且实例化后能产生适应度更高的同类群,则进行生成和替换.

4)重复第 2)步,直至达到结束条件.

1.2　hBOA 概述

hBOA 包含贝叶斯优化算法(BOA)、贝叶斯网络的本地结构和小生境(niching)3 个方面的技术.该算法使用决策树或决策图结构来模拟各变量之间的相关性,这种相关性反映了模式理论中的模式[6].层次贝叶斯优化就是在这 3 种机制作用下能够得到多种贝叶斯网络模型.

BOA 是一个有向无环图,用它来表示一组变量之间概率关系.BOA 通过选择一部分优良个体(样本数据)来构建贝叶斯网络模型.对解集建立贝叶斯网络模型的本质是构建一个有向图,其中每一个节点代表一个变量,节点到节点的有向连接描述了节点间的条件概率关系.因此在已知解集的变量取值概率分布情况及多个变量之间取值的概率关联关系的条件下,可以采用贝叶斯网络模型进行描述.层次贝叶斯网络的本地贝叶斯网络结构使用决策树或决策图结构,能够允许更小的模型构造步骤,并导致更精确的模型.

小生境是生物学中的术语,在自主程序进化算法中引入小生境的思想,使得求解多峰值函数的优化问题时,不仅得到的是局部最优解,而是尽量希望能找到问题的较大范围最优解.

综上所述,在传统 MOSES 框架中,hBOA 算法采用图结构对变量集进行建模,一方面,该建模过程目前无法进行语义解释和预测,很难利用特定的先验知识对其进行约束和指导.因此在问题处理中 hBOA 存在着很多不必要的背景计算,这严重影响算法的速度.另一方面,hBOA 算法尽管采用了小生境技术,但是仍然存在容易陷入局部极值的问题,进一步增加了 MOSES 搜索代价.

当前的研究重点也在于如何使用恰当的优化方法,达到速度和效益的较好的结合点.我们正是通过研究文献[7]中对于程序树优化的搜索空间的分析,结合自动生成子树[8]的思想提出了程序树层次化结构统计模型.

2　程序树层次化结构统计模型

针对 MOSES 特点,利用程序树层次化结构统计模型对同类群进行层次化结构统计,发现有意义的子树,指导同类群的进化方向,具体过程见图 2.

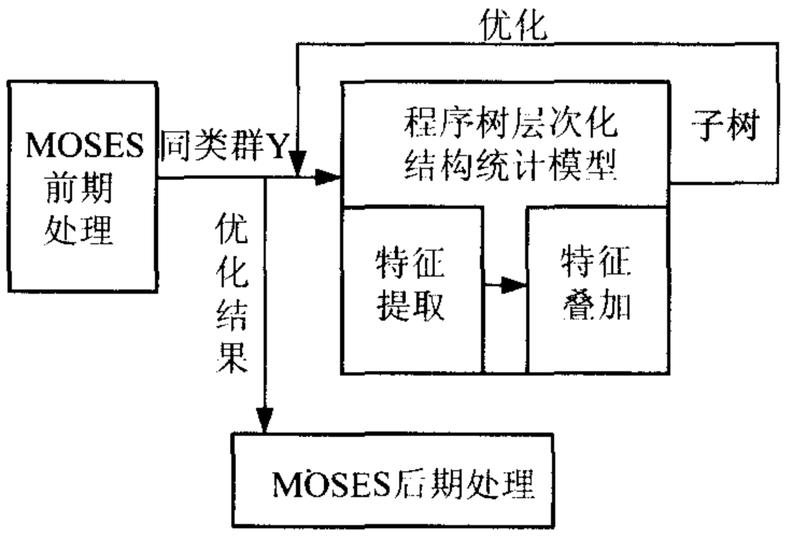

图 2　程序树层次化结构统计模型结构及优化流程图

Fig.2　Flowchart of optimization of hierarchical statistical structure model of program trees

对于 MOSES 产生的待优化的同类群 Y,模型通过特征提取和特征叠加两个步骤可以获得多个子

树,并将这些子树用于对同类群 Y 进行优化. 反复进行上述过程,直至达到结束条件,最后将优化后的同类群进行 MOSES 的后续处理.

2.1 统计模型

程序树层次化结构统计模型用于统计和发现同类群的优秀子树,表达为与程序树存在映射关系的一组层次化的表.若同类群的最高高度为 $h_{\max}$, $k=\{A_1, A_2, \cdots, A_n\}$,则对于每个正整数 $l_{\text{lev}} \leqslant h_{\max}-1$ 可以构造唯一一个表.形式如图 3 所示.若程序树字符集 k 包含 n 个元素,则每个表包含 n^2 个单元,每个单元含有 3 个变量,表示为 cell(l_{lev}, p_{par}, c_{chi}),如图 3 中 A_{ij} 节点表示为 cell(l_{lev}, A_j, A_i). 表明该网格节点对应的所有程序树节点在树中的高度为 l_{lev},本身的值为 A_i,父节点的值为 A_j.程序树的每个节点(除了根节点)对应于唯一的单元节点.

$H=l_{\text{lev}}$	A_1	A_2	…	A_j	…	A_n
A_1						
A_2						
⋮						
A_i				A_{ij}		
⋮						
A_n						

图 3 层次数为 l_{lev}的结构示意图

Fig.3 A structure example of table at level l_{lev}

在模型建立的过程中,同类群中的程序树按照适应度被分为 good 和 bad 两部分,要求 good 部分的样本数略高于 bad 部分.这一点可以保证好的子树特征不会被埋没.如果相对应的节点属于适应度好的树,则 good 计数加 1;否则,bad 计数加 1.

图 1 中所示的同类群对应的程序树层次化结构统计模型包含 2 个 table,每个 table 又包含 n^2 个单元.将同类群 $T_1(A2B1D1)$, $T_2(A3B2C2D3B3)$, $T_3(A3B2C2D1E1)$(节点值后面的数字代表该节点在树中的高度)映射至模型中,其中 $T_1 \in$ bad, T_2、$T_3 \in$ good. $H=1$ 的 table 如图 4 所示.分别用“b”和“g”表示网格节点的 bad 和 good 计数.

$H=1$	A	B	C	D	E
A					
B	b		g		
C					
D	b		gg		
E			g		

图 4 层次为 1 的示意图

Fig.4 An example of table at level 1

2.2 特征提取过程

对 table 进行遍历,对于每个 $g>b$ 的单元,生成了一棵候选子树,单元的 p_{par}值为此候选子树根节点的值,c_{chi}值为根节点的一个孩子的值.遍历图 3 和图 4 得到了 5 个候选子树,并确定了其根节点的值及根节点其中一个孩子的值,它们分别为:h_1&($C2B1$), h_1&($C2D1$), h_1&($C2E1$), h_2&($A2B1$), h_2&($A2C1$),其中标记 h_i 表示根节点的孩子原来位于程序树的 l_{lev}为 i,从 $H=i$ 的表中获得.候选子树的每个孩子节点,都要记录父节点的相关信息,便于后续操作.

通过模型选择出来的候选子树统计发现了同类群中适应度较好部分个体共同拥有的结构.通过特征提取过程就能够得到用于指导优化的子树个体.

2.3 特征叠加过程

通过问题的先验知识可以确定程序树字符集 K 中的终端节点、非终端节点及每个非终端节点对应的孩子数目,然后依次从根节点广度优先遍历候选子树.

对每一个非终端节点 $N \in h_i$&(…),若 N 孩子数不满足应有的孩子数,则首先在 $h=i$ 的候选子树中寻找根节点值与 N 相同但是同样不饱和的候选子树:如果找到,则将找到的候选子树的根节点的子树作为 N 的子树合并入本子树中(注意若 N 达到饱和即停止合并).合并操作后如果 N 饱和则将本子树的 h 标记减 1;否则在 $H=i$ 的表中寻找该节点可能的孩子.

1)如果存在 $g>b$ 的节点,选择 g 最大的节点作为孩子节点,并将以此节点作为根节点的候选子树合并入本子树中.

2)如果存在 $g=b$ 的节点,选择 g 最大的终端节点作为孩子节点.

3)否则选择 $g=0$ 并且 $b=0$ 的终端节点作为孩子节点.

以上 3 种情况均需满足特定问题的语义限制条件.同样的,若该节点孩子数补齐则将该候选子树的 h 标记减 1.

对得到的候选子树进行特征叠加,其中一种可能的结果为($A3B2C2B1D1$)和($C2B1D1$)这两棵子树.此时其 h 标记均为 0.

为保证子树的数目,特别将种群当前最优个体并随机挑选该最优个体的子树作为额外的补充.

获得子树之后,进行如下操作:随机选择一些程序树样本,将获得的候选子树随机替换掉程序树中的一个子树分支,若所得新树适应度变大,则随机保留

新树或者替换掉原来的树,否则以随优化次数增加而逐渐变小的概率接受该树.直到获得满足指标的新树,或者处理完规定数目的程序树.随后若群体数量多于原来的数量,则去除适应度最小的样本,直至达到原来群体数量.

若达到最优结果,成功,终止 MOSES;否则,重复上述操作,直至达到结束条件.在这个过程中起到优化作用的子树将在同类群中被保留下来.

3　人工蚂蚁问题实验比较

3.1　问题描述

图 5 是一张大小为 32 × 32 人工蚂蚁地图,在其上以"#"代表食物,随机布置了 89 个食物.蚂蚁初始位置为[0　0],面向图中右边.从箭头指向开始,寻找食物.蚂蚁的动作可以分为(move)前进一格、(left)左转、(right)右转 3 种,其中每个动作耗费一个时间片.程序所要实现的是:在 600 个时间片内尽可能多的在地图上遍历更多的食物[2],寻找到全部的 89 个食物即为成功的解决了问题.该问题的适应度即在 600 个时间片内吃掉的食物总数.人工蚂蚁问题是一个成熟的检验智能方法的问题.其搜索空间也在文献[7]中被很好的研究.

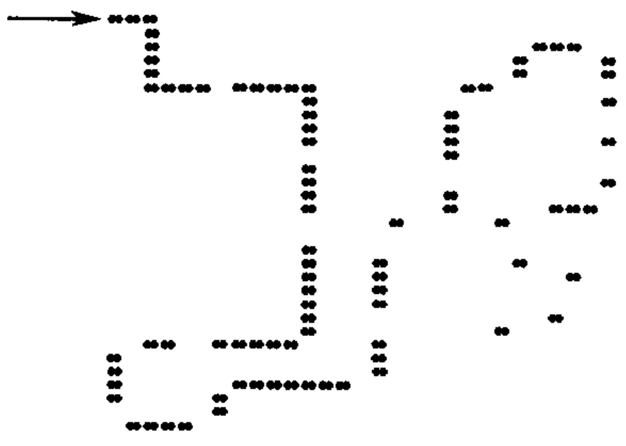

图 5　人工蚂蚁问题示意图

Fig.5　Artificial ant problem example

3.2　实验对比

通过解决人工蚂蚁问题,对 MOSES 基于我们提出的程序树层次化结构统计模型与原来基于 hBOA 的算法进行了性能比较,并与 MRTS(MEMORIZING-RANDOM-TREE-SEARCH)也进行了比较[9].每种算法运行 100 次(注:3 种方法每次运行均得到正确解),记录适应度计算次数的上限值.其中基于 hBOA 的 MOSES 及 MRTS 的种群数在运算过程中自动增加,基于本文模型的种群数为 100,适应度计算次数上限统计实验结果为:hBOA 为 23 000 次;MRTS 为 20 696;改进后的 MOSES 为 13 400.

同时记录了原来的 MOSES 算法和本文改进后算法 100 次运行中的代价分布图,分别如图 6 和图 7 所示.

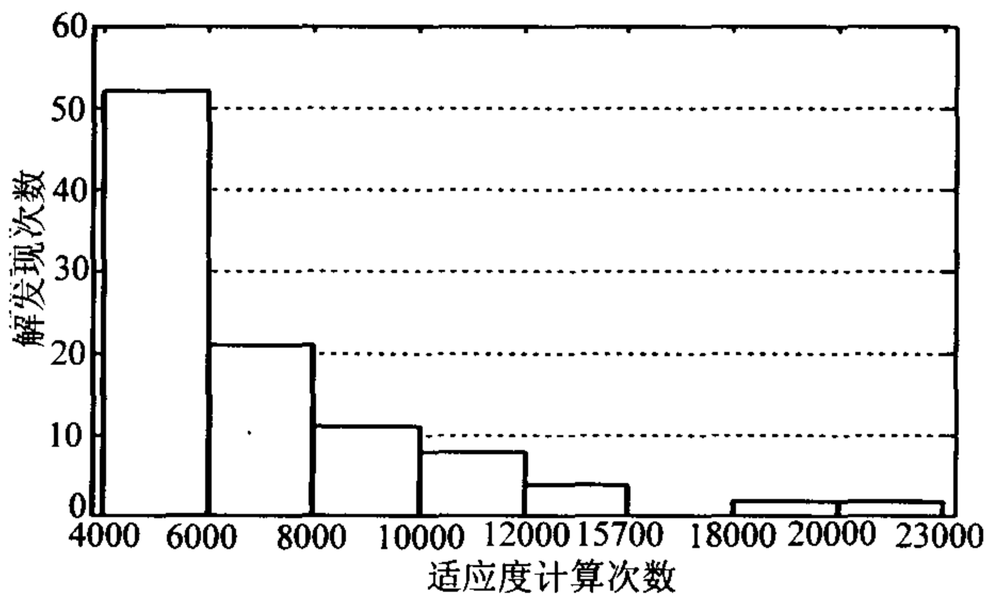

图 6　原 MOSES 100 次运行代价分布图

Fig.6　Cost results of 100 times of running of old MOSES

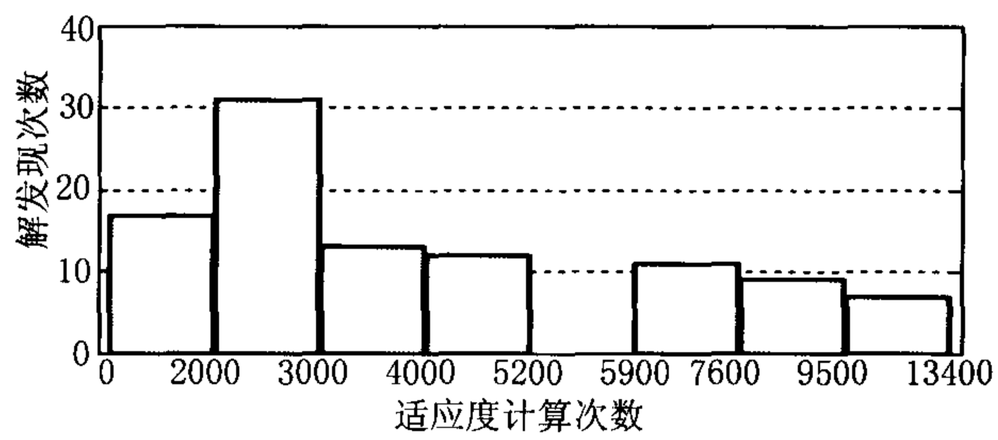

图 7　改进后 MOSES 算法 100 次运行代价分布图

Fig.7　Cost results of 100 times of running of our improved algorithm

3.3　实验分析

通过实验数据的比较可以得出这样的结论:利用子树指导优化的算法,即 MRTS 和本文改进的 MOSES 算法,比 hBOA 这种直接优化的方法更加高效.通过图 6 和图 7 运行代价分布图的比较,可以得出改进后算法比原算法运算效率提高约 50% 的结论.

MRTS 算法与本文提出的模型都是基于发现优秀子树的.但是 MRTS 算法针对节点较少的个体进行研究,寻找其中最优的个体作为候选子树.需要对所有设定规模内的程序树个体全部进行研究.并且没有 MOSES 算法中对程序树进行标准化的处理,会导致无法分辨出大量的具有相同语义但语法不同的冗余树,进一步增加了运算量.

本文提出的改进 MOSES 算法与 MRTS 算法相比具有更大的灵活性和普遍性,能够发现一些潜在的并不直观的特征.MRTS 不能发现某些规模很小、作为独立个体适应度不高但在优秀个体中出现频率很高的子树.然而这些小型子树却是很有意义的.通过文献[7]中 Langdon 和 Poli 对于人工蚂蚁问题搜索空间的分析表明尺度较小的树比尺度更大的树更有可能进化为该问题的解.小规模子树的出现有可能缩小程序树的尺度,在进化的轨迹上具有重要的意义.

虽然本文提出的程序树层次化结构统计模型只

是一个二维模型,但是优选子树的方法减小了有价值的子树被破坏的几率.而且通过适应度的选择,真正优秀的子树被保留了下来,获得了更多生存和发展的机会.该模型通过这样的迭代优化操作,指导群体向适应度更高的方向进化.

4 结束语

根据 MOSES 算法的特点,提出了基于程序树层次化结构统计模型的改进算法.该算法通过建立程序树层次化结构统计模型对同类群进行层次化结构统计,发现有意义的子树,指导群体向更好的方向发展.通过解决人工蚂蚁问题验证了经本文模型改进的算法的效率提高了约 50%,且改进后的算法也优于直接寻找优秀子树的 MRTS 算法.

参考文献:

[1] Moshe Looks. Competent Program Evolution[D]. Washington: Washington University in St. Louis, 2006.

[2] John R Koza. Genetic Programming: On the Programming of Computers by Means of Natural Selection[M]. Cambrid Ge, MA: MIT Press, 1992:1 - 161.

[3] Moshe Looks, Ben Goertzel, L'ucio de Souza Coelho. Clustering Gene Expression Data via Mining Ensembles of Classification Rules Evolved Using MOSES[C]// Genetic and Evolutionary Computation Conference (GECCO), 2007:407 - 411.

[4] Moshe Looks, Ben Goertzel, L'ucio de Souza Coelho. Understanding Microarray Data through Applying Competent Program Evolution[C]// Genetic and Evolutionary Computation Conference (GECCO), 2007:430.

[5] Martin Pelikan, Martin Pelikan, David E Goldberg. Escaping Hierarchical Traps with Competent Genetic Algorithms[C]// Proceedings of Genetic and Evolutionary Computation Conference (GECCO2001), 2001:511 - 518.

[6] Martin Pelikan. Bayesian Optimization Algorithm: From Single Level to Hierarchy[R]. Doctoral Dissertation, University of Illinois at Urbana-Champaign, Urbana, IL. Also IlliGAL Report No. 2002023, 2002.

[7] Langdon W B, Poli R. Why Ants Are Hard. Genetic Programming[R]. Proceedings of the Third Annual Conference. Morgan Kaufmann, Madison, 1998:193 - 201.

[8] Lawrence Davis. Genetic Algorithms and Simulated Annealing [M]. Los Altos, CA: Morgan Kaufmann Publishers, 1987.

[9] Steffen Christensen, Franz Oppacher. Solving the Artificial Ant on the Santa Fe Trail Problem in 20,696 Fitness Evaluations[C]// Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation (GECCO2007), 2007: 1574 - 1579.

(上接第 131 页)

5 结语

本文作者提出了一种基于血流图和双树复数小波域傅里叶变换的方法,把易受环境影响的温谱图转换成血流图,然后进行双树复数小波变换,由于其同时具有方向的信息且信息冗余量小,算法复杂度低,进而提高了识别效率.通过对血流模型的简化,可以在几乎不降低识别率的条件下,进一步减小算法的时间复杂度,因此更接近实际的应用.

参考文献:

[1] Chen Weilong, Er Meng Joo, Wu Shiqian. PCA and LDA in DCT domain[J]. Pattern Recognition Letters, 2005,26 (15):2474 - 2482.

[2] Prokoski F J. Method and Apparatus for Recognizing and Classifying Individuals Based on Minutiae: US Patent, 6173068B1[P]. 2001.

[3] Wu S Q, Song W, Jiang L J, et al. Infrared Face Recognition by Using Blood Perfusion Data[C]// Proc. Audio- and Video-based Biometric Person Authentication. NY, July, 2005 (AVBPA05): 320 - 328.

[4] 赖剑煌,阮邦志,冯国灿.频谱脸:一种基于小波变换和Fourier变换的人像识别方法[J].中国图像图形学报,1999,4(10):811 - 816.
LAI Jianhuang, RUAN Bangzhi, FENG Guocan. Spectroface: A Face Recognition Method Based on Wavelet Transform and Fourier Transform[J]. Journal of Image and Graphics, 1999, 4(10): 811 - 816 (in Chinese)

[5] 胡海平.小波图像压缩编码与小波图像阈值降噪的研究[M].上海:上海大学出版社,2002.
HU Haiping. Research on Image Compressing and Coding and Threshold Denoising Based on Wavelet[M]. Shanghai: Shanghai University Press,2002.(in Chinese)

[6] Hatipoglu S, Mitra S K, Kingsbury N. Texture Classification Using Dual-Tree Complex Wavelet Transform[C]// 7th International Conference on Image Processing and Its Applications. Manchester, UK, 13 - 15 July, 1999: 344 - 347.