基于 MPI 的遥感影像高效能并行处理方法研究

沈占锋[1)] 骆剑承[1)] 陈秋晓[2)] 盛 昊[1)]

[1)](中国科学院遥感应用研究所，北京 100101) [2)](浙江大学区域与城市规划系，杭州 310027)

摘 要 采用基于不同尺度下的面向特征基元的影像分析方法对高分辨率遥感影像进行基于 MPI 的处理，即在对常规的影像数据划分方法进行总结分析的基础上，提出了基于特定环境下的非均匀数据划分策略；在进行基于影像数据库的 MPI 并行处理时，提出了一种新的数据流分配方法。处理结果表明，这两种方法均能够在一定环境下取得比常规方法更高的效率。

关键词 MPI 并行计算 信息提取 尺度 数据划分

中图法分类号：TP393.09 **文献标识码**：A **文章编号**：1006-8961(2007)12-2132-05

High-efficiency Remotely Sensed Image Parallel Processing Method Study Based on MPI

SHEN Zhan-feng[1)], LUO Jian-cheng[1)], CHEN Qiu-xiao[2)], SHENG Hao[1)]

[1)](*Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing* 100101)

[2)](*Department of Regional and Urban Planning, Zhejiang University, Hangzhou* 310027)

Abstract This paper presents the method which improved the efficiency of information extraction based on feature unit of high-resolution remotely sensed image. To improve the precision of image processing, this paper applied image rough-classification based on large scale and precise-segmentation based on different scales. This paper used parallel computing method to improve the speed of image processing. For the data partition method of parallel computing of remotely sensed image, this paper summarized the general data partition methods and gave the general impelmentaiton method of data symmetrical partition method. After the characteristic analysis of the some special of remotely sensed image, this paper gave the mechanism of improving the efficiency of data partition and presented a new scale data asymmetric partition method, and gave the analysis and implementation of the new method. For the image parallel processing based on remotely sensed image database, this paper presented a new data distributing method. The analysis results show that the new methods can improve the efficiency of parallel computing for some special remotely sensed image in the special condition.

Keywords MPI, parallel computing, information extraction, scale, data partition

1 引 言

在对遥感影像进行高效能的处理中，MPI (message passing interface)是常采用的方法之一，而在采用 MPI 进行影像处理过程中，其数据划分及其传输方法也直接影响着影像处理的最终效率。在进行高分辨率遥感影像处理过程中，由于影像数据量大、算法复杂，且影像上目标信息与非目标噪声信息混杂，常常耗用较大的计算量与时间，因此需首先采用基于尺度变化下的特征基元分析的方法，以更好地适应影像信息提取的精度与速度的要求[1,2]；然后在此基础上，对影像上有特征目标(如水域、林地等)的影像并行数据划分策略进行了改进，提出了

基金项目：国家自然科学基金项目(40601057)；中国科学院资源与环境信息系统国家重点实验室开放基金项目(A0615)；天津市科技发展计划重大项目(06YFGZGX17900)

收稿日期：2005-10-30；**改回日期**：2006-09-10

第一作者简介：沈占锋(1977～)，男。2005 年于中国科学院地理科学与资源研究所获博士学位，2005～2007 年在中国科学院遥感应用研究所从事博士后研究。主要从事遥感图像处理与理解、并行计算与分布式计算等方面研究工作，发表相关论文 30 余篇。E-mail：shenzf@irsa.ac.cn

一种新的基于尺度的影像数据不平均分配策略;而对于基于影像数据库的数据处理情况,则摒弃了 MPI 提供的数据流传输方法,并提出一种新的数据流分配方法,以便能够在此基础上更进一步地提高影像并行处理的效率。

2　遥感影像处理方法介绍

2.1　基于特征基元的影像处理方法

常规的遥感影像信息提取过程主要是对组成影像的各像元进行处理与分析的过程,即采用某种信息提取算法对像元进行计算,进而得到遥感影像信息提取的结果。随着遥感信息处理技术的发展,基于特征基元的影像处理方法被提出并得到了更广泛的应用[3]。特征基元是指由遥感影像上相互连通的一系列具有相同或相似特征的像元所组成的较大的斑块区域,这些特征包括光谱、纹理、空间组合关系特征等[4]。与基于像元的影像计算过程相比,基于特征基元的影像计算方法可通过将影像分解成为多个具有不同特征的特征基元,而各基元间仍保持相对独立的特性,并且通过基元特征的分析可以获得影像上地物的特征信息,这更符合人们的目视解译过程,且符合面向对象的特点,其实现技术包括遥感影像的影像分割及边缘检测等。

2.2　基于尺度变换的影像处理方法

在影像分析过程中,采用不同的尺度大小作为研究的基本单元,可以得到不同的影像处理结果。笔者将影像中的不同基元体进行归类,同时采用由大尺度到小尺度过渡的分析过程进行影像分析,首先总结出几种大尺度的区域,即城市建筑区、农业绿地区、山地林业区、裸地区域及水区等[5~7],然后在此基础上进行小尺度特征基元的提取,如楼房、交通枢纽等基元一般存在于城市建筑区内,而轮船基元一般存在于水区等[5]。这一方法的实现原理就是先通过遥感影像的大尺度的分析过程,去除掉用户不感兴趣的区域,并将其作为背景信息,再通过小尺度的影像分析过程来完成影像的处理与理解。它的优点在于节省了大量数据的处理时间,由于避免了用户不感兴趣区域的计算量,从而提高了信息提取的效率。

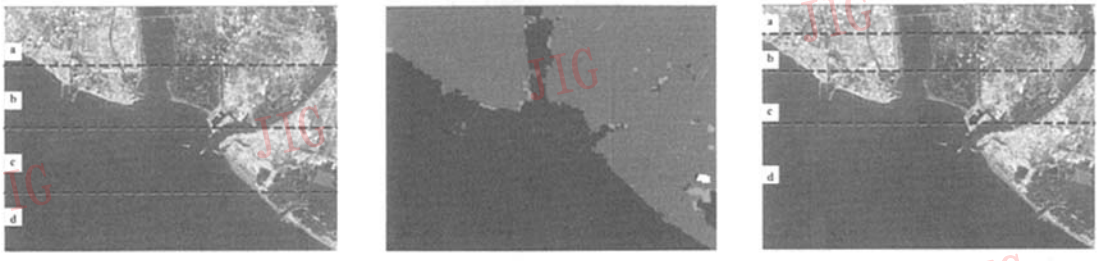
2.3　基于 MPI 的影像并行处理方法

并行计算就是将一个大的计算问题分解为许多部分,在许多处理器(机)单元上同时进行计算的方法,其目的在于加快问题的计算速度,达到影像高效能处理的目的[8]。本文采用基于 MPI 的标准并行编程环境,基于集群的多机处理系统对高分辨率遥感影像进行信息提取。MPI 是基于集群方式应用较多的一种并行计算环境,它在同一操作系统下的并行处理过程具有效率高、自动化程度好、稳定性好等特点,由于数据的划分与传输方法是影响并行计算效率的主要因素之一[2,9],本次研究侧重研究了遥感影像并行处理数据划分方法及传输过程,并结合实验对数据划分结果进行了分析。

3　遥感影像高效能并行处理方法研究

3.1　影像数据不均匀划分策略

基于集群(cluster)的影像并行计算,其主要实现过程是主进程将一个任务及待处理的数据分派到若干个子进程上分别处理,再由主进程负责收集不同子进程的数据处理结果,并进行组合,以达到多处理器共同完成某一任务的目的。由于各处理器在角色上的对等的,所以一般的数据划分方法为数据的平均划分方法[10],如图 1(a)所示(图中以 4 个并行处理机的数据划分为例)。



(a) 常规的数据平均划分法　　(b) 经大尺度影像分割后的结果　　(c) 图(b)再进行数据划分的结果

图 1　基于 MPI 的遥感影像并行处理数据划分策略示意图

Fig. 1　Data partition method of remotely sensed image parallel processing based on MPI

假设影像宽度为 w,高度为 h,并行处理环境由 N 个节点结成。本次研究中提出了一种新的数据不平均划分策略,图 1(a)为用常规的数据平均划分方法分割的结果,图 1(b)为原始影像经大尺度影像分割后所得到的结果,其实现方式主要是在各种先验知识的辅助下,利用纹理、光谱特征对影像进行大尺度的 SVM 分类得到的影像,图 1(c)为基于图 1(b)划分后再进行数据划分的结果。这种方式适用于影像数据量比较大,影像中含有大面积的非目标区域(如水体、林地、农业用地等的图像)。图 1(c)是图 1(a)在经大尺度分类形成如图 1(b)中的水体与非水体后,将水体视为背景区,并将非水体按面积相等的原则进行数据划分的结果,这样各个并行节点处理目标区域上所需的时间就近似相同。假设影像中水体面积为 A,则非水体的面积为 $h\times w-A$,在影像处理过程中,当只需要对非水体区域进行分析时,则可以认为影像中的水体部分中不含任何所需信息,与待划分影像各部分对应的面积应满足

$$A_i=\frac{h\times w-A}{N}\quad(i=1,\cdots,N)\tag{1}$$

根据式(1)所得的面积 A_i 即可确定各处理节点的数据分配。在实际实现算法过程及数据处理的过程中,由于水域部分的处理时间可近似为零,而且不同大小的数据块的传输部分在局域网环境中的时间差别不大,与数据处理所需的时间相比,这部分的时间差也可以近似忽略,因此,可以采用式(1)的非水域区域平均分配的方法进行数据分配。式(1)的实现过程如下:

```
int area = 0;
int row[N];
int num = 0;
for(int i = 0;i < h;i + + )
{
  for(int j = 0; j < w; j + + )
    if(F(i, j) is not water)
      area + + ;
  if(area > = (h × w - A)/N)
  {
    row[num + + ] = i;
    area = area - (h × w - A)/N;
  }
}
```

该算法中函数 $F(i,j)$ 是点 (i,j) 处的像元所对应的特征值,$row[k]\ (k=0,\cdots,N)$ 表示第 k 块数据的起始位置(行)。

对于不重叠的数据划分方法,用上述算法在求取 $row[k]\ (k=0,\cdots,N)$ 后,就可以根据 $row[k]$ 的值来确定不同节点的数据分配,即

$$Data(k)=D(row[k-1],row[k])\quad(k=1,\cdots,N)\tag{2}$$

式(2)中,$Data(k)$ 表示第 k 个节点上应该分配的数据,$D(s,e)$ 表示第 s 行至第 e 行间的影像数据。

类似地,对于基于 $\omega\times\omega$ 大小的窗口的数据处理算法的数据划分方法,其公式可表示为

$$\begin{cases}Data(1)=D\left(0,\dfrac{\omega-1}{2}\right)+D\left(0,row[1]+\dfrac{\omega-1}{2}\right)\\ Data(k)=D\left(row[k-1]-\dfrac{\omega-1}{2},row[k]+\dfrac{\omega-1}{2}\right)\quad(k=2,\cdots,N-1)\\ Data(N)=D\left(row[N-1]-\dfrac{\omega-1}{2},row[N]\right)+D\left(row[N]-\dfrac{\omega-1}{2},row[N]\right)\end{cases}\tag{3}$$

式(2)及式(3)确立了基于尺度的非平均数据块划分策略。基于这种数据分配策略,在局域网的集群环境下实现了基于尺度的遥感影像并行数据划分,相关的测试结果如表 1 所示。

表 1 中的数据为测试 10 次所得结果的平均值,测试环境为:局域网内由 4 台 Windows XP 系统组成的 MPI 并行计算环境,MPI 版本为 1.2.5,采用的基于窗口的影像处理算法(高斯马尔可夫随机场影像分割算法),对 16MB 的高分辨率遥感影像进行测试,表中的 4 个数据分别是主进程的时间消耗与其他 3 个节点的计算时间消耗,由于主进程的数据传输为本进程内的内存复制,所以其数据传输时间为 0;同样由于主进程还需要对各子进程的结

表 1 基于 MPI 的遥感影像并行数据划分策略效率测试表

Tab.1 Efficiency test of different data-partiton method

数据划分方法	两种数据划分方法效率比较	
	数据传输时间(ms)	影像处理时间(ms)
数据平均分配策略	0 360 390 500	32341 25783 30284 32221
数据不平均划分策略	0 272 400 533	23119 17125 19901 19560

果进行收集,所以主进程读出的数据处理时间比其他进程也要多一些。

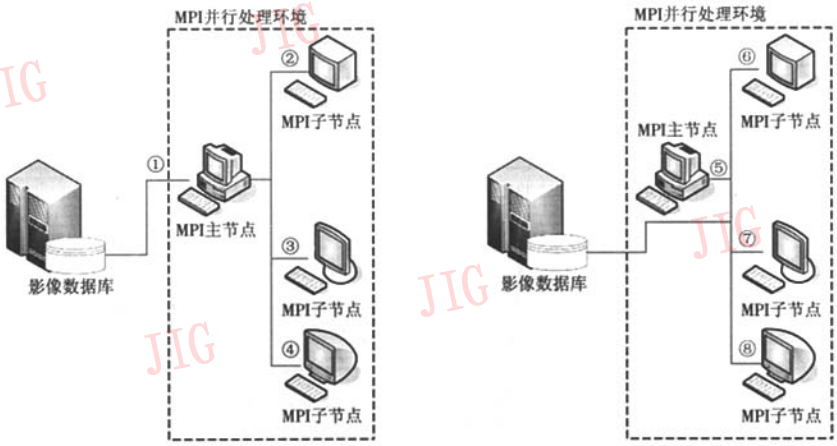
从表 1 中可以看出,由于有一部分数据在处理时无需考虑(本例中为水域部分),所以基于尺度的数据划分方法比常规的平均数据划分方法效率要高一些。考虑到大尺度影像分类的时间为5 119ms,基于尺度的数据划分策略从总体效率上也优于传统的数据平均划分策略,而且总体的效率会随着遥感影像中不需处理的区域比例及算法复杂度的增加而提高。

3.2　基于影像库的并行数据分配与传输

在对以文件形式存储的影像数据进行基于影像库的并行处理时,MPI 程序主进程首先将从影像文件数据块读取数据至其内存中,然后主进程中按照某种方法进行数据划分,并向各计算机子进程进行数据分发。这种方法是目前采用 MPI 实现并行计算的主要方式,其主要优点是流程简单,其底层实现由 MPI 环境负责实现,这样就可以调用 MPI 的相应函数进行数据分发。

当影像数据存储在影像数据库中时,如果仍然采用以上数据获取及传输的方法,即主进程从影像数据库中获取整个影像数据,再采用某种策略分发至其他子节点,那么这一过程存在一个缺点,就是整个处理过程中存在数据的冗余网络传输(见图 2(a)),因为将数据从影像数据库中传输至主进程,再由主进程传输到不同子进程的过程存在着部分数据的冗余传输,特别是当数据量较大时,可以考虑消除此冗余传输,以取得更好的效率。

基于影像数据库的 MPI 并行数据分发策略如图 2 所示,图 2(a)为常规数据处理与分发策略,即图像数据库中的相应图像数据通过 MPI 并行处理环境中的主节点进行数据分发的过程,此数据分发过程的功能由 MPI 的相应函数提供;图 2(b)为本次研究中采用的基于影像数据库的数据分发策略,即图像数据库中的相应图像数据通过影像数据库中的相应服务直接分发至各 MPI 处理节点,这些数据分发过程的功能由影像数据库所在位置的相应服务提供,需由自己实现。



(a) 常规处理数据分发策略　　(b) 本文基于影像数据库的数据分发策略

图 2　基于影像数据库的 MPI 并行处理数据分发策略

Fig. 2　Data distributing method of MPI based on image database

从网络数据流量上看,假设所需要处理的数据量大小为 αMB,则图 2(a)由影像数据库至 MPI 主节点的数据流量也为 αMB,而再由 MPI 主节点根据某种数据分配规则向其他各子节点进行数据分发的过程的数据流量假设为 βMB(如果为等量数据划分原则 $\beta=3\alpha/4$),如果不计算图像处理结果返回主节点的数据流量的话,则总共的网络数据流量为 $(\alpha+\beta)$M。而对于图 2(b)中的数据直接由影像数据库分发的情况,则数据的总流量始终为 αMB,即从影像数据库将总的数据量分成 N 份,并对应地分配到相应的计算机节点处,相对于图 2(a)的情况,图 2(b)的数据减小了 βMB 的流量,由于从总体上降低了网络数据流量,从而提高了系统效率。

从数据传输所需要的时间上看,图 2(a)数据分发策略所需要的时间为 $T_1+\max(T_2,T_3,T_4)$,即必须数据传输至主进程节点后才能继续后面的数据传

输，而图 2(b)数据分发策略所需要的时间仅为 $\max(T_5,T_6,T_7,T_8)$，而实际上各子进程所耗用的时间符合 $T_i \approx T_j,(i,j\in\{2,3,4,5,6,7,8\},i\neq j)$，即总的时间上，图 2(b)也较图 2(a)减少了。

在消除数据传输冗余的方法中，本文采用了基于数据库数据分发策略的实现方法，这种方法消除数据冗余传输的主要方法为：主进程向影像数据库发出数据及数据划分策略的请求(指示各子进程计算机的数据块大小及位置)，影像数据库执行数据分块，同时将各数据块传输至各 MPI 计算节点，并将其以临时文件或内存交换的形式放到子进程计算机中，然后返回主进程相应的数据分配完毕的消息。当主进程收到此消息后，再由主进程激活各子节点进程计算机的相应 MPI 子进程，并由主进程对各子进程的数据处理结果进行收集，并返回数据的并行处理结果。

这种基于影像数据库的直接进行数据分发的策略能够有效地减少传统并行计算处理中的数据冗余传输，进而能够更好地提高并行处理的效率。系统测试的结果是由 MPI 主进程负责数据分配过程所需要的时间大约为 2212 + 533 = 2745ms，其中 2212ms为影像数据库内的数据传输至主进程所需要的时间，533ms 为不同节点所需要的最大传输时间(如表 1 所示)，而由数据库直接进行数据传输所需要的时间为2388ms，由此可以看出，通过这种方式可以提高数据的传输效率，特别是在进行较大数据量传输时。

4 结 论

本文根据遥感影像数据处理的特点，采用了基于特征基元的影像信息提取方法，并给出了基于尺度的影像分割方法及其实现原理，同时结合高分辨率遥感影像数据量、计算量大的特点，本文采用并行的计算方式实现，从而提高了算法的运行速度。

在并行计算过程中，本文还对基于特征基元的尺度分割方法进行了研究，提出了基于特定条件下数据不均匀分配的原理及其数学表达模式，这种方法能够有效地避免算法中由于不必要的全局性像元数据搜索而带来的效率低下的弱点，通过去除背景处理区域的方法来提高算法的处理速度，实验结果也证明了这种数据划分方法能够在一定条件下提高遥感影像处理的效率，也就达到了遥感影像快速处理的目的。在基于影像数据库进行的数据并行处理过程中，本文提出的新的基于影像数据库的直接进行影像数据分配与传输的方法，也能够提高影像处理过程中的效率。

参考文献(References)

1 Yu Li, Wang Run-sheng. Object detection and recognition based on multiscale deformable template[J]. Journal of Computer Research and Development, 2002, **39**(10): 1325 ~ 1330. [余莉，王润生. 基于多尺度变形模板的目标检测与识别[J]. 计算机研究与发展, 2002, **39**(10): 1325 ~ 1330.]

2 Zheng Jiang. Research on Parallel Remote Sensing Image Information Extraction and Analysis Method[D]. Beijing Doctoral Dissertation of Institute of Geographical Sciences and National Resources Research, CAS, 2004. [郑江. 并行遥感影像信息提取与分析方法研究[D]. 北京：中国科学院地理科学与资源研究所博士论文, 2004.]

3 Malay K P, Sanghamitra B, Ujjwal M. A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification[J]. Fuzzy Sets and Systems, 2005, **155**(2): 191 ~ 214.

4 Shen Zhan-feng. Distributed Remotely Sensed Image Computing Model Study and Its Applications on High Resolution Remote Sensing Target Recognition [D]. Beijing: Doctoral Dissertation of Institute of Geographical Sciences and National Resources Research, CAS, 2005. [沈占锋. 分布式影像计算模型及其在高分辨率遥感目标识别中的应用研究[D]. 北京：中国科学院地理科学与资源研究所博士论文, 2005.]

5 Huang Hui-ping, Wu Bing-fang, Li Miao-miao, *et al.* Detecting urban vegetation efficiently with high resolution remote sensing data [J]. Journal of Remote Sensing, 2004, **8**(1): 68 ~ 74. [黄慧萍，吴炳方，李苗苗等. 高分辨率影像城市绿地快速提取技术与应用[J]. 遥感学报, 2004, **8**(1): 68 ~ 74.]

6 Blaschke T, Strobl J. What's wrong with pixels? Some recent developments interfacing remote sensing and GIS[J]. GIS-Zeitschrifi fur Geoinformations Systeme, 2001, (6): 12 ~ 17.

7 Blaschke T, Lang S, Lorup E, *et al.* Object-oriented image processing in an integrated gis/remote sensing environment and perspectives for environmental applications [J]. Environmental Information for Planning, 2000, **2**: 555 ~ 570.

8 Hawick K A, Coddington P D, James H A. Distributed frameworks and parallel algorithms for processing large-scale geographic data[J]. Parallel Computing, 2003, **29**(10): 1297 ~ 1333.

9 Huang Guo-man, Guo Jian-feng. Data partition for distributed-parallel processing of remote sensing imagery[J]. Remote Sensing Information, 2001, (2): 9 ~ 12 黄国满，郭建峰. 分布式并行遥感影像处理中的数据划分[J]. 遥感信息, 2001, (2): 9 ~ 12.

10 Chen Zuo-ning. From high performance computing to high productivity computing[J]. Computer Education, 2004, (6): 26 ~ 28. [陈左宁. 从高性能计算走向高效能计算[J]. 计算机教育, 2004, (6): 26 ~ 28.]