第60卷 第6期 天 文 学 报 Vol.60 No.6
2019年11月 ACTA ASTRONOMICA SINICA Nov., 2019

doi: 10.15940/j.cnki.0001-5245.2019.06.001

基于图像的天体搜索研究与自动化集成软件开发*

陆 扬[1†] 安 涛[1,2] 郭绍光[1‡] 劳保强[1]

(1 中国科学院上海天文台 上海 200030)

(2 中国科学院射电天文重点实验室 南京 210033)

摘要 天体搜索是天文数据处理流程的一个重要环节, 也是以平方公里阵列射电望远镜(Square Kilometre Array, SKA)为代表的下一代射电望远镜在面向海量数据处理中的挑战之一. 现今天体自动搜索算法、软件已日趋成熟并投入应用, 不过在自动化、兼容性等方面仍具有提升空间. 以更自动化、更适应海量数据需求的天体搜索算法研究为宗旨, 以现有算法为研究基础, 天体自动搜索软件系统得到设计和开发. 该系统包含友好的交互式用户操作界面, 具备可视化输出数据显示、兼容不同数据输入和输出并包含为实际应用服务的文件管理功能. 该系统对于大天区图以及图像集, 均能够很好地进行自动化处理. 测试结果显示, 上述方法对于天体搜索的改进有一定成效. 后续将在此基础上对该集成系统做进一步的改进开发, 以适应更多的需求.

关键词 **技术**: **图像处理**, **方法**: **分析**, **星表**

中图分类号: P161; **文献标识码**: A

1 引言

国际大科学工程平方公里阵列射电望远镜(Square Kilometre Array, SKA)是21世纪最具有远见与雄心的科学工程之一, 无论从规模还是数据量来看都是目前人类兴建的最大(射电)天文观测设备, 建成后它将是世界上最大的综合孔径射电望远镜. SKA由分布在3000 km基线内数量庞大的天线组成, 有效接收面积达1平方公里[1]. SKA将产生相当庞大的数据, 按照目前的估计, 平方公里阵列射电望远镜第1阶段(Square Kilometre Array Phase 1, SKA1)每秒将产生TB量级的数据[2–3]. 这对现代科学技术提出了一个跨越式挑战, 面对如此量级的数据, 使用传统算法、软件已无法满足SKA的海量数据处理的要求, 自动化的分析处理将是解决这类问题的必然选择.

2019-02-25收到原稿, 2019-07-11收到修改稿

*国家重点研发计划(2018YFA0404603), 国家自然科学基金项目(11873079、11703069), 国家自然科学基金委员会-中国科学院天文联合基金项目(U1831204)资助

†ylu@shao.ac.cn

‡sgguo@shao.ac.cn

在SKA正式启动前, 在全球范围已逐步建成、更新了4个SKA先导项目和多个探路者项目[1], 其中先导项目包括位于澳大利亚的SKA先导项目(Australian Square Kilometre Array Pathfinder, ASKAP)、默奇森宽视场阵列(Murchison Widefield Array, MWA)、南非台地高原阵列望远镜(MeerKAT射电望远镜阵列)和氢原子再电离时代阵列(Hydrogen Epoch of Reionization Array, HERA), 探路者项目包括中国的宇宙第一缕曙光探测计划−21 cm射电望远镜阵(The 21 Centimeter Array, 21CMA)、欧洲的低频射电阵列(The Low Frequency Radio Array, LOFAR)以及美国的长波长阵列(The Long Wavelength Array, LWA)等, 它们将为SKA望远镜的启用与科学产出提供重要的先验指导. 这些项目以及即将开建的SKA1和全规模的SKA产生的海量规模的数据, 对于数据处理、数据分析、天体搜索等各方面都提出了迫切和更高的要求.

天体搜索通过对天文图像中的星体进行搜索查找, 基于相关算法进行拟合, 然后形成星表[4], 该工作是大型巡天项目的基础. 天体搜索技术在SKA先导望远镜的多个关键科学项目和数据处理工作中均得到应用. 天体搜索对于未来SKA的射电天文数据处理也起着十分重要的作用, 不仅将为多个数据处理流程提供更精确的天空模型, 也将为大规模数据库的建立、数据处理及可视化分析提供相应星表基础[2]. 它的精确度与效率也将在很大程度上影响到数据处理过程以及最终成图等产品. 作为多个SKA科学数据处理管线的起始环节, 也是连续谱成像管线系统、谱线管线系统、科学分析管线系统以及快速成像管线系统的重要组成部分, 天体搜索为建立天空模型提供了关键输入参数. 此外, 天体搜索也为数据库提供了压缩数据产品[5], 搜寻及识别的天体数据也将用于后续的多波段交叉证认、天体分类、光度函数等统计分析.

大多数传统的天体搜索算法、软件在完备度与可靠度方面已达到较高水平[1], 一系列天体自动搜索算法、软件也应运而生, 但在数据处理前后仍需要人工介入进行修正, 在某些程度上需要对天体数据进行后续的人工修正和添加. SKA的数据量比先导望远镜高出几个量级, 这就需要有更适应于未来望远镜设备的数据处理方法, 能有效快速地处理海量的数据, 具备更少的人工干预、更自动化的数据处理、更精确的处理结果等特性[6]. 因此具有高度自动化、可扩展、高准确度、可靠度与完备度的天体搜索算法、软件对于更大数量天体搜索和拟合是不可或缺的.

天体搜索技术发展至今, 已形成了一系列更适应于天文大科学工程的算法、软件, 近年来所开发的一些算法、软件也已被用于SKA先导项目的运行中, 并得到持续的改进. 但迄今, 大多数算法、软件仍然是高度工程化的. 它们能够很好地被系统工程师所用, 作为软件模块和中间算法被嵌套在数据处理系统流程中, 但作为独立软件被科学工作者所用, 并应用于科学目标研究方面仍有着很大的改进空间. 首先, 当今的天文大数据时代, 天体搜索算法无论是作为中间环节还是独立软件, 在面对海量文件和数据时, 迫切需要具备批量处理、自动文件归档等功能及文件管理系统. 而对于大多数软件, 尚需要通过手动操作实现多文件的处理及归档, 即使对于一些具有用户界面的天体搜索软件, 也主要提供单个图像文件的处理. 此外, 现今大多数软件的操作, 包括软件的运行、参数的选择、输入及输出文件的选择等均需通过命令行实现, 这便要求用户首先对于操

[1]https://www.skatelescope.org/precursors-pathfinders-design-studies
[2]https://www.ska-sdp.org

作环境以及每一个所使用的软件较熟悉, 而不同软件的操作又不尽相同. 这给用户使用增添了难度. 因此, 一套适用性强的整合软件系统可以有效解决这个问题. 出于以上各方面的考虑, 基于目前已有的天体自动搜索算法, 通过对算法的改进、批量文件处理和查看等功能的整合, 形成了一套具有交互式用户界面的集成软件, 从而满足更广泛的需求. 本文第2节将简要概述现有的天体自动搜索算法、软件, 第3节将对天体自动搜索集成软件设计及实现进行探讨, 第4节将进行实验测试分析, 第5节将对海量数据处理的挑战进行探讨, 最后对本文进行总结.

2 天体自动搜索算法概述

2.1 方法概述

天体自动搜索自上世纪70年代起发展至今[7], 已产生一系列相关算法以及基于算法形成的软件, 技术也愈加成熟, 不同算法和软件各有其自身的开发背景、侧重领域及优劣.

早期的软件有面向光学图像开发的SExtractor (SE)[8], 适用于大规模巡天数据. 该软件在射电天文领域也得到了相关应用[9], 主要应用在对天体进行搜索、测算及分类上, 搜索方法基于阈值转换Lutz算法[10]. 面向射电波段天体搜索的有Blobcat[11]、Selavy[12]和Duchamp[13]软件及Sfind[14]、Aegean[3][9, 15]算法等. 其中, Duchamp最初面向HI观测数据而开发, 其应用也已扩展到其他波段的天文图像中, 主要用于搜索并绘制谱线数据立方体图像. Selavy是Duchamp的另一个版本, 用在SKA先导项目ASKAP的管线系统的ASKAPsoft架构中, 可处理谱线数据立方体与连续谱图像. Blobcat主要用于处理Stokes I和线偏振射电图像. Blobcat与Aegean均采用了泛洪填充算法(Flood-Fill)[16]进行天体搜索形成像素岛, 但使用不同方法在像素岛中进行天体拟合.

从数据处理的方法及过程的角度来看, 射电天体搜索过程大致包含背景估计与消除、亮源识别、亮源拟合和生成星表等[9].

背景估计是通过对背景与噪声的特性分析, 设立阈值, 将天体与背景进行区分的过程[17], 针对有、无结构背景情况, 分别采用图像滤波器、阈值处理等方法[9]. 在设定阈值的方法中, 经分析比较, 应用错误发现率(False detection rate, FDR)[14]方法能够使算法更完备、可靠[4].

亮源识别过程则采用例如Lutz及Flood-Fill等算法将区分为天体的像素形成像素岛[9]. 这些算法被应用于Duchamp、Aegean和Blobcat等算法和软件中[18].

亮源拟合过程便基于像素岛, 对每一个天体属性进行测算, 根据天体的性质及观测条件, 采用单高斯或多高斯天体进行拟合. 天体区分与拟合的过程, 在一些已开发的天体搜索算法和软件如SExtractor、Selavy、Sfind和Image Search and Destroy (IMSAD)中存在很大差异, 进而会影响处理变星、暂现源等天体的效果[4]. Aegean和Blobcat等算法、软件对此做了改进. Aegean在这一过程中, 采用拉普拉斯核, 更准确地判断天体数量, 设定初始参数, 进行天体拟合. 完成搜索拟合的天体信息将进一步形成星表.

下面将针对在天体搜索过程中涉及的两种上文提到的算法进行更详细的阐述.

[3]https://github.com/PaulHancock/Aegean

2.1.1 FDR算法

在射电图像的处理中, 背景估计一直是比较重要的一步. 在各种处理算法中, 阈值的设定是最重要的一步. 阈值估计得过高会导致射电点源的丢失, 阈值估计得过低就会导致噪声被识别为射电源. 在射电源的阈值选择中, 既可以通过设置为像素的5σ或者像素岛的3σ, 也可以使用FDR算法来自由设定.

FDR算法是一种统计方法, 最早由Benjamini和Hochberg提出[19], 主要用于完善多重检验问题的假设测试, 通过控制FDR来决定P值的阈值. FDR可以灵活调整期望值, 作为阈值的指标, 不同于总体错误率(Family-wise Error Rate, FWER), FWER一般固定设置为0.05. 与FWER相比, FDR采用了更为宽松的标准, 在初始假定都满足的情况下正确率与FWER相当, 其他情况下优于FWER算法[20].

2.1.2 Flood-Fill算法

Flood-Fill算法用在点源搜索的第2个阶段, 即亮源识别, 用于区分前景源与背景噪声, 并将识别的像素点合成为一个射电源. Flood-Fill算法是从一个区域中提取若干个连通的点与其他相邻区域区分开的经典算法. 因为其思路类似洪水从一个区域扩散到所有能到达的区域而得名.

Flood-Fill算法选取两个阈值, σ_s和σ_f, 其中大于σ_s的认为是亮源的一部分, 大于σ_f的认为是亮源关联的判断值, 用于将多个像素点组合为一个亮源.

这里我们使用4路算法(不考虑对角线方向的节点)进行模拟, 使用深度优先的递归方法对9×9个像素点进行处理. 我们的模拟结果如图1所示, 图中黑色区域的值被认为其亮度值大于灰色, 可以看到该模拟算法将从起始点开始的黑色连接块组合为一个像素岛(白色区域).

2.2 软件优势比较

现今各天体自动搜索算法、软件都发展到了一定高度, 并因不同的开发背景及特点, 在不同的应用场景中发挥其各自的优势作用.

Aegean是针对ASKAP开发的天体自动搜索算法, 适用于射电天文领域的数据处理, 并对于SKA的连续谱巡天图像数据处理所面临的问题有着针对性的改进.

Flood-Fill算法作为其亮源识别过程的核心算法, 形成像素岛, 对于采用最小二乘法椭圆高斯拟合的算法和软件, 例如Multichannel Image Reconstruction, Image Analysis and Display (MIRIAD)软件包, Flood-Fill算法帮助解决了这些算法在参数无约束情况下所需的大量人工介入修正问题[11], 提供更高的准确性与自动化程度, 减少了人工介入, 从而更适应于大规模巡天的数据处理.

一些现有天体自动搜索算法或软件对于像素岛中多天体的拟合存在一定的误差, 而在对一个像素岛的天体进行拟合前, 率先预测出天体数目, 对提高拟合的准确度有很大的帮助[9]. 在Aegean算法中, 使用了拉普拉斯转换算法从输入图像中获得曲率图, 结合经过阈值处理的图像, 能更准确地估计每个像素岛中的天体个数, 并确定初始拟合参数, 从而在多天体拟合的准确性方面有了很大的提高.

经过与IMSAD、Selavy、SExtractor和Sfind的比较, Aegean在完备度与可靠度方面均达到很好的测试效果, 并最接近于理想标准[9], 而根据Hopkins等[6]对天体自动搜索

的需求所进行的测试比较中, Aegean对于致密亮源与暗弱源的搜索拟合方面都保持较高的可靠度与完备度. 因此本文选取Aegean作为研究基础以对天体自动搜索算法、软件进行更进一步的研究、提升与开发.

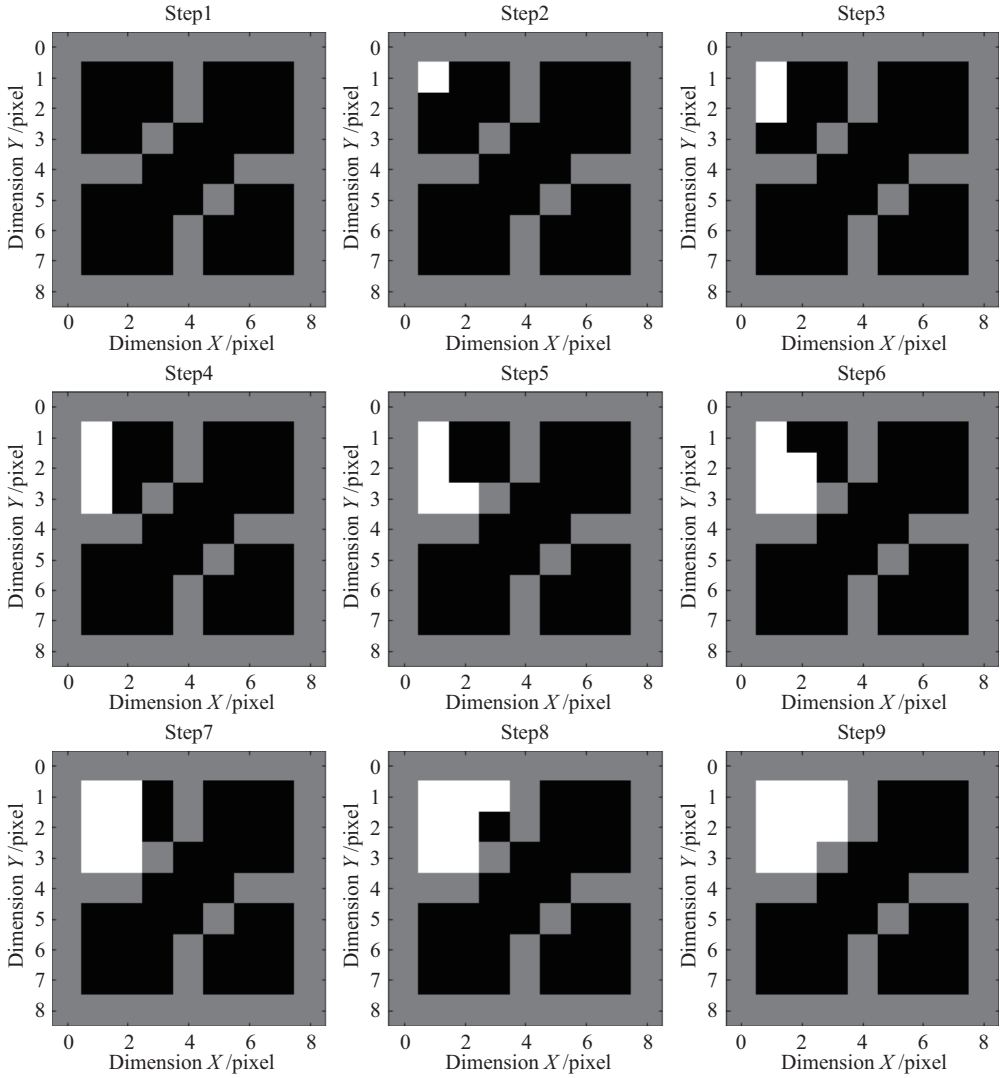

图 1　Flood-Fill算法模拟, 其中黑色区域的值被认为其亮度值大于灰色, 左上图为原始图像, 从坐标(1, 1)开始迭代处理, step$N$为迭代的每一步, 每一步可以识别出一个满足条件的像素点, 可以看到该模拟算法将从起始点开始的黑色连接块组合为一个像素岛(白色区域).

Fig. 1　Demonstration of the Flood-Fill Algorithm. The flux density values of black areas are considered to be greater than grey areas. The upper left image is the original image. The iterations start from $(1, 1)$ coordinates. Each step of iteration is step$N$. At each iteration, a pixel satisfying the requirements is identified. As shown in the graph, the Flood-Fill algorithm groups the connected block starting from $(1, 1)$ into a pixel island (white areas).

3 天体自动化搜索软件

3.1 设计理念

平方公里阵列的数据规模对天空自动搜寻技术提出了较高的要求, 所开发的算法、软件需要满足如下的条件: 适应于SKA科学数据处理管线系统并与各模块顺利衔接; 为不同的数据处理流程提供精确的天空模型; 能够提供多种类型的输出数据以助于进一步的科学研究; 具有高度的延展性以适应例如超级计算机等大规模数据处理环境; 满足上述性能的同时在运行与算法、软件方面需具备高效率以适应SKA的管线系统, 尤其针对快速成像、前期校准等重要工作的需求.

根据上述需求, 通过对SKA天体搜索的算法设计进行分析, 可知目前需要改进的为输入输出文件格式支持、算法完备度、自动化程度、执行效率、可延展性等. 根据现有算法、软件普遍具有的特点与优势, 在扩充相关功能的前提下, 着重考虑自动化搜索在海量数据自动处理和计算的应用. 在后续软件设计中, 我们着重对以下几方面进行了功能改进.

3.1.1 软件结构兼容性需求改进

在输入文件的读取支持方面, 首先改进了默认支持读取缺省波束信息的Flexible Image Transport System (FITS)图像文件的功能. 在天体搜索起始阶段, 波束信息为必要的输入参数, 经初始测试, 例如Very Large Array (VLA)数据, 该信息在部分FITS图像文件中无法自动读取, 手动设置则影响数据处理效率, 该功能的改进有助于拓展自动算法应用面, 使更多FITS图像文件可被直接读取, 从而达到流程一体化;

其次, 考虑到SKA及其先导望远镜和管线系统输出数据的多样性, 增加了天体搜索拟合结果图的输出功能, 支持Portable Network Graphic Format (PNG)等图片格式以及Comma Separated Value (CSV)等文件格式的输出. 天体自动搜索的结果既衔接于工程系统, 也具有科学用途, 对于巡天观测也尤为重要. 因此, 设计搜索拟合可视化效果帮助工程的判断与科学需求, 同时也增加了输出图像格式类型以便于后续的进一步检查处理.

3.1.2 软件自动化与普适性

为加强软件自动化性能及其对工程和科研的普适性, 改进了软件的功能, 形成天体自动搜索软件系统, 并设计了交互式的用户界面. 软件系统及界面的设计主要从可操作性、延展性与功能完备度等方面进行了考虑. 软件系统拥有友好的用户界面, 易于操作并便于功能的实现, 能够直观且快速展示所需的图像结果, 并可以多种格式保存输出数据. 基于SKA对天体搜索技术的要求, 软件系统及用户界面的设计也提升了处理大数据量的能力, 软件可一次性读取多个文件进行批量处理, 并按文件类型自动对输出文件进行归档. 基于SKA数据处理的复杂度, 也考虑使用户界面在功能扩展方面具有一定的灵活性, 软件系统设计具备数据库接口, 从而可与SKA成像管线系统与数据管理系统形成衔接.

3.2 软件介绍

根据天体搜索算法、软件的特性以及界面友好等用户界面指标, 开发了一款天体自动搜索集成化软件, 该用户界面原始界面如图2所示. 界面共包含4个区域: 输入区域、

输出区域、系统提示区域以及图形区域. 如图3所示.

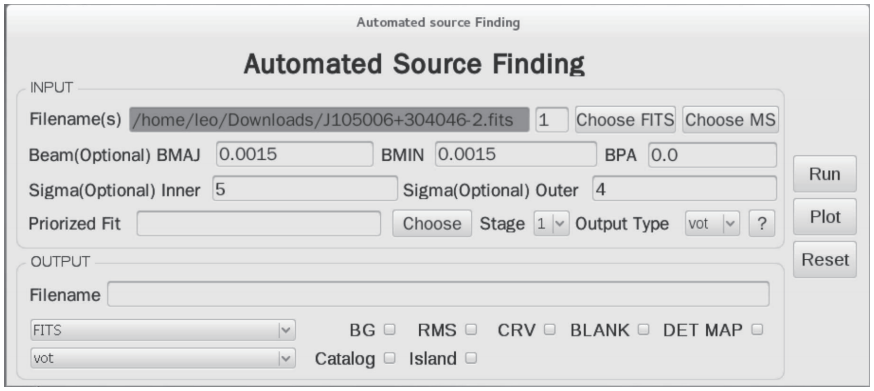

图 2　天体自动搜索用户原始界面

Fig. 2　The initial Automated Source Detection User Interface

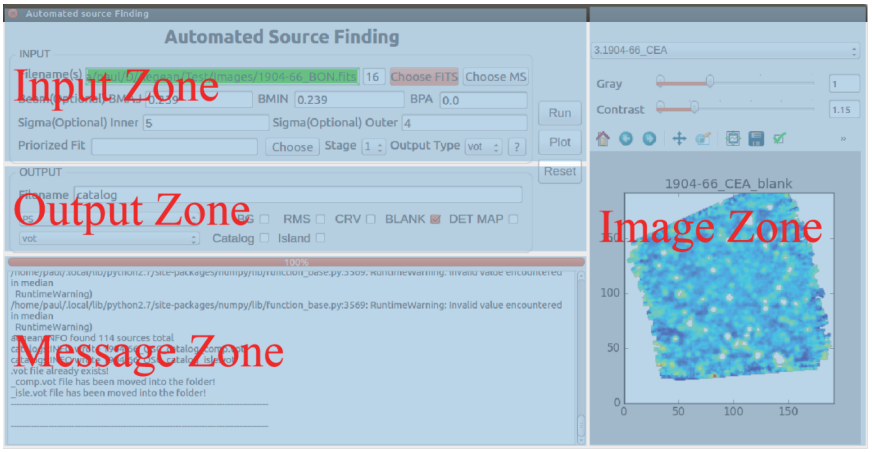

图 3　天体自动搜索用户界面区域

Fig. 3　Regions of the Automated Source Detection User Interface

输入区域为输入参数读取部分: 用户可选择所需读取拟合的输入图像数据, 可对波束参数进行调整, 默认值为FITS表头信息, 可对阈值σ参数进行调整, 并可在原图拟合的基础上对图像做特定区域的拟合;

输出区域为输出数据参数的设置部分: 用户可选择所需输出的文件及其格式, 可选择背景(BK)、噪声(RMS)、曲率(CRV)、残差(BLANK)及拟合(DET MAP)图像文件, 并以FITS、PNG、PS等格式输出归档, 并可通过多种文档格式输出保存星表(包含所有搜索天体的CATALOG星表和基于像素岛的ISLAND星表)文件;

系统提示区域显示软件运行过程中的所有系统信息, 并实时显示运行进度;

图形区域可显示指定文件对应的图像, 并可对图像进行灰度调节.

3.3　开发环境与架构

软件基于Linux操作系统研发, 兼容并支持目前大多数天文软件及集群的处理环境.

软件使用Python 2.7开发, 部分代码参考目前国际SKA先导阵列的成熟软件, Graphical User Interface (GUI)开发使用PyQt4.

Python语言是一个互动性及面向对象的脚本语言, 因其具有丰富的标准库, 而被称为胶水语言[4]. 该用户界面使用了如下模块:

(1)用户界面由PyQt模块编译;

(2)系统提示使用pprocess模块编译;

(3) MS格式使用CASA模块编译;

(4)图形显示使用PyFITS编译;

(5)数据库接口使用PySqlite编译.

运行环境: Linux. 需安装Python开发环境以及Qt4、Matplotlib、Pyfits、Numpy、Astropy等相关Python库. 自动化搜索的用户操作流程图如图4所示.

4 应用案例

为了验证天体自动搜索软件系统进行多项改进后的性能, 本节将展示该系统的测试结果. 所有测试均采用默认参数设置以保持测试的一致性, 并选择所有输出文件类型的文件输出以验证其完备度和时效性. 在数据的选取上, 考虑采用不同观测数据来源、望远镜天区、不同图像文件大小以及不同数量的两个批次数据进行测试. 测试基于2个E5-4610v4、12核CPU、256 GB内存的运行环境.

对于SKA量级望远镜, 自动处理庞大数据是一项重要需求. 因此, 选取第1组大天区图像对系统的运行效率及延展性能进行验证与展示. 数据来源于银河系与河外星系全天默奇森宽场阵列巡天项目(The Galactic and Extra-Galactic All-Sky MWA Survey, GLEAM)已公开数据[21]. 所测试及展示的数据是其中覆盖南天5 $\deg^2$的天区图像. 使用天体自动搜索系统, 可对该尺度天区FITS图像的所有图像(包括背景、噪声、曲率、残差及拟合图像文件)与星表进行自动连贯输出. 图5展示了输出的结果拟合图.

同样, 快速自动处理批量射电图像文件也是下一代望远镜对软件系统提出的需求. 第2组数据选取VLA Faint Images of the Radio Sky at Twenty-cm (FIRST)和APEX (Atacama Pathfinder EXperiment) Telescope Large Area Survey of the Galaxy (ATLASGAL)多张图像对该系统的自动化性能及时效性进行测试与演示. FIRST基于NRAO (National Radio Astronomy Observatory) VLA望远镜阵列而形成的南北1 × 10^4 $\deg^2$的射电巡天, 其角分辨率为$5''$[5], 从中随机选用了不同天区不同尺度60个FITS图像进行验证, 见表1. ATLASGAL是基于南天亚毫米波望远镜的巡天, 从中随机选取了不同大小的20个FITS图像进行验证[6].

天体自动搜索系统可一次性读取这80个FITS文件作为输入文件, 自动对这一组数据进行了天体识别与拟合, 并自动归档于相对应的文件目录, 每个FITS文件的拟合图像均可在系统界面上进行查看, 大大地方便了用户的操作. 图6展示了其中一个图像的拟合图. 图7为自动归档示意图. 系统在读取每一个FITS输入文件时将创建对应的文件目录, 归档输出图像文件, 并创建子目录用以归档所有的输出星表文件, 方便输出结果用作

[4]https://www.python.org/doc/essays/omg-darpa-mcc-position

[5]http://first.astro.columbia.edu

[6]https://atlasgal.mpifr-bonn.mpg.de

数据库管理.

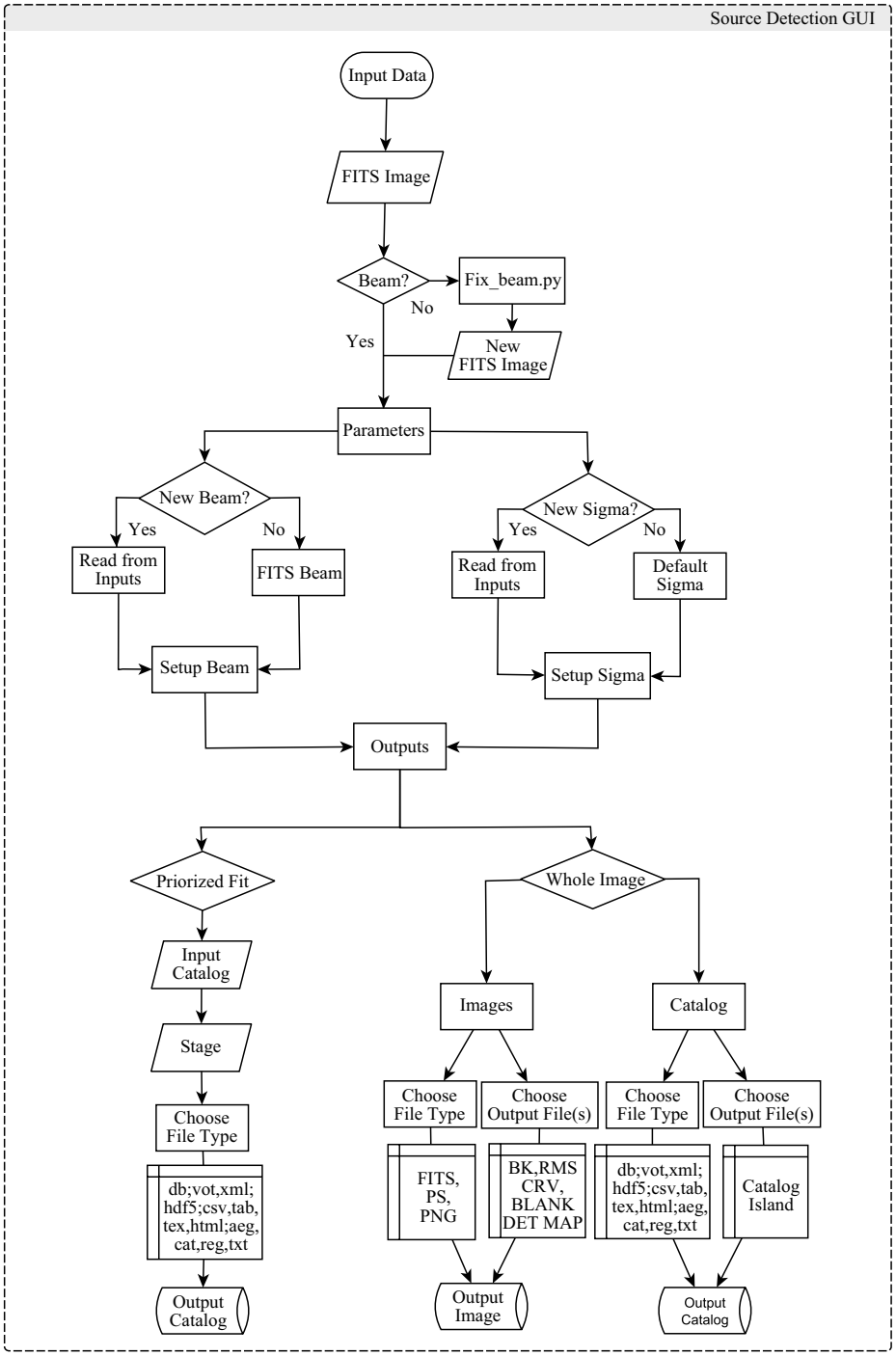

图 4　天体自动搜索用户界面操作流程图

Fig. 4　Operation flow of the Automated Source Detection User Interface

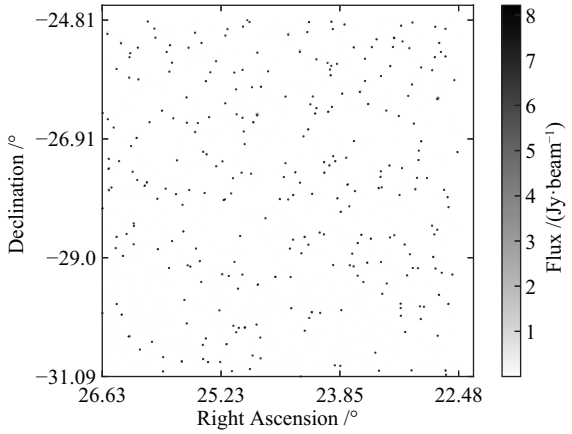

图 5 GLEAM图像的搜索拟合输出图

Fig. 5 The output image of Detection Map from the fitting of a GLEAM input image

表 1 VLA和ATLASGAL测试图像信息

Table 1 VLA and ATLASGAL images and Source Detection Results

Input Images	Number of Images	Image Size/MB	Coverage/ arcmin	Processing Time	Output files
VLA FIRST	10	0.10	4.5		
VLA FIRST	1	0.17	5		Output Images:
VLA FIRST	44	0.45	10	Approx.	BK, RMS, CRV,
VLA FIRST	4	1.75	20	5 min	BLANK, DET MAP
VLA FIRST	1	3.82	30	and	
ATLASGAL	10	0.96	5	58 seconds	Output Cataloge:
ATLASGAL	9	1.41	20	in total	CATALOG, ISLAND
ATLASGAL	1	3.17	30		
Total	80				

5 海量数据天体搜索应对方案探讨

天文学已逐渐进入一个大数据的时代, 伴随着更先进观测设备与技术的出现, 天文数据量已从PB量级向着EB量级跨越[3, 22]. 其数据量的庞大与多样性对于传统数据处理方法都是一个挑战, 天文界持续探寻着能够处理更大量级数据处理的技术方法, 并随着如今人工智能算法和计算力的演进, 进一步将新技术与天文数据处理方法进行融合, 使更多新的应用和算法软件得到探索和尝试.

SKA以其更高灵敏度、时间、频率与空间分辨率、巡天速度、大视场, 足以产生每秒TB量级数据[3, 22]. SKA所需要的是高效、准确、高度自动化的数据处理流程. 一方面, 天体搜索作为多个数据管线的开端, 关系后续多个环节的数据处理; 另一方面, 实时产生的庞大数据规模以及有选择性的数据存储, 让后续的错误数据修正以及遗漏数据的

重新添加变得不现实[6, 9], 因此, 少量甚至无需人工干预的自动化流程以及强大的计算能力, 是应对SKA量级数据的天体搜索的关键.

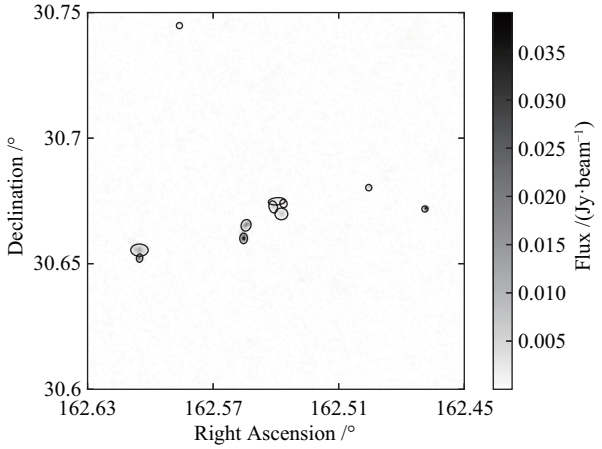

图 6　FITS图像的搜索拟合输出图示例

Fig. 6　An example of the output image of Detection Map from the fitting of a FITS input image

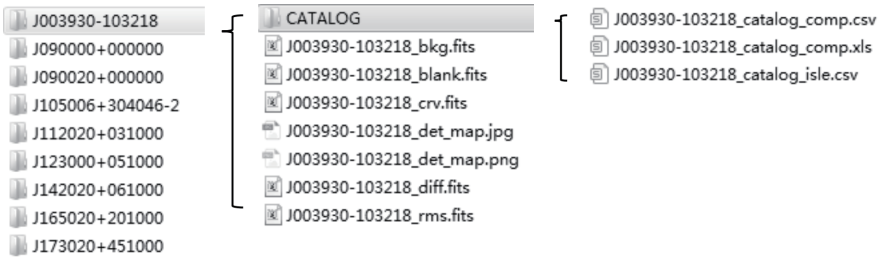

图 7　多图文件自动归档

Fig. 7　Automatic archiving of multiple output files

5.1　准确度

减少人工干预、修正, 让算法和软件实现最大可能的自动化对海量数据天体搜索工作十分重要. 因此, 天体搜索算法、软件的准确度尤为重要, 算法、软件需要具备高可靠度和高完备度, 大幅减少错勘、遗漏等. 为实现并不断提高天体搜索的准确度, 天文领域产生了一系列针对不同望远镜和科学目标的天体搜索算法, 这些算法也持续被探索、开发、归纳和改进, 如第2节所提到的基于光学开发的SExtractor, 基于射电天文、SKA先导项目MWA和ASKAP所开发的Aegean、Duchamp、Selavy等, 都在一代又一代的软件基础上进行有针对性的算法改进和迭代更新. 为了进一步帮助天体搜索算法更好地适应大型望远镜的需求, 天体搜索软件也越来越多地在大型望远镜巡天项目中得到应用、比较和性能评估. 例如Popping等[23]对可用于ASKAP HI巡天项目的Deep Investigation of Neutral Gas Origins (DINGO)和Widefield ASKAP L-band Legacy All-sky Blind surveY (WALLABY)的5种天体搜索算法、软件, 着重于点源和展源搜索的可靠度和完备度方面进行了测试比较. Hopkins等[6]也基于SKA探路者项目ASKAP的

宇宙演化巡天(The Evolutionary Map of the Universe, EMU)项目开展了数据竞赛, 选用了11个算法、软件对包含亮源、暗弱源、展源的3组类型天体的模拟图像进行了测试, 由此得到对算法准确度等性能更充分的了解, 并使算法、软件、ASKAP管线得到优化. 基于SKA射电谱线巡天, Norris等[4]对天体搜索的总体情况和技术挑战进行了探讨, 总结了致密源搜索算法的类型、优劣和改进需求, 也分析了展源和弥散源搜索面临的挑战, 为展源算法的进一步开发提供了意见. 而对于海量数据的计算性能方面, Dehghan等[18]则以Aegean和Duchamp作为研究对象, 尤其针对SKA管线和环境的适应性, 着重就运行环境部署, 包括计算成本、计算效率方面探讨了当前天体搜索算法的性能. SKA的数据竞赛也于2018年开始发布, 旨在帮助完善数据处理及科学数据分析, 其首个数据竞赛(Square Kilometre Array Science Data Challenge 1, SDC1)[24]即着重于考察数据处理算法、软件对于图像中的天体搜索与拟合的可靠度、完备度与准确度. 此外, 为了帮助天体搜索算法的可靠度和完备度指标更为精确化, 也产生了相应的评估算法和软件. ComEst[25]是为SExtractor所开发的主要用于光学和近红外图像完备度的评估软件, Serra等[26]提出了一种基于噪声的对称性, 使用负流量探测来判断天体搜索结果的可靠度的方法, Westerlund等[27]设计了基于谱线数据天体搜索准确度的评估软件Source Finder Accuracy Evaluator (SFAE).

这些评估和测试比较都为天体搜索算法、软件的升级和更先进算法、软件的开发提供了实验基础和改进方向. 总体而言, 对于算法、软件本身, 致密源天体搜索算法、软件的发展已使之具备较好的准确度(包括可信度、完备度、参数的准确度等)[27], 并得到不断的完善, 将在未来更多、更大数据量的测试评估中循序渐进加以优化; 对于弥散源、展源的搜索拟合, 无论是准确度方面还是技术的计算成本仍存在较大改进空间, 同时处理大型望远镜产生的海量数据也是比较大的挑战, 因此仍处在算法、软件发散式探索和基于现有算法、软件的改进中. 未来将需要更多场景的测试、比较, 并逐渐形成比较完善且系统的兼顾点源和展源等多种天体类型的算法、软件; 对于算法、软件在运行环境上的部署, 面对数据规模如SKA的需求, 已有算法、软件在延展性能方面仍有差距, 需要天文与软件计算领域更多的协作[18].

5.2 弥散源

在天体搜索中, 复杂结构天体(展源、弥散源、暗弱源等)的搜索拟合是现阶段海量数据背景下天体搜索算法的挑战之一.

下一代巡天望远镜, 包括SKA先导项目The Westerbork Synthesis Radio Telescope (WSRT)望远镜的The Westerbork Observations of the Deep APERTIF (The new Phased Array Feed receiver system for the Westerbork Synthesis Radio Telescope) Northen-Sky (WODAN)[28], ASKAP阵列的EMU巡天[29]和MeerKAT的The MeerKAT International GHz Tiered Extragalactic Exploration (MIGHTEE)巡天[30]等, 将在灵敏度、分辨率和视场等多个方面较先前的设备有重大突破[31]. 例如, EMU将实现对整个南天的深度巡天, 其灵敏度与角分辨率将分别高于The NRAO VLA Sky Survey (NVSS)巡天45倍和4.5倍. 而EMU与WODAN将共同提供频率在1.3 GHz、角分辨率达10″–15″、灵敏度达10 μJy$\cdot$beam^{-1}的全天空成像[29].

这些巡天望远镜的使用, 将有助于天文学家们对暗弱源及延展结构的天体进行观

测, 获取更精确的图像, 并展开进一步的研究和分析. 而这需要天体搜索算法和软件能够对更大数据量的天文图像中的弱源、展源天体进行快速自动识别并拟合.

即使是目前点源搜索相对表现较好, 可靠度和完备度较高的天体搜索算法和软件, 在对大型望远镜巡天的数据进行天体搜索时仍存在较大的不足[6]. 针对最考验天体搜索软件和算法的特殊天体—展源、弥散源、暗弱源、暂现源等, 天文界从不同角度, 甚至借助了机器学习等人工智能算法进行了探索, 针对展源搜索的算法和软件也相继提出. 多层次贝叶斯算法[32]、(小波分解的)传统阈值方法、压缩感知算法、基于霍夫变换的圆检测算法等[33–35]是针对展源、弥散源所开发的软件和算法, 但现有的技术也可能面临较高的计算成本, 需要算法与技术的融合加以实现[4]; 针对暗弱源的一些算法、软件正在进一步研究、开发中, 基于Giant Metrewave Radio Telescope (GMRT)和The Australia Telescope Compact Array (ATCA)数据的测试分析也在进行中[36]; 针对SKA及下一代大型望远镜科学数据处理这一重要领域的暂现源成像系统, 所需的天体搜索算法、软件也进行了探讨和分析[37–38]. 而近年来, 随着机器学习等人工智能算法得到更多的应用, 天文界也逐步尝试了这些算法, Gheller等[39]将深度学习神经网络应用于弥散射电源的搜索并开发了名为Cosmodeep的软件, 同样结合深度学习, Sadr等[40]开发了针对低信噪比天体搜索的DeepSource算法. 这些方法也需要结合现有的算法、软件加以融合形成更完备的软件系统, 以完善整体性能, 一方面应用于巡天项目、数据处理管线, 另一方面, 通过测试评估, 分析各算法、软件的优劣, 从而能对算法、软件进行更好的改进.

5.3 计算力和效率

面对海量数据, 天体搜索算法的性能表现也高度依赖于科学数据处理管线的兼容、衔接以及运行环境的部署.

以SKA为代表的数据密集型计算需实现实时数据流的处理, 对于每一个环节数据处理的时效性有很高的要求, 在算法、软件的设计和运行环境的选择上, 需考虑追求准确度所带来的软件复杂度与运行效率之间的权衡, 对于数据的读写速率(I/O吞吐量)、内存的消耗都需要很好的考量. 未来SKA量级的数据处理均需要部署于高性能计算机, 当前科学数据处理(Science Data Processor, SDP)研发任务正在开展中, 欧洲和站址国正在为SKA及先导项目建设数据中心, 中国科学院上海天文台也正展开对SKA区域中心原理样机的研发[41], 因此, 这也需要算法、软件的设计适用于不同高性能运算环境, 可进行并行化等加速处理. 本文所述天体搜索软件系统也将根据此需求进行改进优化, 并基于原理样机进行测试. 此外, 与其他软件和管线系统的无缝衔接能力, 拥有高效的文件管理系统也是处理海量数据的重要需求, 这也是本文所述软件系统设计的重要考量.

SKA建成后将产生海量数据, 预估SKA第1阶段(建成10%)的数据规模所需的运算能力已远超于当前世界最快的超级计算机[22]. 面对如此庞大的数据量, 顺利进行科学数据处理对于目前高性能计算机的计算能力都是极大的挑战, 尤其是对系统I/O、传输带宽、共享资源调度分配、数据管理等紧密关系到计算效率的环节提出很高的要求.

例如中国科学院上海天文台的科研团队针对SKA海量数据处理要求, 正在研发SKA数据中心原理样机[3, 22], 在计算架构方面, 不同于传统高性能计算架构将数据移至缓存进行并行处理, 而采用适应于SKA数据密集型计算的数据岛计算架构, 提高时效并保障数据处理的流畅和稳定. 应对海量数据的数据管理, 采用了由西澳大学牵头合作

研发的数据流管理系统[42](Data Activated Flow (Liu) Graph Engine, DALiuGE), 实现数据密集型计算的高效性、实时性、连续性和低能耗.

6 总结与展望

本文探讨了天体自动搜索算法、软件的发展现状, 并基于现有的算法, 改进、开发了一套更具适用性和自动化程度更高的集成软件, 为SKA科学数据处理提供了软件验证支持. 测试表明, 该软件对于不同类型图像具有良好的自动化处理效果, 能够实现自动批量处理不同大小图像并处理大天区图像, 具有更好的交互能力, 为将来软件的进一步开发与发展提供了参考.

SKA由于具有极高的灵敏度、时间、频率与空间分辨率、巡天速度、大视场, 而产生了海量的数据, 高效、准确、高度自动化的数据处理流程是目前应对大数据处理所必须的, 本文讨论的解决方案已经可以部分应对自动化的处理流程, 但是在海量数据的高速处理上还有改进空间.

另外, 相对于点源而言, 复杂结构的天体流量密度更低, 因此复杂结构搜索的工作更具有挑战性, 并且由于具有其自身的发射特性, 后续将进一步考虑优化算法, 完善软件对延展源的支持.

致谢 感谢中国SKA区域数据中心原型样机给本项目提供的硬件平台支持. 感谢Paul Hancock等提供的相关软件. 感谢Paul Hancock对本项目的建议.

参 考 文 献

[1] Dewdney P E, Hall P J, Schilizzi R T, et al. Proceedings of the IEEE, 2009, 97: 1482
[2] 郭绍光, 郑小盈, 毛羽丰, 等. 科研信息化技术与应用, 2018, 9: 3
[3] An T. SCPMA, 2019, 62: 989531
[4] Norris R P, Afonso J, Bacon D, et al. PASA, 2013, 30: e020
[5] Hollitt C, Johnston-Hollitt M, Dehghan S, et al. An Overview of the SKA Science Analysis Pipeline//Lorente N P F, Shortridge K, Wayth R. Astronomical Data Analysis Software and Systems XXV. San Francisco: Astronomical Society of the Pacific, 2017, 512: 367-370
[6] Hopkins A M, Whiting M T, Seymour N, et al. PASA, 2015, 32: e037
[7] Masias M, Freixenet J, Lladó X, et al. MNRAS, 2012, 422: 1674
[8] Bertin E, Arnouts S. A&AS, 1996, 117: 393
[9] Hancock P J, Murphy T, Gaensler B M, et al. MNRAS, 2012, 422: 1812
[10] Lutz R K. CompJ, 1979, 23: 262
[11] Hales C A, Murphy T, Curran J R, et al. MNRAS, 2012, 425: 979
[12] Whiting M, Humphreys B. PASA, 2012, 29: 371
[13] Whiting M T. MNRAS, 2012, 421: 3242
[14] Hopkins A M, Miller C J, Connolly A J, et al. AJ, 2002, 123: 1086
[15] Hancock P J, Trott C M, Hurley-Walker N. PASA, 2018, 35: e011
[16] Roerdink J B T M, Meijster A. Fundamenta Informaticae, 2000, 41: 187
[17] Huynh M T, Hopkins A, Norris R, et al. PASA, 2012, 29: 229
[18] Dehghan S, Johnston-Hollitt M, Hollitt C. Point Source Detection Software in the SKA Era//Lorente N P F, Shortridge K, Wayth R. Astronomical Data Analysis Software and Systems XXV. San Francisco: Astronomical Society of the Pacific, 2017, 512: 233-236

[19] Benjamini Y, Hochberg Y. Journal of the Royal Statistical Society: Series B (Methodological), 1995, 57: 289
[20] Shaffer J P. Annual Review of Psychology, 1995, 46: 561
[21] Wayth R B, Lenc E, Bell M E, et al. PASA, 2015, 32: e025
[22] 安涛. 科技纵览, 2017: 52
[23] Popping A, Jurek R, Westmeier T, et al. PASA, 2012, 29: 318
[24] Bonaldi A, Braun R. arXiv:1811.10454
[25] Chiu I, Desai S, Liu J. A&C, 2016, 16: 79
[26] Serra P, Jurek R, Flöer L. PASA, 2012, 29: 296
[27] Westerlund S, Harris C, Westmeier T. PASA, 2012, 29: 301
[28] Röttgering H, Afonso J, Barthel P, et al. JApA, 2011, 32: 557
[29] Norris R P. Evolutionary Map of the Universe//Proceedings of the International Astronomical Union, IAU Symposium. International Astronomical Union, Cambridge University Press, 2012, 284: 489
[30] Van der Heyden K, Jarvis M J. MIGHTEE proposal to Meerkat, 2010
[31] Riggi S, Ingallinera A, Leto P, et al. MNRAS, 2016, 460: 1486
[32] Butler-Yeoman T, Frean M, Hollitt C P, et al. Detecting Diffuse Sources in Astronomical Images//Lorente N P F, Shortridge K, Wayth R. Astronomical Data Analysis Software and Systems XXV. San Francisco: Astronomical Society of the Pacific, 2017, 512: 217-220
[33] Peracaula M, Lladó X, Freixenet J, et al. Segmentation and Detection of Extended Structures in Low Frequency Astronomical Surveys using Hybrid Wavelet Decomposition//Evans I N, Accomazzi A, Mink D J, et al. Astronomical Data Analysis Software and Systems XX. San Francisco: Astronomical Society of the Pacific, 2011, 442: 151-154
[34] Dabbech A, Ferrari C, Mary D, et al. A&A, 2015, 576: A7
[35] Hollitt C, Johnston-Hollitt M. PASA, 2012, 29: 309
[36] Peracaula M, Torrent A, Masias M, et al. NewA, 2015, 36: 86
[37] Lucas L, Staley T, Scaife A. A&C, 2019, 27: 96
[38] Mooley K P, Frail D A, Ofek E O, et al. AJ, 2013, 768: 165
[39] Gheller C, Vazza F, Bonafede A. MNRAS, 2018, 480: 3749
[40] Sadr A V, Vos E E, Bassett B A, et al. MNRAS, 2019, 484: 2793
[41] 洪晓瑜, 武向平, 安涛, 等. 天文学进展, 2018, 36: 348
[42] Wu C, Tobar R, Vinsen K, et al. A&C, 2017, 20: 1

Research on Source Detection Algorithm Based on Astronomical Images and the Implementation of an Automated Software System

LU Yang[1] AN Tao[1,2] GUO Shao-guang[1] LAO Bao-qiang[1]
(1 Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030)
(2 Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Nanjing 210033)

Abstract Source detection is an important part of data processing pipeline, and it is also one of the challenges that Square Kilometre Array (SKA) and other next generation telescopes are facing with when dealing with massive data. To date, source detection algorithms have become quite mature and been applied in various data processing. Meanwhile, there are still areas such as automation that can be further improved on, and more tests are necessary for the fulfillment of the requirement of SKA data processing. The purpose of this paper is to conduct the research on source detection algorithms that are more automated and adaptive to massive data processing. Based on it, the research team has made improvements of source detection algorithm, and designed and developed a set of automated source detection software system, which is highlighted with a user-friendly interactive interface, output display function, more compatible data input and output, and improved data management. Integrating multiple functions together enables it to have good performance for automatic processing of large sky image and image sets, and the test results show that the improvements are effective. The research team will make further improvements and develop functions to meet the needs of SKA.

Key words techniques: image processing, methods: analytical, catalogs