第60卷 第5期 天 文 学 报 Vol.60 No.5
2019年9月 ACTA ASTRONOMICA SINICA Sept., 2019

doi: 10.15940/j.cnki.0001-5245.2019.05.009

基于图像相减和随机森林的AST3巡天暂现源及变源搜寻方法*

黄天君[1,2] 孙天瑞[1,2] 胡 镭[1,2] 宁宗军[1] 吴雪峰[1†]
王力帆[3,1,4] 王晓峰[5,1] 朱镇熹[1] UDDIN Ashraf Syed[6]
ASHLEY Charles Brewster Michael[7]

(1 中国科学院紫金山天文台 南京 210033)
(2 中国科学技术大学天文与空间科学学院 合肥 230026)
(3 George P. and Cynthia Woods Mitchell Institute for Fundamental Physics & Astronomy, Texas A. & M. University, Department of Physics and Astronomy, Texas TX 77843)
(4 中国南极天文中心 南京 210033)
(5 清华大学物理系/清华天体物理中心 北京 100084)
(6 The Observatories of the Carnegie Institution for Science, California CA 91101)
(7 Department of Astrophysics, University of New South Wales, New South Wales NSW 2052)

摘要 AST3-2 (Antarctic Survey Telescopes)光学巡天望远镜位于南极大陆最高点冰穹A, 其产生的大量观测数据对数据处理的效率提出了较高要求. 同时南极通信不便, 数据回传有诸多困难, 有必要在南极本地实现自动处理AST3-2观测数据, 进行变源和暂现源观测的数据处理, 但是受到低功耗计算机的限制, 数据的快速自动处理的实现存在诸多困难. 将已有的图像相减方案同机器学习算法相结合, 并利用AST3-2 2016年观测数据作为测试样本, 发展一套的暂现源及变源的筛选方法成为可行的选择. 该筛选方法使用图像相减法初步筛选出可能的变源, 再用主成分分析法抽取候选源的特征, 并选择随机森林作为机器学习分类器, 在测试中对正样本的召回率达到了97%, 验证了这种方法的可行性, 并最终在2016年观测数据中探测出一批变星候选体.

关键词 恒星: 变星: 普通, 方法: 数据分析, 技术: 图像处理
中图分类号: P141; **文献标识码**: A

2019-04-14收到原稿, 2019-04-24收到修改稿

*国家自然科学基金项目(11725314、11673068、11325313、11633002、11761141001), 中国科学院前沿科学重点研究项目(QYZDB-SSW-SYS005), 中国科学院战略性先导科技专项(XDB2300000), 江苏省第5期“333工程”培养资金项目资助

†xfwu@pmo.ac.cn

1 引言

时域天文已经成为目前天体物理研究的关键领域, 它的研究对象包括暂现源和变源, 如超新星、伽马射线暴、变星、活动星系核、系外行星等, 是我们了解宇宙中极端物理现象的主要途径[1]. 时域天文巡天是时域天文的主要研究手段, 对预定天区进行重复观测, 以期获得天体的时变信息, 现在已经有多个采用上述手段进行的时域天文巡天项目, 如PTF (Palomar Transient Factory)巡天[2], TNTS[3] (Tsinghua University NAOC Transient Survey)以及本文涉及的AST3 (Antarctic Survey Telescope, AST3)巡天. AST3巡天的选址位于南极大陆冰穹A. 南极大陆的大气环境极为寒冷、干燥且稳定, 因此是非常理想的天文选址. 2008年, 中国首套南极巡天设备CSTAR (Chinese Small Telescope Array)开始在南极大陆冰穹A投入运行, 其继任者AST3系列是中国的第2代南极光学望远镜, 计划部署3台[4], 分别于2012年和2015年部署了系列中的第1台AST3-1和第2台AST3-2, AST3-3仍在测试中. AST3-2是南极大陆目前最大的光学望远镜, 入瞳直径0.5 m, 像面对应视场4.25 deg^2, 主要用于包括新星[5–6]、变星[7]、系外行星[8–9]、活动星系核[10]等在内的变源和暂现源[11]的研究, 冰穹A的气候条件和地理位置[12]使得此处非常适合开展长时间不间断的时域天文观测活动[13–15]. 本文工作使用AST3-2在2016年度观测得到的数据作为测试样本.

时域天文巡天广泛使用现代化的大视场望远镜, 望远镜的长期运行积累了大量的观测数据, 如何高效处理这些数据成为一个新的问题. Alard等[16]提出的图像相减法是目前时域天文领域应用比较广泛的一个方法, 通过比较两个时刻的星像流量差异来获得天体的时变信息, 可以准确找出流量发生变化的源. 理想情况下相减后得到的残差图像上应该只有流量发生变化的暂现源和变源, 但是在实际操作中不可避免地受到各种因素的干扰, 如仪器影响、对齐偏离、卷积异常等, 这使得我们需要对相减后的残差进行区分, 而仅靠人力去处理观测产生的海量数据是不现实的. 在数据科学领域常使用机器学习代替人类执行一些分类和预测的工作, 天文数据处理方面也已有此尝试, 其中先驱当属Bailey等人在2007年处理超新星工厂(Super Novae Factory, SNFactory)中提出的分类方法[17]. 机器学习算法有许多种, 其中Breiman[18]提出的随机森林是目前性能较好的机器学习算法, 优点是极为准确、能够处理高维数据等. 现有的变星探测方法是从候选源的光变信息出发, 通过光变曲线的特定参数判断候选源是否为变星以及对应的变星分类, 我们希望能够结合前述图像相减法和随机森林这两个成熟算法的优点, 发展一套基于这两者的变星搜寻流程, 提高数据处理的效率.

2 数据处理流程

AST3-2属于施密特式望远镜[19], 入瞳直径0.5 m. 为了适应低温环境和无人值守的情况, AST3系列通过卫星通讯进行远程操控[20–22], 在软硬件方面均有一些特殊的设计[20–21, 23]. AST3-2相对于AST3-1的改进主要集中于伺服系统. AST3-2的视场大小为4.25 deg^2, CCD物理分辨率为10560 × 10560像素, 折合每像素对应1″. 为了避免可能的机械故障, AST3系列均没有安装机械快门, CCD以帧转移模式工作, 这使得实际可用视场和分辨率减半, 为10560 × 5280像素. AST3-2采用SLOAN数字化巡天标准i波段滤光片[24]. 在2016年度的巡天任务中, AST3-2覆盖了571个天区, 合2352.52 deg^2, 通常曝光

时间为60 s, 读出时间20 s. 每个天区的观测次数从15次到125次不等, 多数天区重复观测次数在35次左右, 共计得到25000张以上图像, 总数据量达到5.8 TB. 我们统计了全年观测数据的大气质量, 分布情况如图1, 从中得知我们的巡天观测计划安排得当, 主要利用大气质量较小的区域来观测.

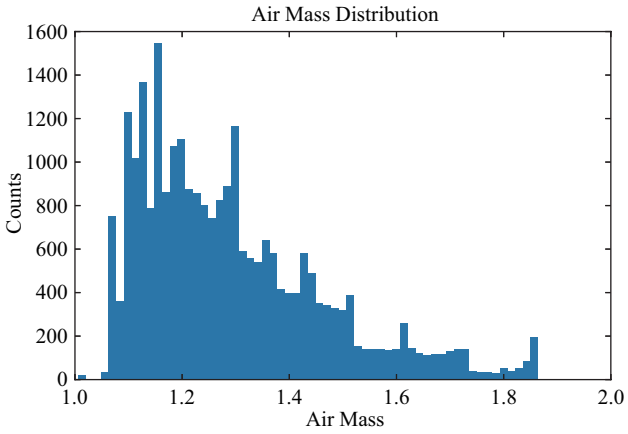

图 1　AST3-2 2016年度观测数据大气质量分布图

Fig. 1　Air mass distribution of AST3-2 2016 dataset

以中心指向赤经RA = 15:52:00, 赤纬DEC = −44:32:23的2152–4454天区为例, 2016年度测光误差分布和星像半高全宽(Full Width at Half Maximum, FWHM)状况如图2和图3所示, 可见在一幅典型的图像中, 测得星等亮于17 mag的测光误差小于0.17 mag, 星像的FWHM多在5像素左右.

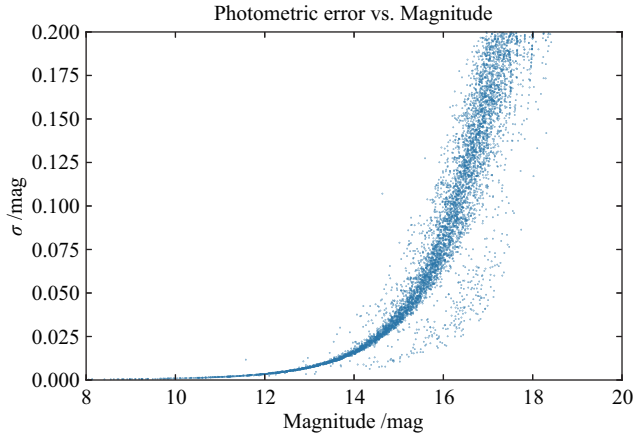

图 2　AST3-2 2016年度观测数据典型图像的测光误差分布图

Fig. 2　Photometric error distribution of a typical image of AST3-2 2016 dataset

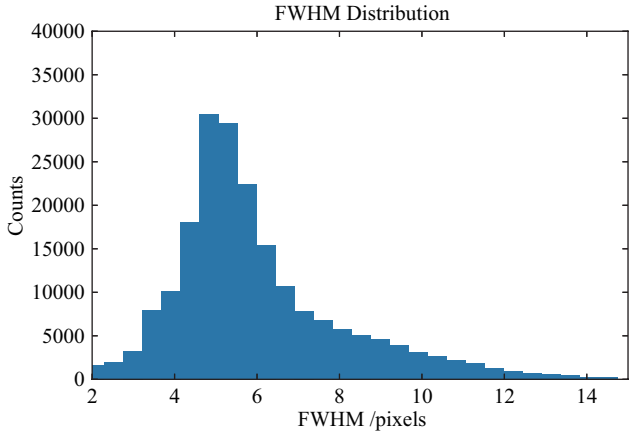

图 3 2016年数据2152-4454天区星像FWHM分布情况

Fig. 3 FWHM distribution of the images in the sky area 2152-4454 of AST3-2 2016 dataset

天体在经过一段时间后其亮度发生变化, 那么在此之后拍摄的图像上将会反映出这一变化. 在理想情况下, 相减之后背景会被全部扣除, 只剩下星像之间的残差, 因此可以根据两张图片之间的差别来判断是否存在变源, 图像相减法[16, 25]即是根据这种差异来检测变源的, 通过将已知模板图像同待检测图像进行相减, 通过判断图像间的差异来寻找可能存在的变源. 用$\boldsymbol{D}$代表差异, $\boldsymbol{T}$代表模板图像, $\boldsymbol{I}$代表待检测图像, $\boldsymbol{K}$代表卷积核, 其基本思想如下式:

$$\boldsymbol{D} = \boldsymbol{T} - \boldsymbol{I} \otimes \boldsymbol{K} \,. \tag{1}$$

点扩散函数(Point Spread Function, PSF)是光学系统的输入为点光源时其输出像的光场分布, 常用于描述光学系统对点源的响应. 由于相减的两张图片并不是拍摄于同一时刻, PSF因外部条件的不同有所差异, 不能直接相减, 需要经过卷积使得两者的PSF尽可能相似, 然后再进行相减. 在执行图像相减之前需要做一些预先的准备工作, 包括图像对齐和模板制作. 本文采用图像相减法初步获得观测范围内天体的星等变化信息.

2.1 图像对齐

在图像的预处理方面, 我们利用CCD上的过扫描(overscan)区域进行图片本底(bias)修正, 然后进行平场处理[26]. 我们采用SExtractor[27]软件对所有图像进行自动孔径测光, 测光星表作为下一步的基础. 望远镜在不同时段对同一天区进行拍摄时, 由于望远镜的指向存在误差, 观测所产生的图像可能会产生偏差, 在对观测数据进行处理的过程中需要做图像对齐, 将同一天区的所有图像统一到一个坐标系中去. 我们借助FITSH[28]软件包中的grmatch指令对图像进行对齐. FITSH程序包提供了一整套用于天文图像数据分析的工具, 可以完成包括图像校准、源的认证、测光、图像组合、空间变换等一系列图像处理中的步骤. 在这一步我们输入需要进行校准的星表和图像, FITSH会自动完成星表对齐和图像的变换. FITSH读取需要校准的星表之后, 使用三角对齐算法[28–29]找出两张星表在空间上的变换关系, 随后对其中每一个点进行交叉认证, 不断重复上述步骤, 直到达到令人满意的匹配率. 这套算法能够很好地适应大视场高分辨率图像. 两张星表之

间的转换信息单独写入一个文件中, 程序会根据这个文件生成一个转换后的图片, 新生成的图像已经对齐到参考图像的参考系中. 我们选择像质最佳的图像作为参考图像, 将同一天区的其他所有图像对齐到参考图像的坐标系中去. 图像对齐的效果视图片质量而定, 偶尔也会出现同一个源在不同图片上位置偏离的情况, 我们最后整理数据的时候设置0.5倍FWHM的匹配半径. 图4展示了图像对齐的效果.

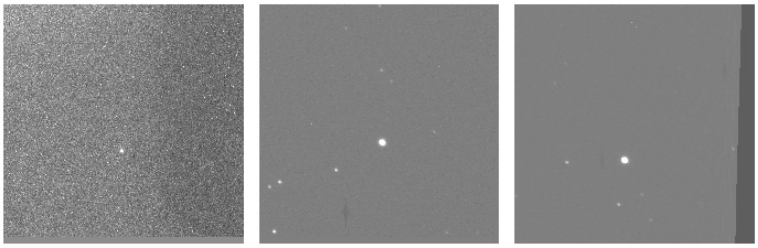

图 4　图像对齐事例, 图上列出了参考图像(a), 待检测图像(b)和对齐后的图像(c).

Fig. 4　Example of image registration, reference image (a), input image (b), and remapped image (c) are shown on the panels.

2.2　模板制作

图像相减前需要制作相减的模板图像. 每张图片的拍摄时间和拍摄条件均有不同的差异, 这些差异反映到成像结果上体现为PSF的区别, 成像质量最佳的图像具有最小的视宁度和最为锐利的PSF. 在图像相减的过程中, 一般是通过卷积来使得待检测图像的PSF轮廓尽可能匹配模板图像的PSF[27], 也就是说模板图像的PSF决定了图像相减的执行质量. 因此我们希望模板图像的质量尽可能好, PSF尽可能锐利. 我们通过叠加生成一张高信噪比的模板图像[30–31], 把叠加生成的图片称作生成模板.

在模板制作之前首先确定每张待叠加图像的权重, 我们从信噪比出发得到图像的权重. 图像的信噪比S/N可以用如下式子表示:

$$\mathrm{S/N} = \frac{N_{\mathrm{photon}}}{\sqrt{A_{\mathrm{psf}}}\sigma_{\mathrm{sky}}}\,, \tag{2}$$

N_{photon}代表光子数目, A_{psf}表示源所占的像素数目, σ_{sky}表示每个像素的天光背景. 显然N_{photon}应该和相对的透明度T成正比, A_{psf}应当与FWHM^2成正比. 我们引入这样定义的每张图的权重w_i, 并用T_i、FWHM_i^2、σ_i分别表示每张图的透明度、FWHM和天光背景:

$$w_i = \frac{T_i}{\mathrm{FWHM}_i^2\sigma_i^2}\,, \tag{3}$$

其中相对透明度T可以通过如下过程比较两幅图像的零点来得到. 用E_1、E_2分别表示进行零点比较的两张图像的流量, 对应的流量关系为$E_2 = E_1T$, 对应到用零点ZP_1、ZP_2表示的星等之间的关系$\mathrm{ZP}_2 - \mathrm{ZP}_1 = -2.5\lg\frac{E_2}{E_1}$, 进而得到透明度的表达式:

$$T = 10^{-0.4(\mathrm{ZP}_2-\mathrm{ZP}_1)}\,. \tag{4}$$

根据FWHM选取8张成像质量最佳的图像, 将其中最佳的一张作为参考图像, 利用FITSH将其他7张统一到该参考图像的坐标系中去, 并按归一化后的权重进行叠加, 最后产生的图像有着比原始图像更小的FWHM, 效果如表1和图5.

表 1 原始图像同模板图像的FWHM比较. I_1至I_7表示连同参考图像进行叠加的7张图片, Ref表示参考图像

Table 1 FHWM comparison between the original images and coadded template, I_1 to I_7 are 7 original images and ref represents the reference image

Sky Area	Coadd	Ref	I_1	I_2	I_3	I_4	I_5	I_6	I_7
0453-4750	4.6	4.91	5.09	7.15	5.27	4.82	5.37	5.70	5.64
0534-5046	4.76	5.47	4.94	4.92	5.67	5.59	5.67	6.27	6.11
1925-4750	5.02	4.47	5.37	5.52	5.09	5.64	5.08	5.01	4.93
0523-4158	4.77	5.34	5.45	6.82	5.52	5.05	5.49	4.98	4.97

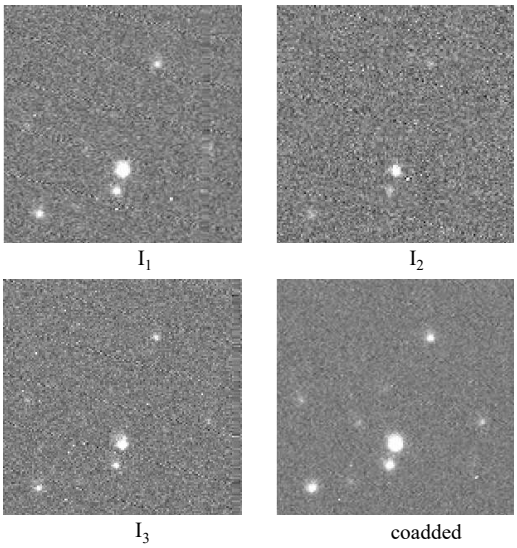

图 5 叠加生成模板同原始图像的比较, I_1、I_2、I_3是原始叠加图片其中的3张, coadded为叠加生成的模板图像.

Fig. 5 Comparison of the coadded template image and original images. I_1, I_2, I_3 show three original images, and coadded shows the coadded template image.

2.3 图像相减

我们借助HOTPANTS[32]这一软件来进行图像相减. HOTPANTS是高阶PSF变换与模板相减(High Order Transform of Psf ANd Template Subtraction)的缩写, 专门用于执行图像相减, 其核心算法和代码来自Alard等[16, 25]的工作, 可以自动完成求解卷积核、卷积、图像相减等步骤, 对于同一张图各处的PSF有所不同的情况也有考虑.

减图完成后我们对残差图像进行测光, 测光阈值设为2σ, 将流量高于背景流量标准差两倍以上的区域视作星像, 测光结果除用来标记源的位置以外不做其他用途. 增亮的源在残差图上留下的是正值, 用通常的测光方法就可以检测出来, 变暗的源留下的则是

负值, 为此我们对整张图的数值做一次翻转, 对翻转后的图像进行测光, 这样就可以将变暗的源检测出来.

本文描述的工作所处理的是2016年度档案数据, 所有待检测图像的质量都是已知的, 而在实时数据处理中, 一般是事先生成一张模板, 再将观测得到的图像同模板进行相减, 这就涉及到模板和待检测图像的视宁度关系. 假如出现待检测图像视宁度优于模板图像的情况, 需要把待检测图像和模板图像的关系反转, 对模板图像做卷积后同待检测图像相减, 这是处理档案数据和实时数据的不同之处.

不用待处理图像减模板图像的原因是, 虽然可以通过反减的方式将残差为负的源检测出来, 但相减的过程中需要将相减项的PSF卷积到被减项的PSF, 所以残差图像的PSF同被减项的PSF一致, 这样残差图像的质量比正减所得到的要差.

为了验证图像处理流程的可靠性, 我们将减图和测光的结果同VSX (Varibale Star Index)星表做对比. VSX星表是已知最完整的变星星表, 如果我们的方法能将VSX星表里对应天区的全部变星都检测出来, 那么就可以认为这一套流程是有效的. 同时, 如果有一部分已知变星没有被检测到, 就意味着我们设置的参数需要调整. 我们根据比较的结果反复调整了测光参数的设定. 为了保证所有的变源都被包括进来, 我们选取了较为宽松的限制, 比如将阈值设得比较低. 表2是选取较低阈值时的比较结果, 我们挑选了0007-4158天区作为展示, 匹配半径设为5 arcsec, 约合5像素, 已知变星可以多次检出, 说明测光阈值设置较为合适.

表 2 残差图像测光结果中已知变星的赤经(RA)、赤纬(DEC)、星等、星等变化(MAG VAR)及其出现次数

Table 2 Right ascension (RA), declination (DEC), magnitude, magnitude variation (MAG VAR), and occurrence number of known variable stars in the result of residual image photometry

RA	DEC	magnitude	MAG VAR	occurrence
00:04:00	−42:43:56	14.68	0.62	15
00:04:09	−41:08:10	13.68	0.96	15
00:04:51	−42:30:40	15.17	1.07	15
00:05:28	−41:53:14	7.31	7.63	15
00:05:36	−42:11:55	17.29	1.02	13
00:05:53	−42:55:19	17.18	0.99	13
00:06:42	−42:27:03	15.81	0.86	14
00:07:24	−42:33:36	9.48	0.06	15
00:10:26	−42:33:13	15.68	0.96	16

2.4 残差图像处理

在确保所有变源都被包括进来之后需要对残差图像的测光结果做初步的筛选, 我们对候选源做如下方面的限制: 首先把靠近图像边缘的源舍去. 受光学系统成像原理的限制, 远离光轴、靠近图像边缘的星像质量通常是比较差的. 同时图像处理的过程中, 如果模板图像中靠近边缘的某些星像在同它进行对齐的图像中没有对应的源, FITSH会在创建的新图像中将对应位置的数值设为0, 无论该源的亮度是否变化, 相减之后都会在残差图像上留下痕迹. 图像相减的方法本身不能区分这些残差的来源, 位于模板图像上的源同待检测图像中的异常值相减后得到的残差同正常情况下的残差并没有显著的差别, 所以需要根据实际星像进行判断. 其次是形态方面的限制, 将残差形态与原始星像形态差异过大的或FHWM太小的源全部舍去. 在减图的过程中假如求解卷积核出现异常, 会导致残差的形态偏离点源PSF, 据此我们对残差的形态参数做出限制.

我们把变源和暂现源分开处理, 本文只讨论变源. 我们把同时出现在模板图像和待检测图像中的源看作变源, 不满足条件的视为暂现源. 进一步地, 把多次检测到亮度出现变化的源看作变星. 叠加生成的模板图像不是一个真实时刻所拍摄的图像, 待检测图像同模板图像相减后得到残差并不代表源的亮度一定发生了变化, 只能说明同模板星像的亮度有差异. 对于一个残差均为正值或者负值的源, 由于不好辨别残差的来源, 不能简单地认为这个源在增亮或者减弱. 但是如果不同时刻的图像同模板图像相减后的残差中同时包含了正值和负值, 则可以确认源在整个观测周期内呈现增亮和变暗的过程, 可以看作变星的候选源.

2.4.1 样本构成

我们利用机器学习[33−35]对残差图像做进一步的分类, 根据残差图像的2维信息确定变星的候选源. 首先创建机器学习的训练集. 为简单起见, 我们将训练样本单纯地设为正负两类而不考虑造成残差的原因, 分别对应真源和假源. 我们认为残差形状较为规整的图像相减执行的效果比较好, 包含了星像亮度变化的信息, 可能是真源; 而残差形态偏离点源PSF, 可能对应对齐偏离、相减过程中卷积异常、宇宙线干扰、成像质量差、亮源泊松噪声影响等多种情况, 我们将其视作假源. 用于训练的正样本由挑选出来的170颗VSX星表中已知变星所对应的残差图像构成, 其特征均符合对真源的要求, 负样本由上述异常情况所对应的有代表性的图像构成, 挑选后的正负样本数量分别为2000和10000, 用于训练的样本图像大小统一为51 × 51像素, 目标星像居中, 样本不做标准化处理. 正负样本示例如图6–7.

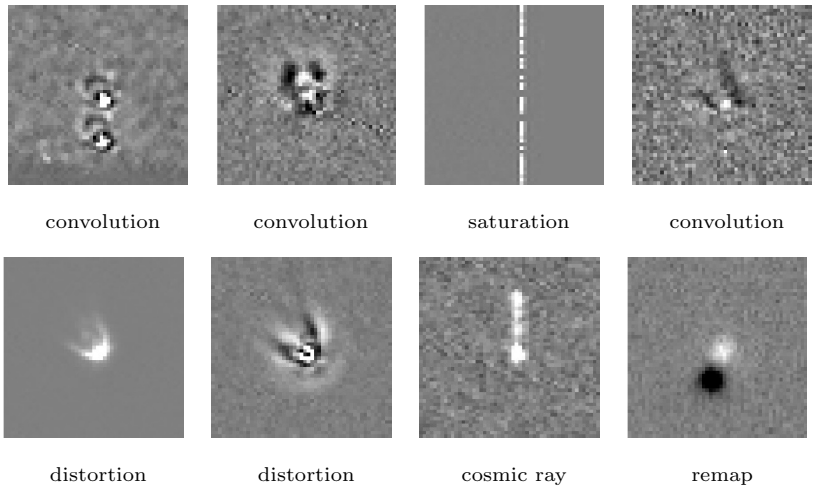

图 6　具有不规则形态的典型假源对应的残差, 可能由溢出、对齐或卷积的异常、宇宙线、望远镜系统畸变等原因造成. 每张子图的标签描述了假源产生的原因.

Fig. 6　Residual images of the typical negative samples with an irregular morphology, could be caused by saturation, errors in image registration and convolution, cosmic ray, distortion of telescope, etc. The label of each subfigure describes the cause of negative sample.

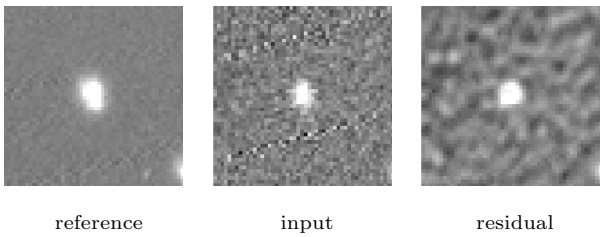

图 7　典型真源对应的模板图像、待检测图像、残差图像，可见规整的圆形轮廓

Fig. 7　Template, input, and residual images of a typical postive samples，a regular and circular profile can been seen

2.4.2　主成分分析法

机器学习需要将样本的特征作为模型的输入量, 我们的样本是分辨率51 × 51像素的灰度图像, 共计2601个特征, 数量较大, 同时图像很大一部分都是目标星像周围的背景, 并没有包含很多有用的信息, 具有较大的可压缩性, 因此可以使用主成分分析法(Principal Component Analysis, PCA)对训练集进行预处理, 提取数据的主要特征.

主成分分析法多用于数据降维, 通过将数据投影到若干个新的基矢的方向构成新的数据集, 从而达到降维的目的.

定义D维数据集:

$$\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\} \ , \tag{5}$$

$\boldsymbol{x_n}$代表了N条数据中的一条记录. PCA的目标是将数据$\boldsymbol{X}$投影到线性M维空间中去, 为了找出M维空间的基矢, 首先需要构建数据集的协方差矩阵$\boldsymbol{S}$:

$$\boldsymbol{S}=\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x_n}-\overline{\boldsymbol{x}})(\boldsymbol{x_n}-\overline{\boldsymbol{x}})^{\mathrm{T}}\,, \tag{6}$$

其中$\overline{\boldsymbol{x}}$代表了数据集的平均值. 解出该协方差矩阵的特征值和特征向量, 选择前M个最大的特征值所对应的特征向量, 将原D维数据集投影到这M个特征向量所构成的M维空间:

$$a_n=\boldsymbol{U}^{\mathrm{T}}(\boldsymbol{x_n}-\overline{\boldsymbol{x}})\,, \tag{7}$$

a_n是投影得到的M维空间中的新变量, $\boldsymbol{U}$是$D\times M$维的矩阵, 该矩阵的列对应最重要的M个主成分. 我们首先使用PCA将原始训练集降维至200维.

我们接下来对每个子类单独做主成分分析, 计算每个样本I_n在每个子类PCA方法中的重构误差ε_n:

$$\varepsilon_n=\sqrt{\frac{(I_n-\tilde{I}_n)^2}{m}}\,, \tag{8}$$

其中$\tilde{I}_n$表示使用原始图像的前20个主成分重构得到的图像, m代表图像的像素数量. 每个样本得到两个重构误差ε_{T}和ε_{F}, 分别对应真源和假源这两个子类的PCA方法, 这两个重构误差和之前得到的200个主成分一起作为样本的特征, 构成训练集.

2.4.3 随机森林

我们选择随机森林同上述主成分分析法结合使用. 随机森林由Breiman[18]于2001年提出, 是一种有监督学习算法, 其核心是自助采样法和决策树的集成. 自助采样通过有放回地从训练集中随机抽取不同的样本组成多个不同的训练集, 这种随机性可以避免出现过拟合, 同时赋予模型较强的抗噪能力; 随机森林模型中包括了许多独立工作的决策树, 各个决策树各自根据输入样本生成预测, 最后再结合各个决策树的预测生成单预测. 与单一决策树相比, 随机森林输入的是训练集的子集, 其对应的每一棵子树同决策树相比要浅, 这也使得其不容易出现过拟合. 决策树的结点依据选择的多个特征进行分裂, 使得模型的准确率得到提升. 我们借助Python机器学习库scikit-learn构建和训练随机森林模型, 子树的数量设置为1000, 训练样本的数量共10200个. 我们使用十折交叉验证法评估模型的准确性和泛化能力. K折交叉验证法(K-fold cross-validation)是指将训练集分割成K个子样本, 选取一个子样本用作验证模型的数据, 其他$K-1$个样本用来训练模型. 交叉验证重复K次, 使得每个子样本验证一次, 对K次验证的结果做平均最终得到一个单一估测. 我们将训练集分割成10个子样本, 做十折交叉验证, 并绘制对应的受试者工作特征曲线(receiver operating characteristic curve, ROC cureve)及混淆矩阵(confusion matrix), 结果如图8、表3、表4所示. 最后我们用训练得到的模型对5万多个源的残差图像进行了预测, 挑选出多次被判定为真源, 同时光度相对模板时刻有一定起伏的源, 最后我们一共得到1721颗变星候选体.

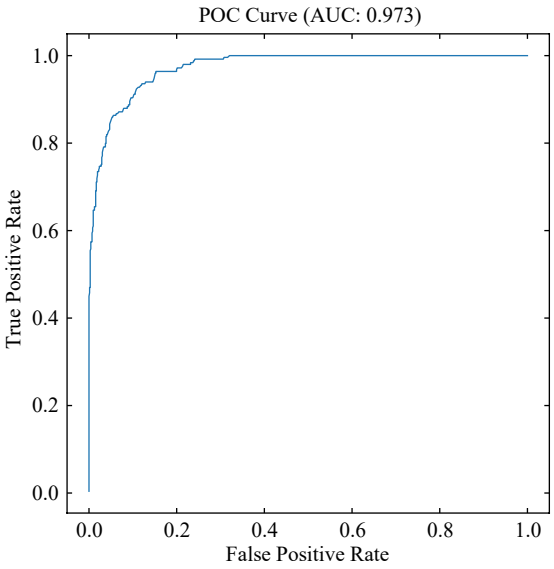

图 8　随机森林分类器的ROC曲线

Fig. 8　The ROC curve of random forest classifier

表 3　随机森林分类器的十折交叉验证法结果

Table 3　The 10-fold cross-validation result of random forest classifier

ACU	Prec	Recall
0.914	0.640	0.912
0.926	0.635	0.946
0.919	0.667	0.937
0.919	0.654	0.926
0.931	0.721	0.907
0.933	0.703	0.948
0.924	0.665	0.925
0.933	0.689	0.938
0.920	0.664	0.928
0.938	0.723	0.949

表 4　随机森林分类器的混淆矩阵

Table 4　The confusion matrix of random forest classifier

		Positive	Negative
Actual Value	True	125	10
	False	85	823

2.5 变星认证

我们采用较差测光的方法得到目标天体的光变曲线, 通过比较目标天体和周围参考星在同一时刻的星等, 得到目标天体星等的变化信息. 这么做的前提有: 目标天体和参考星之间的观测条件相近, 即参考星比较靠近目标天体, 同时假设参考星是恒星. 同时为了减小测光误差, 应该选择星等较为接近目标天体的参考星. 我们利用SExtractor软件测光得到的FLUXRADIUS参数和MAGBEST参数在以目标天体为中心半径512″的范围确定星等和星像大小与目标天体最为接近, 同时重复观测次数最多的天体中挑选3颗作为参考星, 通过APASS (The AAVSO Photometric All-Sky Survey)星表给出的参考星星等来确定变星的星等. 我们找出的变星候选源共1721颗, 其中部分源因为所在位置的关系缺乏合适的参考星, 因此我们绘制了871颗变星候选体的光变曲线, 人工对其进行确认其中52颗为已知变星. 因为观测次数的限制, 我们难以确定候选源的周期和分类, 这里仅展示部分变星候选源的位置、星等和相应的光变曲线, 如表5和图9. 部分候选源的特征比较明显, 如RA:2:19:54、DEC:−54:15:15处的候选源可能是激变变星, RA:6:32:36、DEC:−49:48:31处的源可能是长周期变星. 部分已知变星的光变曲线也一并展示, 如图10.

表 5 部分变星候选源的赤经、赤纬、星等及分类

Table 5 The RA, DEC, magnitude and classification of several candidates for variable stars

RA	DEC	magnitude	class
2:19:54	−54:15:15	14.3	CV
2:29:26	−54:42:51	14.0	LPV
5:21:56	−56:42:48	10.7	Candidate
6:32:36	−49:48:31	11.8	LPV
19:42:25	−44:21:51	13.4	Candidate
19:57:13	−41:50:11	16.2	Candidate
20:3:18	−48:44:22	14.4	Candidate
20:11:27	−45:42:35	15.6	CV
20:12:23	−43:7:52	15.7	CV

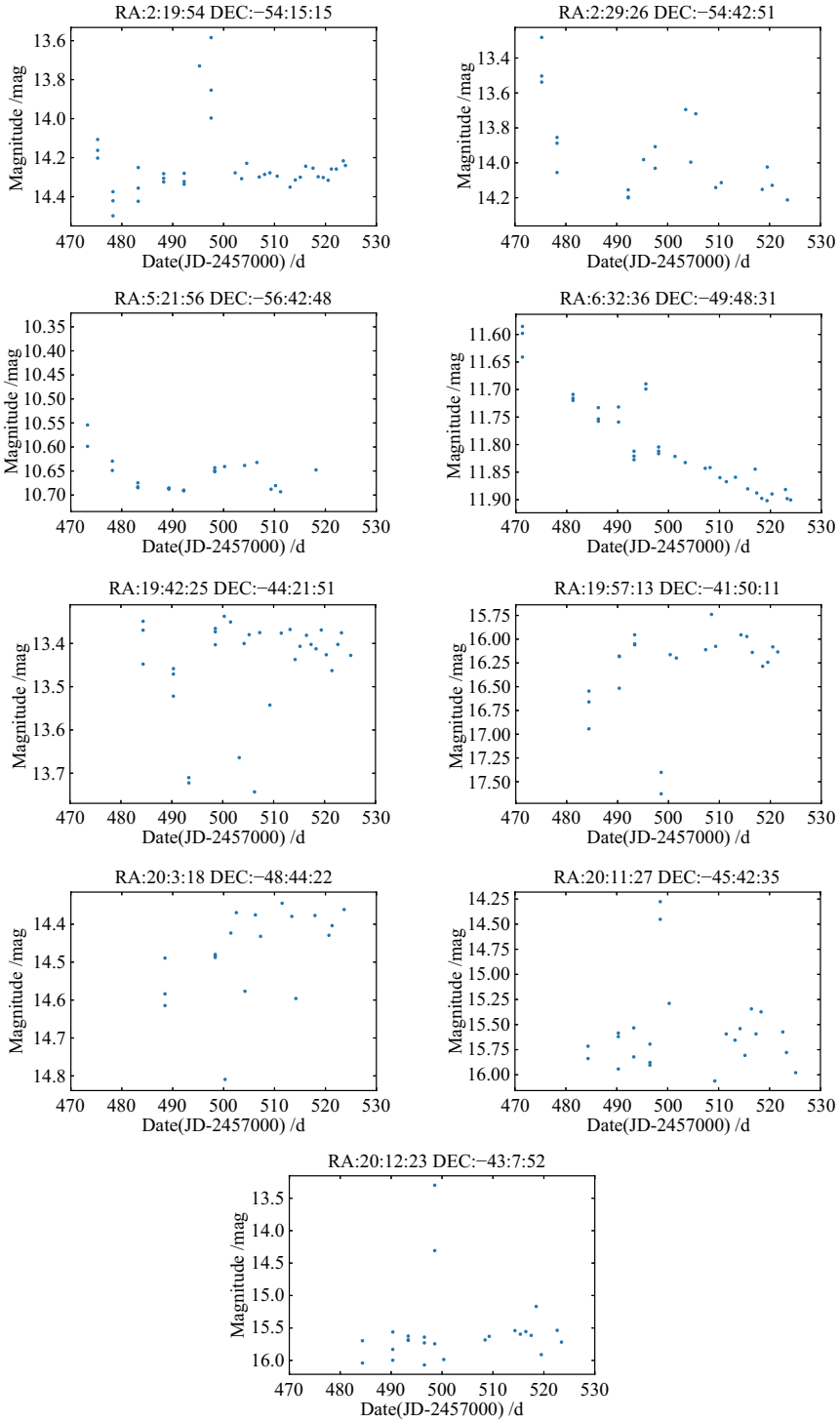

图 9　具有代表性的变星候选源光变曲线

Fig. 9　The light curves of the representative candidates for variable stars

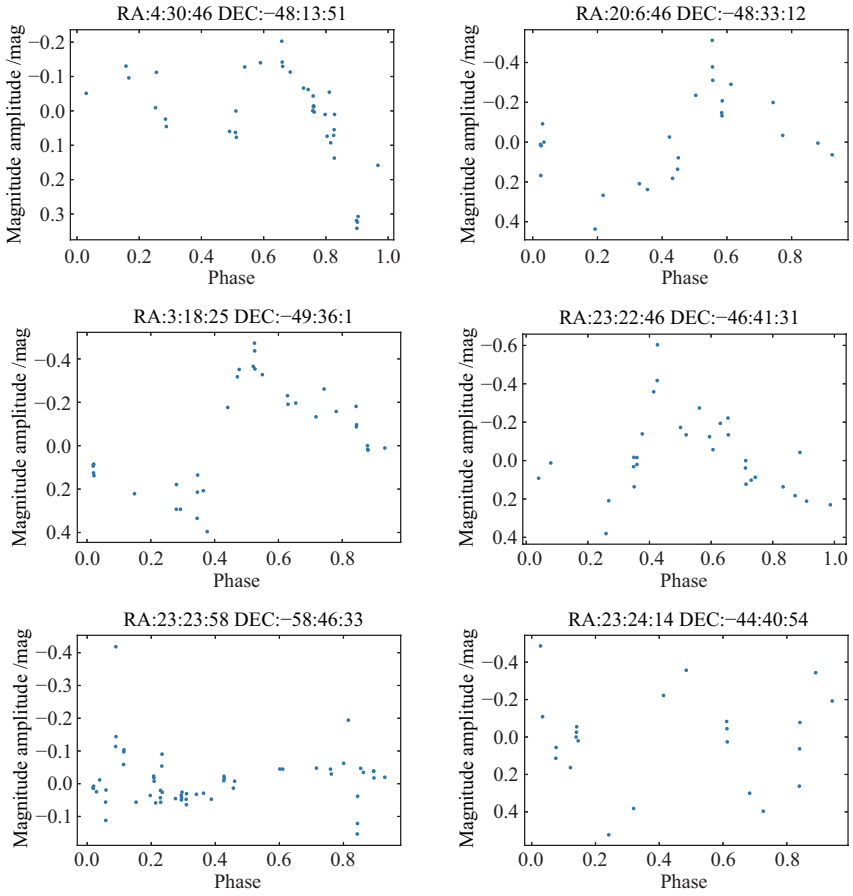

图 10 探测到的部分VSX星表中的变星的相位图

Fig. 10 The phase diagram of several variable stars detected and known in VSX catalog

3 讨论

我们用于图像相减的模板通过叠加生成之后带来一个问题: 模板的星像及其星等并不对应一个真实时刻的值, 待检测图像“时刻”的星等和模板“时刻”的星等之间出现差异不代表待检测图像时刻的星等一定发生了变化, 所以必须比较多个时刻的残差图像才能判断该源的星等是否真正发生了变化. 理想情况下, 图像相减后的残差应该已经包含了我们所需要的所有光变信息, 图像相减之前的卷积已经消除了不同时刻的零点差异, 可以通过残差图像的测光数据直接得到以模板时刻为基准的光度相对变化的信息. 我们考虑到较差测光的需求, 选择从原始测光数据中得到光变曲线. 如果能直接从残差图像中得到光变信息, 就可以一定程度上简化程序, 提高运行效率.

我们的方法对变星的探测能力有所不足, 检测出的52个已知变星均有较大的振幅和较亮的星等, 如表6所示, 对于暗弱和振幅不大的变星的探测能力较为弱.

表 6 部分探测到的VSX星表中的已知变星的赤经、赤纬、周期

Table 6 The RA, DEC and period of several variable stars detected and known in VSX catalog

RA	DEC	Period	RA	DEC	Period
4:30:46	−48:13:52	0.36	20:6:46	−48:33:15	0.56
3:18:25	−49:36:1	0.64	23:22:46	−46:41:32	0.55
23:23:58	−58:46:34	0.88	23:24:14	−44:40:54	0.55
18:54:56	−53:37:43	62.11	18:45:39	−59:38:25	0.30
4:42:3	−51:33:35	0.73	20:46:50	−44:58:5	0.57
23:14:11	−46:48:55	0.27	4:53:19	−43:13:27	0.39
19:37:3	−48:53:50	0.48	20:0:30	−45:9:6	0.31
5:29:28	−58:54:46	0.15	22:47:12	−42:44:38	0.36
5:13:59	−45:46:55	0.29	19:56:23	−48:56:58	0.64
19:16:25	−53:42:15	410.74	19:44:54	−47:0:37	0.81
23:22:46	−42:48:38	0.26	5:14:22	−53:53:56	0.55
22:44:45	−42:21:42	0.54	20:19:57	−46:40:38	0.35
19:47:44	−45:39:37	0.60	18:52:23	−54:9:58	62.40
20:5:56	−48:50:32	1.11	19:33:52	−53:1:59	241.77
20:46:7	−47:19:9	0.51	19:15:7	−53:32:39	1.34
23:20:16	−44:3:54	0.37	19:46:11	−51:31:18	444.00
19:15:50	−54:34:54	223.70	22:58:7	−40:33:18	0.31
17:31:38	−60:30:24	286.00	19:54:45	−45:13:1	0.28
19:11:59	−53:0:23	25.26	19:18:13	−54:23:32	60.23
19:43:23	−48:37:35	0.44	20:55:11	−43:21:6	3.38
5:20:45	−56:28:23	41.6	4:49:8	−49:8:6	0.40
19:41:32	−45:30:36	0.83	22:59:17	−43:49:16	0.57
20:22:1	−44:13:10	0.38	20:44:23	−45:55:244	0.70
5:17:0	−55:55:26	0.79	23:38:29	−45:12:23	0.40
4:49:8	−49:8:6	0.40	5:42:41	−53:16:33	0.23
6:10:33	−48:44:25	0.23	4:46:55	−59:3:38	0.49

从图3可以看出, 我们用于模型训练的2016年年度观测数据质量并不理想, 作为训练集的残差图像, 其PSF同模板图像相同, 考虑到PSF的差异, 训练得到的模型可能不适用于质量较好的观测数据, 泛化能力受到了训练集质量的限制. 同时较差的星像质量使得一部分图像在整个数据处理的流程中直接被弃用, 这也造成了数据量的损失. 将来加入更多的观测数据, 通过增加训练集的数量和PSF的变化范围, 则有望大幅改善这些问题, 使得模型具备更强的泛化能力, 应用到未来的观测数据处理中去.

传统的变星搜寻方法是通过目标源光变曲线参数来进行判断, 但是这种方法依赖于对周期的测量, 因此只适用于探测周期性变星, 对非周期性的激变变星没有很好的分辨能力. 我们的方法不依赖于测量周期, 因此在对激变变星的探测上具有优势.

4 总结与展望

我们通过整合图像相减法、机器学习及其他一些天文数据处理软件, 设计了一套自动处理观测数据, 进行变星搜寻的程序. 其优点是整个流程无需人工干预, 我们只需对最后的候选源进行判读, 整套程序基于成熟的算法和软件整合而来, 具备较高的可靠性, 参数调整也很灵活; 缺点是目前对暗弱及振幅较小的变星的探测能力不强, 需要调整. 受制于训练集的数量, 模型的泛化能力暂未得到验证. 我们的工作证明了通过图像相减法和机器学习进行变星搜寻的可行性, 这套程序经过完善后可以在AST3-2南极巡天望远镜的数据处理计算机上运行, 进行搜寻变星的工作, 省去大批原始数据回传的步骤.

接下来的工作重点将集中于参数调整和扩大训练样本, 以期提高整套方法的灵敏度和泛化能力, 并将该方法用于今年正在执行的2019观测季.

致谢 感谢审稿人非常有帮助的建议, 感谢袁祥岩研究员和李正阳副研究员提出的宝贵意见. 本文使用的数据来自中国南极天文中心AST3南极巡天项目. Ashley C. B. Michael感谢AAD (Australian Antarctic Division)以及Astronomy Australia Limited管理下的NCRIS (Australian National Collaborative Research Infrastructure Strategy)支持.

参 考 文 献

[1] Grindlay J, Tang S, Los E, et al. Proceedings of the International Astronomical Union, 2011, 7: 29
[2] Law N M, Kulkarni S R, Dekany R G, et al. PASP, 2009, 121: 1395
[3] Zhang T M, Wang X F, Chen J C, et al. RAA, 2015, 15: 215
[4] Yuan X Y, Cui X Q, Gu B Z, et al. SPIE, 2014, 9145: 91450F
[5] Ma B, Hu Y, Shang Z H, et al. ATel, 2016, 9033: 1
[6] Ma B, Wei P, Shang Z H, et al. CBET, 2014, 3796: 1
[7] Wang L Z, Ma B, Li G, et al. AJ, 2017, 153: 104
[8] Zhang H, Yu Z Y, Liang E S, et al. ApJS, 2019, 240: 17
[9] Zhang H, Yu Z Y, Liang E S, et al. ApJS, 2019, 240: 16
[10] Han X M, Tian Q G, Ji T, et al. Variability of the Radio-Loud NLS1 Galaxy PKS 0558-504. 224th AAS, Boston, June, 2014: 417.05
[11] Liu Q, Wei P, Shang Z H, et al. RAA, 2018, 18: 005
[12] Hu Y, Shang Z H, Ashley M C B, et al. PASP, 2014, 126: 868
[13] Li Z Y, Yuan X Y, Cui X Q, et al. SPIE, 2018, 10700: 107001L
[14] Ma B, Shang Z H, Hu Y, et al. MNRAS, 2018, 479: 111
[15] Shang Z H, Hu K L, Yang X, et al. SPIE, 2018, 10700: 1070057
[16] Alard C, Lupton R H. ApJ, 1998, 503: 325
[17] Bailey S, Aragon C, Romano R, et al. ApJ, 2007, 665: 1246
[18] Breiman L. Machine Learning, 2001, 45: 5
[19] Yuan X Y, Su D Q. MNRAS, 2012, 424: 23
[20] 李运, 杨世海. 天文学报, 2017, 58: 24
[21] Li Y, Yang S H. ChA&A, 2018, 42: 448
[22] Li X Y, Wang D X. Proceedings of the International Astronomical Union, 2012, 8: 329
[23] Hu Y, Shang Z H, Ma B, et al. SPIE, 2016, 9913: 99130M
[24] Li X Y, Yang S H, Du F J, et al. SPIE, 2018, 10700: 107005P
[25] Alard C. A&AS, 2000, 144: 363

[26] Wei P, Shang Z H, Ma B, et al. SPIE, 2014, 9149: 91492H
[27] Bertin E, Arnouts S. A&AS, 1996, 117: 393
[28] Pál A. MNRAS, 2012, 421: 1825
[29] Pál A, Bakos G Á. PASP, 2006, 118: 1474
[30] Annis J, Soares-Santos M, Strauss M A, et al. ApJ, 2014, 794: 120
[31] Hu L, Wu X F, Andreoni I, et al. Science Bulletin, 2017, 62: 1433
[32] Becker A. HOTPANTS: High order transform of PSF ANd template subtraction. 2015, record ascl: 1504.004
[33] du Buisson L, Sivanandam N, Bassett B A, et al. MNRAS, 2015, 454: 2026
[34] 黄超, 马月华, 赵海斌, 等. 天文学报, 2017, 57: 526
[35] Huang C, Ma Y H, Zhao H B, et al. ChA&A, 2017, 41: 549

An Automatic Method for Detecting Transients and Variable Sources in AST3 Survey Based on Image Subtraction and Random Forest

HUANG Tian-jun[1,2] SUN Tian-rui[1,2] HU Lei[1,2] NING Zong-jun[1]
WU Xue-feng[1] WANG Li-fan[3,1,4] WANG Xiao-feng[5,1] ZHU Zhen-xi[1]
UDDIN Ashraf Syed[6] ASHLEY Charles Brewster Michael[7]

(1 Purple Mountain Observatory, Chinese Academy of Sciences, Nanjing 210033)
(2 Department of Astronomy, University of Science and Technology of China, Hefei 230026)
(3 George P. and Cynthia Woods Mitchell Institute for Fundamental Physics & Astronomy, Texas A. & M. University, Department of Physics and Astronomy, Texas TX 77843)
(4 Chinese Center for Antarctic Astronomy, Nanjing 210033)
(5 Physics Department and Tsinghua Center for Astrophysics, Tsinghua University, Beijing 100084)
(6 The Observatories of the Carnegie Institution for Science, California CA 91101)
(7 Department of Astrophysics, University of New South Wales, New South Wales NSW 2052)

Abstract AST3-2 (Antarctic Survey Telescopes) Telescope locates in Dome A, the loftiest ice dome on the Antarctic Plateau. It produces huge amount of observation data which requires more efficient data reduction program to be developed. Also data transmission in Antarctica is much difficult, thus it is necessary to perform data reduction to detect variable sources and transient sources remotely and automatically in Antarctica, but this attempt is restricted by the poor computer performance in Antarcitca. For the realization of this aim, developing a new method based on pre-existing image subtraction method and random forest algorithm, taking the AST3-2 2016 dataset as test sample becomes an alternative choice. This method performs image subtraction on data set, then applies principle component analysis to extract the features of residual images. Random forest is used as a machine learning classifier, and a recall rate of 97% is resulted. Our work verifies the feasibility and accuracy of our method, and finally finds out a batch of candidates for variable stars in the AST3-2 2016 dataset.

Key words stars: variables: general, methods: data analysis, techniques: image processing