

引文格式: 方雷, 姚申君, 包航成, 等. 遥感影像并行处理的数据划分及其路径优化算法[J]. 测绘学报, 2019, 48(5): 572-582. DOI: 10.11947/j. AGCS. 2019. 20160524.
FANG Lei, YAO Shenjun, BAO Hangcheng, et al. An algorithm for optimizing routing of remote sensing image parallel processing based on data partitioning[J]. Acta Geodaetica et Cartographica Sinica, 2019, 48(5): 572-582. DOI: 10.11947/j. AGCS. 2019. 20160524.

遥感影像并行处理的数据划分及其路径优化算法

方 雷¹, 姚申君², 包航成³, 康俊峰⁴, 刘 婷⁵

1. 复旦大学环境科学与工程系, 上海 200438; 2. 华东师范大学地理科学学院, 上海 200241; 3. 金华市规划与地理信息中心, 浙江 金华 321000; 4. 江西理工大学建筑与测绘工程学院, 江西 赣州 341000; 5. 杭州师范大学理学院, 浙江 杭州 311121

An algorithm for optimizing routing of remote sensing image parallel processing based on data partitioning

FANG Lei¹, YAO Shenjun², BAO Hangcheng³, KANG Junfeng⁴, LIU Ting⁵

1. Department of Environmental Science and Engineering, Fudan University, Shanghai 200438, China; 2. School of Geography, East China Normal University, Shanghai 200241, China; 3. Jinhua Planning and Geomatics Center, Jinhua 321000, China; 4. School of Architectural and Surveying and Mapping Engineering, Jiangxi University of Science and Technology, Ganzhou 341000, China; 5. College of Science, Hangzhou Normal University, Hangzhou 311121, China

Abstract: Parallel processing technologies have been widely applied to remote sensing images processing. While previous research has developed many parallel algorithms for processing images, few studies have been focused on synchronous parallel processing for multiple computing tasks when one copy of remote sensing image has many redundant backups under the cloud computing environment. To bridge the research gap, this research proposes a routing optimization algorithm for parallel processing of remote sensing image. Based on data segmentation, the method is developed to solve the dynamic routing optimization problem when applying the parallel technology to remote sensing image distributed storage and processing. Following the introduction of 8 definitions (e.g. model data state, model elements, relative information quantity and matrix mapping) and 6 properties (e.g. directed, transitive, reproductive, multi-dimensional properties), a mathematical model is proposed. Under the framework, the ratio of average computation costs is used as the flag to control horizontal or vertical parallel processing. In addition, typical examples such as quadtree index generation, and quadtree-based target detection are presented for illustrating the application of our model on parallel processing. Finally, through the experiments, we verify the effectiveness of the algorithm, discussing the characteristics and influential factors of the algorithm.

Key words: parallel; remote sensing image; data generation; optimal path; GIS

Foundation support: The National Key Research and Development Program of China (No. 2016YFC0803105); The National Natural Science Foundation of China(No. 41301423)

摘 要: 研究了一种基于数据划分的遥感影像并行处理的路径优化算法, 用于解决将并行技术应用于海量遥感影像分布式存储和处理领域时其处理模型所具有的多路可达性所引起的路径动态、最优选择问题。在栅格数据可分解性分析及并行模型数据态、元素、相对信息量和映射等 8 个基本定义和 6 个性质的基础上, 给出并行处理一般数学模型。以该模型为基础获得在一般并行处理情况下, 以平均计算代价变量的比值作为控制横向并行与纵向并行选择方式的标志, 并进一步给出四叉树索引并行生成、基于四叉树的目标检测并行处理等具体示例。最后, 通过试验验证了算法的有效性, 分析了算法的特点及影响因素。

关键词: 并行; 遥感影像; 数据生成; 最优路径; 地理信息系统

中图分类号:P237

文献标识码:A

文章编号:1001-1595(2019)05-0572-11

基金项目:国家重点研发计划(2016YFC0803105);国家自然科学基金(41301423)

海量空间信息处理和应用具有信息密集和运算密集的特征,目前广泛依赖于并行技术和高性能计算机的应用。遥感影像处理(如影像的校正、配准、识别)是并行技术在空间信息处理领域的一大主要应用。早在 10 年前,国内外就已广泛开展图像的并行化处理研究^[1-10],在该领域取得多方面的成果。例如,结合同行技术和遗传算法有学者提出利用低分辨率像机获取高分辨率图像的方法^[6];部分研究人员提出能高效实现遥感图像预处理的分布式共享存储并行处理系统^[7]、遥感卫星图像几何粗校正的数据并行方法^[8]、基于小波的遥感图像全局配准算法^[9]、遥感多图像配准中自动提取特征点的并行算法^[10]、基于 GPU 的并行算法^[11-12]等,此外还有针对遥感影像识别、分割、分类、融合等高性能算法^[13-18]。针对数据量呈指数增长的全球遥感图像,必须研究基于并行化的快速、有效、高精度的自动图像配准算法,为此,有学者提出了相应的遥感图像自动配准的并行策略^[19-20]。而随着遥感影像并行化方法的研究深入,近 5 年来,越来越多的学者不满足于某一类操作或者数据的并行化处理算法,而倾向于提出并行算法的通用模型^[21-24]。

可以看出,目前研究主要还是利用遥感图像的矩阵可分解性质,设计了大量关于影像并行算法和并行处理的环境。但是,在云计算环境下,一份遥感数据存在多个冗余备份,具备同步并行处理多个计算任务,解决了对一份数据处理时输入输出(I/O)资源争用的问题。例如,可以同时同一份遥感数据的不同备份进行数据划分和压缩操作,快速生成金字塔索引。过往并没有针对该环境的最佳并行处理路径选择的深入研究。本文在总结前人研究的基础上,给出一个一般情况下的基于数据划分的最佳并行处理路径的数学模型,实现了在相同的遥感影像处理算法和同等的计算资源条件下,计算机自主选择最优处理路径,达到最优的遥感影像处理效率。

1 基于数据划分的最佳并行处理路径的数学模型

并行分为功能分解和数据分解。基于数据分解的并行处理,在遥感图像的并行处理中使用较

为普遍。例如,创建金字塔索引时,对原始影像划分的过程中同时伴随着数据压缩。本文根据栅格数据的可分解性给出栅格影像数据并行处理的相关定义,最后提出最佳并行处理路径选择算法的数学模型,即旋转门模型。

1.1 并行处理算法数学模型及其相关定义

一般的栅格影像划分存储及并行处理过程均可以抽象成如图 1 所示的模型。它包含以下四要素:数据态 R_i 、元素个数 n_i 、相对信息量 r_j 、映射 f 及其对应的单位信息量下的计算代价 ϕ_p 。

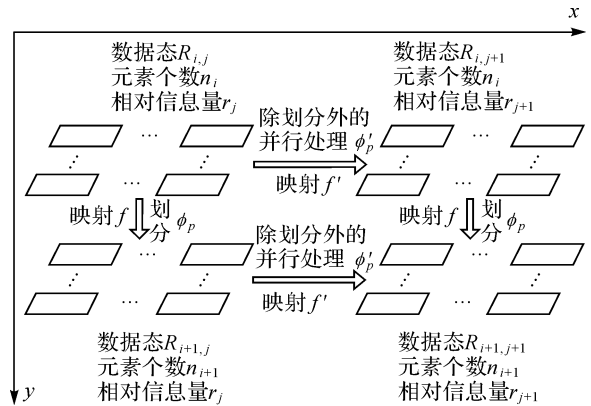


图 1 栅格影像的一般并行处理过程

Fig.1 General parallel processing model of a raster image

如果要想从源数据态 $R_{i,j}$ 生成 3 个数据态 ($R_{i,j+1}$ 、 $R_{i+1,j}$ 和 $R_{i+1,j+1}$) 必须经过一次 f 和 f' 操作以及一次 $R_{i,j}$ 到 $R_{i+1,j}$ 的操作;或一次 f' 和 f 操作以及一次 $R_{i,j}$ 到 $R_{i,j+1}$ 的操作。这两次操作可以是串行的操作也可以是并行的操作。任何一条独立的处理链无法同时(并行的)获得 3 个结果数据态,所以并行操作必须将图中 x 方向和 y 方向的操作相结合,形成多路并行处理过程:既生成 3 个结果数据态,又可以充分利用云计算平台中的多个计算节点的计算资源,提高数据处理效率。此外,虽然有两条处理路径均能生成数据态 $R_{i+1,j+1}$,但是这种情况在实际的并行过程中显然是不被允许的,因为该方式造成了计算资源的浪费。于是对并行模型进行化简,得到纵向并行和横向并行。

图 2 给出了更为一般的纵向并行和横向并行处理,显示了从源数据态经过处理链得到的不同的数据态(划分数据态和其他数据态)的过程。

图 2(a)和(b)均是一个不完全二叉树。易知,不完全二叉树上的每一个分叉都表示一种实行并行的可能性。若将每一个映射过程视为一个并行任务,则并行维数为 d 的映射过程,可以形成 d 条并行的子任务。若生成所有的数据态,均有横向并行和纵向并行两种并行路径的选择。易知,若

映射 f 不发生变化,并行路径不发生改变。在具体的并行处理实施过程中,如何让计算机根据映射 f 动态选择最佳的并行路径将是“遥感影像并行处理并对处理后的数据进行分布式存储,再对存储后的遥感影像进行下一次的并行处理”这种循环往复操作的关键问题。

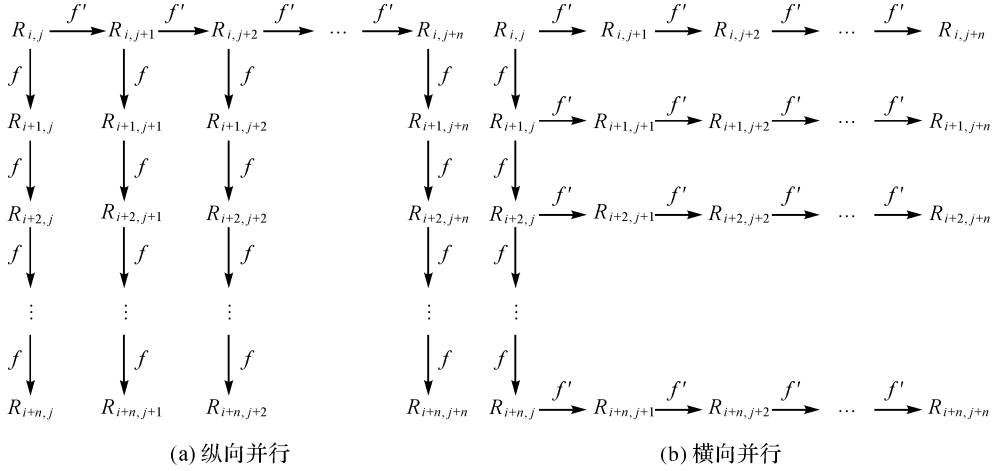


图 2 一般性纵向并行与横向并行处理

Fig.2 General longitudinal and horizontal parallel processing

经过推导^[25],采用式(1)对并行处理路径进行选择

$$\left. \begin{aligned}
 &\frac{\phi_p}{\phi'_p} > \frac{r_j \times n_i}{r_{j+1} \times n_{i+1}}, \text{ 则 } \phi_{\text{纵}} > \phi_{\text{横}}, \text{ 采用横向并行} \\
 &\frac{\phi_p}{\phi'_p} < \frac{r_j \times n_i}{r_{j+1} \times n_{i+1}}, \text{ 则 } \phi_{\text{纵}} < \phi_{\text{横}}, \text{ 采用纵向并行} \\
 &\frac{\phi_p}{\phi'_p} = \frac{r_j \times n_i}{r_{j+1} \times n_{i+1}}, \text{ 则 } \phi_{\text{纵}} = \phi_{\text{横}}, \text{ 任一种并行策略均可}
 \end{aligned} \right\} \quad (1)$$

式中, ϕ_p 为划分操作单位信息量下的计算代价; ϕ'_p 为其他操作单位信息量下的计算代价; r 为相对信息量; n 为元素个数。当 $\phi_{\text{纵}} = \phi_{\text{横}}$ 时,为临界状态。

1.2 示例及讨论

(1) 四叉树索引并行生成过程中(图 3), $\frac{n_i}{n_{i+1}} = \frac{1}{4}, \frac{r_j}{r_{j+1}} = 4$ 。 ϕ_p 代表划分操作单位信息量下的计算代价, ϕ'_p 代表压缩操作单位信息量下的计算代价。则有

$$\left. \begin{aligned}
 &\phi_p > \phi'_p, \text{ 则 } \phi_{\text{纵}} > \phi_{\text{横}} \\
 &\phi_p < \phi'_p, \text{ 则 } \phi_{\text{纵}} < \phi_{\text{横}} \\
 &\phi_p = \phi'_p, \text{ 则 } \phi_{\text{纵}} = \phi_{\text{横}}
 \end{aligned} \right\} \quad (2)$$

即①划分的计算代价大于压缩时,采用横向并行

(先划分再压缩);②划分的计算代价小于压缩时,采用纵向并行(先压缩再划分);③划分与压缩的计算代价近似时,为临界状态,采用横向并行和纵向并行均可。

(2) 基于四叉树的目标识别并行处理过程中(图 4), $\frac{n_i}{n_{i+1}} = \frac{1}{4}, \frac{r_j}{r_{j+1}} = 1+k$ 。 k 值为识别目标的信息量,大于 0,与检测出的目标及生成的矢量图形有关,具有单调递增性(即需要检测的目标越多,生成的矢量图形越多 k 值越大)。 ϕ_p 代表划分操作单位信息量下的计算代价, ϕ'_p 代表目标操作单位信息量下的计算代价。则有

$$\left. \begin{aligned}
 &\phi_p > \frac{1+k}{4} \phi'_p, \text{ 则 } \phi_{\text{纵}} > \phi_{\text{横}} \\
 &\phi_p < \frac{1+k}{4} \phi'_p, \text{ 则 } \phi_{\text{纵}} < \phi_{\text{横}} \\
 &\phi_p = \frac{1+k}{4} \phi'_p, \text{ 则 } \phi_{\text{纵}} = \phi_{\text{横}}
 \end{aligned} \right\} \quad (3)$$

即①划分的计算代价大于 $\frac{1+k}{4}$ 倍的目标识别计算代价时,采用横向并行(先划分再目标识别);②划分的计算代价小于 $\frac{1+k}{4}$ 倍的目标识别计算代价时,采用纵向并行(先目标识别再划分);③划

分的计算代价与 $\frac{1+k}{4}$ 倍的目标识别计算代价近似

时,为临界状态,采用横向并行和纵向并行均可。

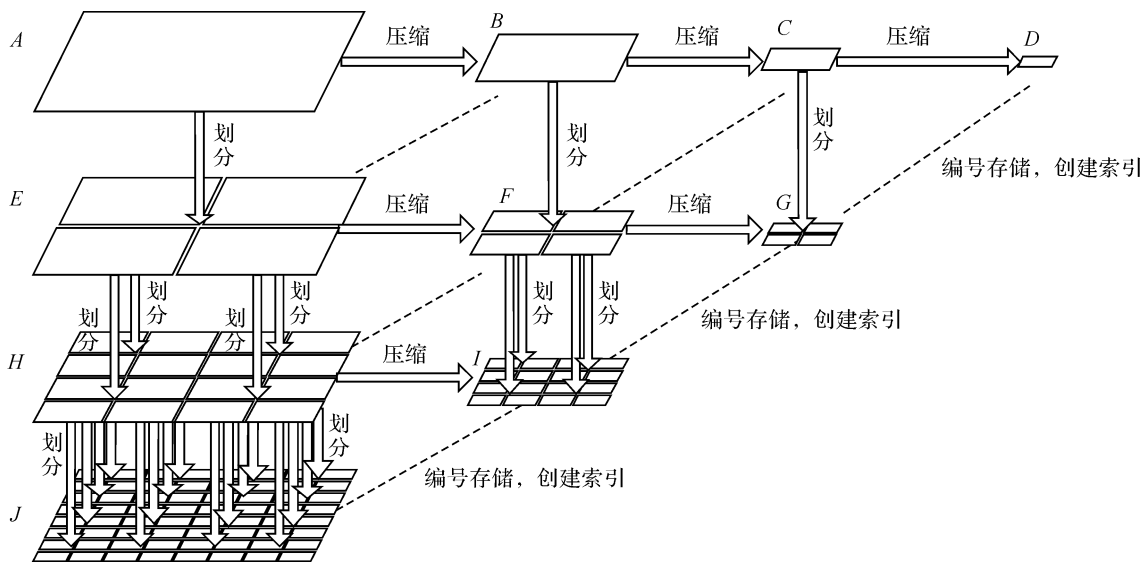


图 3 生成四叉树索引(4层)

Fig.3 Generating quadtree spatial index to image pyramid

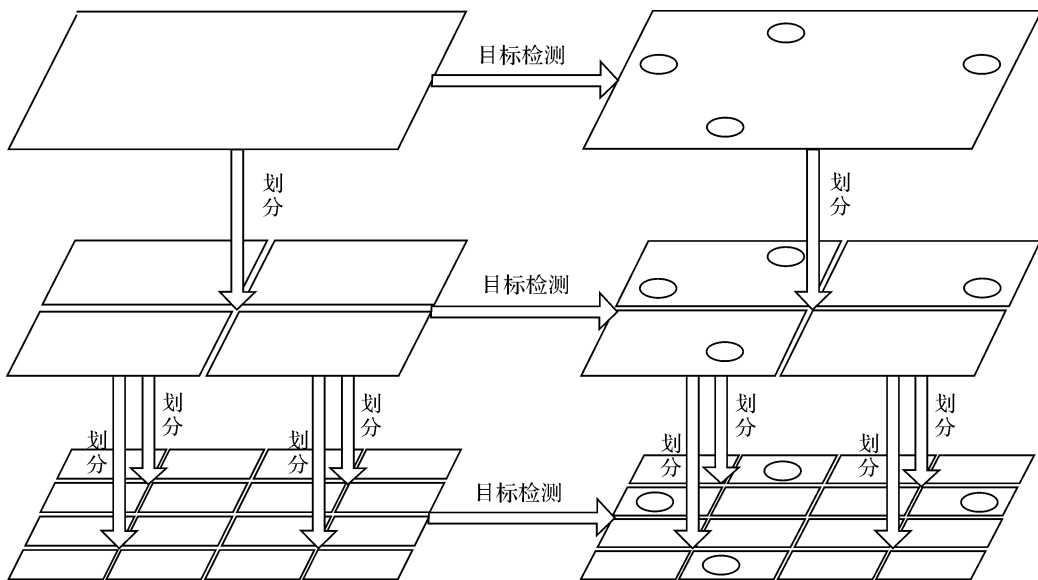


图 4 栅格影像目标识别

Fig.4 Detection and recognition parallel processing of target

对于式(3)进行进一步讨论: $k \rightarrow \infty$ 时,由于 ϕ_p 和 ϕ'_p 均为大于 0 的常数,无论 ϕ_p 和 ϕ'_p 的取值如何,均有 $\phi_3 < \phi_4$ 恒成立。也就是说,当检测结果很多时,无论划分映射过程和目标识别映射过程的效率如何,采用纵向并行(先目标识别再划分)的策略才能实现最优化并行。

(3) 横向并行和纵向并行的多种并行操作组合(图 5)。上述两个示例均为一个并行操作和划

分操作组合的并行策略分析,本文提出的最佳并行策略可以推广到多个并行。例如,图 5(a)给出了 3 个并行操作和划分操作,包含了两个纵向并行和一个横向并行组合,称为旋转门模型。旋转门模型是解决多个并行处理操作组合策略的问题,本质上是最优路径规划,而不是先后组合的问题。与“传统的、没有路径规划的”并行算法相比,旋转门模型的优势在于:可以实现在相同的遥感

影像处理算法和同等的计算资源条件下,计算机自主选择最优处理路径,达到最优的遥感影像处理效率。旋转门模型以遥感影像数据划分为基础,是这个旋转门的主轴,每一扇门都是一个纵向或横向的选择(必定包含数据划分操作部分),并且由于数据有多个冗余备份,每一扇门与其他扇门的操作都是可以同时进行的、独立并行的关系。如果不以划分为轴,这种组合起来的策略还可变

成一个彼此为边的立方体模型(图 5(b))。

此外,进一步讨论式(2)和式(3)可知,旋转门模型的路径判断准则可简单概括为“何种操作的单位计算代价大则先进行何种操作”。即,要获得全级别的栅格影像这一前提下,要实现多种并行处理操作组合的时间最优,则要优先进行计算代价大的操作。这一结论与传统的认知和经验不同,是旋转门模型的另一理论贡献。

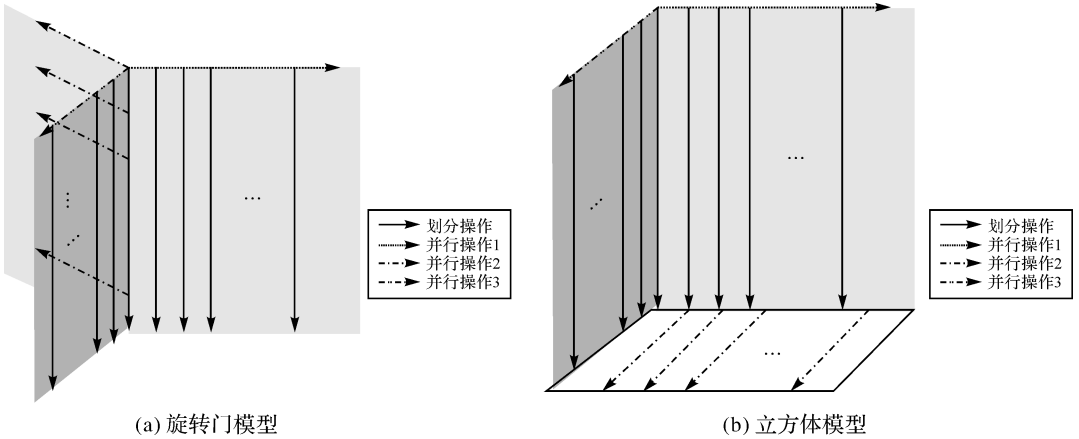


图 5 横向并行和纵向并行的多种组合

Fig.5 Combined longitudinal and horizontal parallel processing

2 算法实现流程描述及讨论

旋转门模型的编程实现过程如图 6 所示,其核心是计算输入遥感影像并行处理程序的单位信息量下的计算代价比值 ϕ_p / ϕ'_p 和每步处理遥感

影像的相对信息量和相对元素个数的比值 $(r_j \times n_i) / (r_{j+1} \times n_{i+1})$ 。据此,计算机自主选择先分割再进行其他并行操作的并行路径,还是选择先进行其他并行操作再进行分割的并行路径。

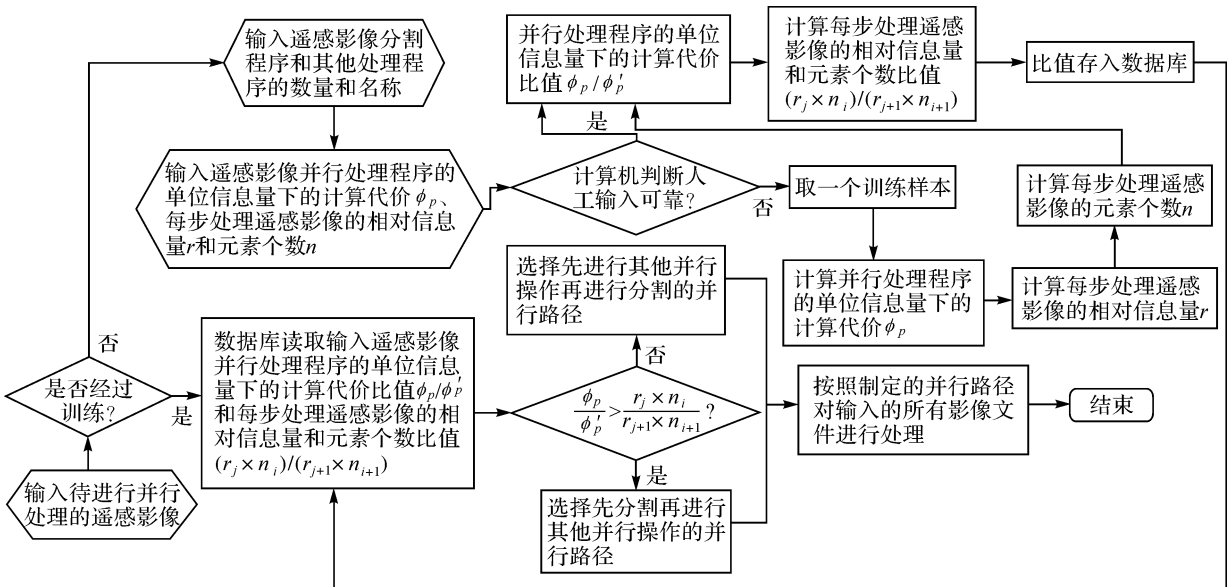


图 6 算法流程

Fig.6 Algorithm flow diagram

总之,通过实现算法的描述可知,旋转门模型原理简单实用,可以使计算机集群平台自动根据用户输入的遥感影像和指定的遥感影像处理程序准确地得到并行处理时间最短的最优路径,解决了将并行技术应用于海量遥感影像分布式存储和处理领域时其处理模型所具有的多路可达性所引起的路径动态、最优选择问题,满足了并行处理时耗时最短、最高效的要求。

特别的,在云计算或集群平台上,图 2 的纵向并行和横向并行的旋转门模型在计算节点上的实际运行过程示意图如图 7 所示。由图 7 可知,在实际运行过程中,除第 1 个计算节点之外,旋转门模型分配每个计算节点“非常有秩序地”都在做相同的操作,而且任何一个节点又都是一个不完全

二叉树的子节点的根节点,它们遵守相同的规则,环环嵌套。这样的“秩序”是根据“计算时间最少”这一标准,由公式自动计算出来,没有人工干预。每个节点都做同样的操作对庞大的云计算平台有着诸多优势。例如,在同一个计算节点内,若指令快速缓冲存储区(cache)和数据 cache 都有限时,做同样的事情可以有机会让代码和数据一直在 cache 里,从而提高效率;再如,减少数据传输次数,规避数据传输瓶颈;还有,容错率高,不会影响其他计算节点的计算结果,容易找到出错节点或问题所在,容易恢复中断的处理和操作等等。综上,旋转门模型虽然是遥感影像处理领域的一个算法,但是从一个侧面反映了“秩序即效率”的自然或社会规律。

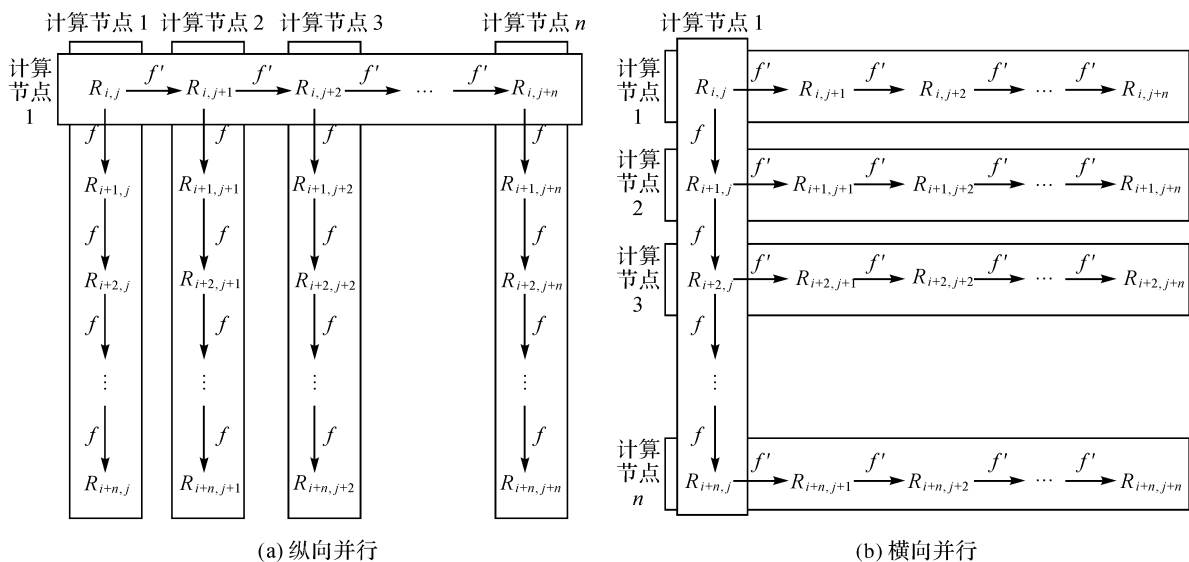


图 7 纵向并行与横向并行的旋转门模型在计算节点上的实际运行过程
Fig.7 Longitudinal and horizontal parallel processing on compute nodes

3 结果验证与分析

前文已经通过严密的理论论证提出了旋转门模型。旋转门模型的本质是多种并行处理的组合基础上的最优路径选择策略。前文在分析传统的遥感影像并行处理的缺点时,已经指出现有的并行处理策略只注重单一操作的并行处理,若完成数据划分、目标识别、数据压缩、投影变换、特征提取等处理过程,只能依次先完成数据划分再进行目标识别,然后数据压缩等。所以,第 2 节的理论模型就是在“多种遥感影像处理并行策略一定优于这些并行处理过程各自完成再串行”这一前提假设的基础上提出的。即旋转门模型与现有并行

处理策略相比具有天然的优势。换句话说,旋转门模型可以实现在相同的遥感影像处理算法和同等的计算资源条件下,计算机自主选择最优处理路径,达到最优的遥感影像处理效率。所以本文试验的目的是为论证实际遥感影像并行处理过程中基于数据划分的最佳并行路径选择模型的正确性及其性能,并未重点测试比较旋转门模型与现有并行处理策略的高低(讨论部分间接说明了旋转门模型与现有并行处理策略的优势)。试验内容以四叉树金字塔并行生成过程和目标识别为例,测试了“先压缩再划分和先划分再压缩”、“先识别再划分和先划分再识别”的运行时间,最后结合压缩算法、目标识别算法和划分算法的计算代价,对上述两个

测试结果进行深入分析。需要说明的是,数据量和软硬件环境不同会影响测试时间,但是最优路径的选择结果和结论不会发生变化。

3.1 测试环境

测试采用名字节点 3 个,配置为表 1 中第 1 及第 2 类共 3 台 PC 机;计算节点 32 个,配置为表 1 中 8 类共 32 台 PC 机。本文试验的软件环境简单,Windows 操作系统运行稳定、操作方便、安全性较高,为保证系统的正常运行,服务器端操

作系统采用 Windows 2008 Server (64 bits) 或 Windows 2003 Server (32 bits)。其中,选择 Windows 2008 Server 是为运行 Dryad 平台,选择 Windows 2003 Server (32 bits) 是用其作为 UDDI 的服务目录服务器。对客户端理论上没有要求,任何安装了网络浏览器的普通 PC 机均可作为客户端。

数据选择表 2 所示数据,其数据量从 200 MB 到 5.58 GB 不等。

表 1 PC 机及网络设备配置

Tab.1 Configuration of PC and network equipments

类别编号	CPU	内存/GB	硬盘/TB	网卡	操作系统	数量
1	Intel Core i5 @ 3.2 GB	8	2	千兆网络适配卡	Windows	1
2	Intel Core2 Duo E4500 @ 2.2 GB	4	0.25	千兆网络适配卡	2008 Server	2
3	Intel Core2 Duo E7400 @ 2.8 GB	4	0.5	千兆网络适配卡	(64 bits)	3
4	Intel Core2 Quad6600 @ 2.4 GB	4	1	千兆网络适配卡		3
5	Intel Core2 Quad8200 @ 2.3 GB	4	1	千兆网络适配卡		11
6	Intel Core2 Quad8300 @ 2.5 GB	4	1	千兆网络适配卡		1
7	Intel Pentium D @ 2.6 GB	2	0.5	百兆网络适配卡	Windows 2003	1
8	Intel Pentium4 @ 2.4 GB	1	0.27	百兆网络适配卡	Sever(32 bits)	10

表 2 试验用栅格数据

Tab.2 Experimental data

类别编号	名称	数据格式	数据量/MB	尺寸/像素	像元深度/bit	描述
A	某近海岸影像数据	TIF	197.80	7129×7151	8	多波段(4 个波段)SPOT5 校正影像
B	SAR 海洋遥感影像数据	TIF	1010	23 088×23 567	16	单波段合成孔径雷达
C	2007 年 1 月 27 日某岛融合校正图像	TIF	828.48	15 036×14 444	8	多波段(4 个波段)SPOT5 校正影像
D	嘉善县数字正射影像图 1	TIF	323.30	12 061×9373	8	1:10 000 标准分幅多波段(3 个波段)DOM
E	上虞市数字正射影像图 1	TIF	82	6097×4701	8	1:10 000 标准分幅多波段(3 个波段)DOM
F	青海互助县无人机影像图	TIF	5580	24 098×92 452	8	无人机影像(3 个波段)

3.2 测试内容

3.2.1 生成二叉树金字塔(压缩和划分组合策略测试)

测试 1:使用 10 个计算节点,对测试数据 A 采用先压缩再划分的并行策略,构建 8、10、12 层二叉树金字塔,记录总时间;使用 10 个计算节点,对测试数据 A 采用先划分再压缩的并行策略,同样构建 8、10、12 层二叉树金字塔,记录总时间。

测试 2:测定压缩和划分操作的单位信息量

下的计算代价 ϕ_p 。分别请求压缩和划分操作,则云端的计算节点分别使用 ENVI/IDL 算法执行压缩(将一份影像数据按二叉树金字塔的压缩方法进行压缩)和划分(将同一份影像数据划分 64 份)处理,分别记录栅格数据 A—F 压缩和划分时间,并计算出相同环境下单位信息量下的计算代价。

3.2.2 目标识别(目标识别和划分组合策略测试)

测试 1:本文采用 Tensorflow 与 CNN 卷积神经网络结合的目标识别方法*,使用 10 个计算

* 基于 Tensorflow 的深度学习,试验首先对遥感影像进行处理得到样本影像训练集和测试集,再将参与训练识别的影像数据输入 CNN 进行特征提取,利用分类概率和边框回归实现识别概率统计,最终实现目标识别。识别率超过 80%。试验的测试时间不包括预处理和训练测试时间,只包括最终目标识别时间。

节点,对测试数据 F 采用先目标识别再划分的并行策略,构建 2、3、4 层数据,最终将 4、16、64 份数据均匀存储在计算节点上为止,记录 1—2 级、2—3 级、3—4 级的时间;使用 10 个计算节点,对测试数据 F 采用先划分再目标识别的并行策略,同样构建 2、3、4 层数据,最终将 4、16、64 份数据均匀存储在计算节点上为止,记录 1—2 级、2—3 级、3—4 级的时间。

测试 2:测定目标识别和划分操作的单位信息量下的计算代价 ϕ_p 。分别请求目标识别和划分操作,则云端的计算节点分别使用 ENVI/IDL 算法执行目标识别(将一份影像数据按 Tensorflow 与 CNN 卷积神经网络结合的目标识别方法进行目标识别)和划分(将同一份影像数据划分 64 份)处理,分别记录栅格数据 $A—F$ 目标识别和划分时间,并计算出相同环境下单位信息量下的计算代价。

为了客观起见,本测试试验运行 50~100 次,取其均值作为试验结果。表 3 和表 5 结果为实际并行处理过程模拟,故包含传输时间。

3.3 测试结果

3.3.1 生成二叉树金字塔(压缩和划分组合策略测试)

栅格数据 A 采用不同的并行策略获得的测试结果如表 3 所示。

表 3 栅格数据 A 构建二叉树金字塔并行策略测试结果
Tab.3 Results of building quadtree pyramid parallel processing of data A min

数据	先压缩再划分	先划分再压缩
8 级	2.823	4.961
10 级	3.263	7.585
12 级	7.750	12.573

测定压缩和划分操作的单位信息量下的计算代价如表 4 所示。

表 4 单位信息量下的计算代价测定结果

Tab.4 Results of calculation cost under unit information amount

编号	划分时间 /s	时间变化 /s	压缩时间 /s	时间变化 /s	划分计算代价 /((s/(字节×像元)))	压缩计算代价 /((s/(字节×像元)))
A	6.432	—	26.499	—	1.810×10^{-8}	7.423×10^{-8}
B	25.218	18.786	103.394	76.895	1.811×10^{-8}	7.413×10^{-8}
C	9.447	3.015	38.803	12.304	1.814×10^{-8}	7.403×10^{-8}
D	7.566	1.134	31.120	4.621	1.827×10^{-8}	7.445×10^{-8}
E	6.025	-0.407	24.845	-1.654	1.824×10^{-8}	7.411×10^{-8}
F	143.690	137.258	602.220	572.721	1.819×10^{-8}	7.435×10^{-8}

3.3.2 目标识别(目标识别与划分组合策略测试)

栅格数据 F 采用不同的并行策略获得的测试结果如表 5 所示。

测定目标识别和划分操作的单位信息量下的计算代价如表 6 所示。

表 5 栅格数据 F 目标识别并行策略测试结果

Tab.5 Results of target detection and recognition parallel processing of data F h

数据	先识别再划分	先划分再识别
1 至 2 级	11.351	12.393
2 至 3 级	2.971	3.764
3 至 4 级	1.521	2.536

表 6 单位信息量下的计算代价测定结果

Tab.6 Results of calculation cost under unit information amount

编号	划分时间 /s	时间变化 /s	识别时间 /h	时间变化 /h	划分计算代价 /((s/(字节×像元)))	识别计算代价 /((s/(字节×像元)))
A	6.432	—	0.254	—	1.810×10^{-8}	1.530×10^{-5}
E	25.218	18.786	4.717	4.463	1.811×10^{-8}	1.508×10^{-5}
C	9.447	3.015	1.008	0.754	1.814×10^{-8}	1.525×10^{-5}
D	7.566	1.134	0.556	0.302	1.827×10^{-8}	1.511×10^{-5}
E	6.025	-0.407	0.145	-0.109	1.824×10^{-8}	1.550×10^{-5}
F	143.690	137.258	11.120	10.866	1.819×10^{-8}	1.536×10^{-5}

3.4 结果讨论

3.4.1 生成四叉树金字塔(压缩和划分组合策略测试)

由前文图 3 可知,若创建 n 级金字塔索引,则该索引中的文件总数 S_n 为

$$S_n = 1 + 5 + 21 + \dots + \frac{1}{3}(4^n - 1) = \frac{1}{3} \left[\frac{4}{3}(4^n - 1) - n \right] \quad (4)$$

那么,栅格数据 A 在构建 8 级、10 级、12 级金字塔索引过程中新生成 29 123 份、466 029 份、7 456 535 份图像文件。同理,若原始图像的数据量为 a ,则数据总量 m 为

$$m = \frac{4}{3} \left[n - \frac{1}{3} \left(1 - \frac{1}{4^n} \right) \right] a \quad (5)$$

那么,原数据量为 197.80 MB 的栅格数据 A 在构建 8 级、10 级和 12 级金字塔索引过程中的总数据量为 2 021.96 MB、2 549.42 MB 和 3 076.89 MB,数据量增大了 9.2 倍、11.9 倍和 14.6 倍。面对如此大量的数据增倍(数据份数和数据量),表 3 表明无论采用先压缩再划分的并行路径还是先划分再压缩的并行路径均得到较高的处理效率。表 3 还表明“先压缩再划分”的并行路径确实优于“先划分再压缩”的并行路径,从而验证了第 2 节的理论,而且金字塔生成的级数越多,先压缩再划分的并行路径的优势越明显。表 4 中对单位信息量下的划分计算代价 ϕ_p 与压缩计算代价 ϕ'_p 的测定($\phi_p < \phi'_p$),也验证了本文提出的“并行路径选择标志”来确定最优并行路径方法的正确。此外,从表 4 中的时间变化秒数和单位信息量下的计算代价可以推知,如果在进行多种遥感影像操作时,若采用传统的并行处理策略,先完成所有的数据划分再对划分好的数据进行压缩处理,其时间已经比旋转门模型的数据处理时间长,所以,从时间处理长短来说,旋转门模型是优于现有的并行处理策略的。

3.4.2 目标识别(目标识别和划分组合策略测试)

目标识别是一个数据量增加的过程,易从表 5 结果获知“先识别再划分”的并行策略远远优于“先划分再识别”的并行策略。从表 6 可知,目标识别的单位信息量下的计算代价远远大于划分,所以进行先识别再划分的策略是正确的。“先划分再识别”的策略(横向并行),相对“先识别再划分”策略会让计算节点过多地处理复杂的识别

处理操作,从而会降低整体的计算效率。此外,从表 6 可知,由于目标识别算法的单位计算代价过大,且与数据量和像元数量正相关,所以传统方法中,将原数据分割成几份再进行并行目标识别会减少数据处理的总时间。但是,旋转门模型却告诉我们,在生成全级别的栅格影像并同时实现目标识别操作时,要“先识别再划分”,然后再将各级别的结果分布存储于云计算平台上,既能保证数据完整性和目标识别的准确性,又能保证效率最高。

3.4.3 目标识别、数据压缩和数据划分综合并行操作组合

综上,若同时进行目标识别、数据压缩和数据划分综合并行组合操作,生成全级别的栅格影像时,最优路径应是“目标识别—数据压缩—数据划分”或“目标识别—数据划分”与“数据压缩—数据划分”两种选择,此路径即能保证时间效率最高,又能保证数据操作的正确性。若先进行数据压缩或数据划分操作,则目标识别则需要重新训练样本,除时间效率低下之外,还会增加多余的操作时间。

4 结论

得益于云计算理论和技术的提出,遥感影像的处理可以实现对同一份数据进行多任务的并行处理。本文在此前提下,深入研究了基于数据划分的遥感影像并行处理问题,提出了 8 个定义(数据态、信息量、元素、映射过程、表达式、并行维数、计算代价与纵向/横向并行),6 个性质(有向性、传递性、繁殖性、多维性、同一性与多路可达性)。以这些定义和性质为基础,从相邻数据态之间的一次映射逐步推导致“基于数据划分的遥感影像并行处理”的一般情况,提出最佳并行路径选择模型——旋转门模型,用于解决将并行技术应用于遥感影像分布式存储和处理领域时,其处理模型所具有的多路可达性所引起的路径动态、最优选择问题。本文的贡献在于:在批量处理海量遥感数据的情形下,同时实现遥感数据生成、目标识别等并行任务时,最优并行路径的选择只与平均计算和数据态的相对信息量、元素个数的比值这一标志有关,而且从原始数据(源数据态 1)到最终数据(最终数据态 n)的诸多并行路径中,看似复杂,其实只有独立上行和独立下行组成的单向最优路径。

本文还进一步给出四叉树索引并行生成、基于四叉树的目标识别并行处理的研究结果。指出在四叉树索引并行生成过程中,如果划分的计算代价

大于压缩时,采用先划分再压缩的横向并行;如果划分的计算代价小于压缩时,采用先压缩再划分的纵向并行。基于四叉树的目标识别并行处理操作中,当检测结果很多时,无论划分映射过程和目标识别映射过程的效率如何,采用纵向并行(先目标识别再划分)的策略才能实现最优化并行。

本文提出了基于数据划分的遥感影像并行处理问题的基本概念和性质,最终立足于最佳并行路径选择问题上。在模型的提出上,使用了简单直观的遥感影像二维矩阵来抽象表示遥感数据(可以理解为 TIFF 格式),并在此基础上提出概念、性质和公式推导,最后得出结论。实际上,遥感数据在云计算平台中存储方式多样,如果把一个二维矩阵按行重新组织成一维直线(1个像元高度)。原本二维矩阵里读取一行或一列都很方便,现在,在一维直线中读取一行可以直接根据偏移量连续读取,但是如果读一列需要先算好起始点和长度,然后每隔固定像元读取一个像元,这种情况下就无法批量读取;同样如果将二维矩阵按列重新组织成一维直线,则按行批量读取时也会有问题。得益于云计算环境中同一份数据的多个冗余备份特性,在实际应用中,将遥感影像既按行存储也按列存储,并设计调度器对不同的遥感影像处理程序进行调度,依据需要来处理按行存储或按列存储的一维数据。所以,实际存储方式并不影响本文提出的算法模型及相关理论。只是基于不同的遥感数据存储结构或存储方式,需要对模型进行具体化。同时针对不同的遥感影像处理方法,也需要对模型进行具体化,还要考虑很多算法在实际应用时的细节,特别是容错性和健壮性问题,这些都将是未来的研究工作的重点。此外,基于本文提出的概念和性质,遥感影像并行处理问题还有很多其他的内容可以研究,例如,负载均衡、质量检验等,而这些将在未来的研究工作中深入展开。

参考文献:

[1] 吕捷,张天序,张必银. MPI并行计算在图像处理方面的应用[J]. 红外与激光工程, 2004, 33(5): 496-499.
LÜ Jie, ZHANG Tianxu, ZHANG Biyin. Applications of MPI parallel-computing on image processing[J]. Infrared and Laser Engineering, 2004, 33(5): 496-499.

[2] 方金云,何建邦. 并行栅格数据处理网格服务节点软件的关键技术[J]. 地球信息科学学报, 2004, 6(1): 17-21.
FANG Jinyun, HE Jianbang. Key technology of implementation of computing node software for parallel processing of raster data[J]. Geo-Information Science, 2004, 6(1):

17-21.

[3] CROOKES D. Architectures for high performance image processing: the future[J]. Journal of Systems Architecture, 1999, 45(10): 739-748.

[4] RITTER G X, GADER P D. Image algebra techniques for parallel image processing [J]. Journal of Parallel and Distributed Computing, 1987, 4(1): 7-44.

[5] CHALERMWAT P. High-performance automatic image registration for remote sensing [D]. Fairfax, Virginia: George Mason University, 1999.

[6] 刘志军,丁明跃,周成平,等. 基于并行遗传算法的图像超分辨率复原[J]. 中国图象图形学报, 2004, 9(1): 62-68.
LIU Zhijun, DING Mingyue, ZHOU Chengping, et al. A parallel genetic algorithm for image super-resolution restoration [J]. Journal of Image and Graphics, 2004, 9(1): 62-68.

[7] 陆松,刘光明. 分布共享存储的遥感图像并行预处理系统结构研究[J]. 计算机工程与科学, 2004, 26(10): 56-59.
LU Song, LIU Guangming. Research on the parallel pre-processing architecture with distributed shared memory for remote sensing images[J]. Computer Engineering & Science, 2004, 26(10): 56-59.

[8] 张发存,王忠,赵晓红,等. 遥感卫星图像几何粗校正的数据并行方法研究[J]. 计算机研究与发展, 2004, 41(7): 1200-1206.
ZHANG Facun, WANG Zhong, ZHAO Xiaohong, et al. A data parallel algorithm on geometric correction of remote-sensing satellite images [J]. Journal of Computer Research and Development, 2004, 41(7): 1200-1206.

[9] 周海芳,唐宇,何凯涛,等. 基于小波的遥感图像全局配准算法研究及其并行实现[J]. 自动化学报, 2004, 30(6): 880-889.
ZHOU Haifang, TANG Yu, HE Kaitao, et al. An automatic global registration algorithm based on wavelet and its parallel implementation[J]. Acta Automatica Sinica, 2004, 30(6): 880-889.

[10] 郑明玲,周海芳,刘衡竹,等. 遥感多图像配准中自动提取特征点的并行算法[J]. 计算机工程与科学, 2004, 26(10): 45-48.
ZHENG Mingling, ZHOU Haifang, LIU Hengzhu, et al. A parallel algorithm of automatically selecting the characteristic points in the multiple image registration for remote sensing[J]. Computer Engineering and Science, 2004, 26(10): 45-48.

[11] 肖汉,张祖勋. 基于 GPGPU 的并行影像匹配算法[J]. 测绘学报, 2010, 39(1): 46-51.
XIAO Han, ZHANG Zuxun. Parallel image matching algorithm based on GPGPU [J]. Acta Geodaetica et Cartographica Sinica, 2010, 39(1): 46-51.

[12] 王宗跃,马洪超,明洋. 基于 GPGPU 的全波形并行分解算法[J]. 遥感学报, 2014, 18(6): 1217-1222.
WANG Zongyue, MA Hongchao, MING Yang. GPGPU-

- based parallel algorithm for full waveform decomposition [J]. *Journal of Remote Sensing*, 2014, 18(6): 1217-1222.
- [13] 郑肇葆. 遗传算法与单纯形法组合的影像纹理分类方法 [J]. *测绘学报*, 2003, 32(4): 325-329. DOI: 10.3321/j.issn:1001-1595.2003.04.008.
- ZHENG Zhaobao. Classification of image texture by combination of genetic algorithm and simplex method [J]. *Acta Geodaetica et Cartographica Sinica*, 2003, 32(4): 325-329. DOI: 10.3321/j.issn:1001-1595.2003.04.008.
- [14] 常方正, 赵银娣, 刘善磊. 遥感影像 CVA 变化检测的 CUDA 并行算法设计 [J]. *遥感学报*, 2016, 20(1): 114-128.
- CHANG Fangzheng, ZHAO Yindi, LIU Shanlei. CUDA parallel algorithm for CVA change detection of remote sensing imagery [J]. *Journal of Remote Sensing*, 2016, 20(1): 114-128.
- [15] 董志鹏, 王密, 李德仁, 等. 利用对象光谱与纹理实现高分辨率遥感影像云检测方法 [J]. *测绘学报*, 2018, 47(7): 996-1006. DOI: 10.11947/j.AGCS.2018.20170690.
- DONG Zhipeng, WANG Mi, LI Deren, et al. A cloud detection method for high resolution remote sensing imagery based on the spectrum and texture of objects [J]. *Acta Geodaetica et Cartographica Sinica*, 2018, 47(7): 996-1006. DOI: 10.11947/j.AGCS.2018.20170690.
- [16] 刘婧, 李培军. 结合结构和光谱特征的高分辨率影像分割方法 [J]. *测绘学报*, 2014, 43(5): 466-473. DOI: 10.13485/j.cnki.11-2089.2014.0087.
- LIU Jing, LI Peijun. A high resolution image segmentation method by combined structural and spectral characteristics [J]. *Acta Geodaetica et Cartographica Sinica*, 2014, 43(5): 466-473. DOI: 10.13485/j.cnki.11-2089.2014.0087.
- [17] 董志鹏, 王密, 李德仁. 一种融合超像素与最小生成树的高分辨率遥感影像分割方法 [J]. *测绘学报*, 2017, 46(6): 734-742. DOI: 10.11947/j.AGCS.2017.20160514.
- DONG Zhipeng, WANG Mi, LI Deren. A high resolution remote sensing image segmentation method by combining superpixels with minimum spanning tree [J]. *Acta Geodaetica et Cartographica Sinica*, 2017, 46(6): 734-742. DOI: 10.11947/j.AGCS.2017.20160514.
- [18] 黄波, 赵涌泉. 多源卫星遥感影像时空融合研究的现状及展望 [J]. *测绘学报*, 2017, 46(10): 1492-1499. DOI: 10.11947/j.AGCS.2017.20170376.
- HUANG Bo, ZHAO Yongquan. Research status and prospect of spatiotemporal fusion of multi-source satellite remote sensing imagery [J]. *Acta Geodaetica et Cartographica Sinica*, 2017, 46(10): 1492-1499. DOI: 10.11947/j.AGCS.2017.20170376.
- [19] 周海芳, 刘光明, 郑明玲, 等. 遥感图像自动配准的串行与并行策略研究 [J]. *国防科技大学学报*, 2004, 26(2): 56-61.
- ZHOU Haifang, LIU Guangming, ZHENG Mingling, et al. A research on serial and parallel strategies of the automatic image registration for remote sensing [J]. *Journal of National University of Defense Technology*, 2004, 26(2): 56-61.
- [20] 胡晓东, 骆剑承, 沈占锋, 等. 高分辨率遥感影像并行分割结果缝合算法 [J]. *遥感学报*, 2010, 14(5): 917-927.
- HU Xiaodong, LUO Jiancheng, SHEN Zhanfeng, et al. Data sewing algorithm for parallel segmentation of high-resolution remotely sensed image [J]. *Journal of Remote Sensing*, 2010, 14(5): 917-927.
- [21] 杨景辉. 遥感影像像素级融合通用模型及其并行计算方法 [J]. *测绘学报*, 2015, 44(8): 943. DOI: 10.11947/j.AGCS.2015.20150059.
- YANG Jinghui. The generalized model and parallel computing methods for pixel-level remote sensing image fusion [J]. *Acta Geodaetica et Cartographica Sinica*, 2015, 44(8): 943. DOI: 10.11947/j.AGCS.2015.20150059.
- [22] 史园莉, 李海涛, 宋朝达, 等. 一种基于通用模型的遥感影像并行处理算法——以 PCA 融合为例 [J]. *遥感信息*, 2011(3): 14-18, 120.
- SHI Yuanli, LI Haitao, Song Chaoda, et al. A remote sensing data parallel processing algorithm based on general model: with an example of PCA fusion [J]. *Remote Sensing Information*, 2011(3): 14-18, 120.
- [23] 刘坡, 龚建华. 大规模遥感影像全球金字塔并行构建方法 [J]. *武汉大学学报(信息科学版)*, 2016, 41(1): 117-122.
- LIU Po, GONG Jianhua. Parallel construction of global pyramid for large remote sensing images [J]. *Geomatics and Information Science of Wuhan University*, 2016, 41(1): 117-122.
- [24] 祁昆仑. 基于视觉特征的高分辨率光学遥感影像多任务分类研究 [J]. *测绘学报*, 2017, 46(6): 802. DOI: 10.11947/j.AGCS.2017.20170081.
- QI Kunlun. Multi-task classification of high resolution optic remote sensing images based on visual features [J]. *Acta Geodaetica et Cartographica Sinica*, 2017, 46(6): 802. DOI: 10.11947/j.AGCS.2017.20170081.
- [25] 方雷. 基于云计算的土地资源服务高效处理平台关键技术探索与研究 [D]. 浙江大学 2011.
- FANG Lei. Exploration and Study on Key Technologies of High Performance Platform for Land Resource Service Based on Cloud Computing [D]. Zhejiang University 2011.

(责任编辑:丛树平)

收稿日期: 2016-10-20

修回日期: 2018-08-22

第一作者简介: 方雷(1982—), 男, 博士, 研究方向为遥感影像高效处理及信息提取。

First author: FANG Lei(1982—), male, PhD, majors in high-efficiency remote sensing image processing and information extraction.

E-mail: fanglei@fudan.edu.cn

通信作者: 姚申君

Corresponding author: YAO Shenjun

E-mail: shenjun_yao@126.com