论著

文章编号:1000-5404(2007)09-0840-03

分布式并行计算对数字化人体图像重采样时间的影响

冉 旭[1],宋文强[1],谭立文[2] (第三军医大学基础医学部:[1]计算机教研室,[2]人体解剖学教研室,重庆市计算医学研究所,重庆 400038)

提 要:目的 探讨不同客户机规模的分布式并行计算对数字化人体图像重采样时间的影响。**方法** 将计算机分为1个串行计算组与5个并行计算组,对同一个数字化人体数据集重采样生成冠状面与矢状面图像,比较各组重采样时客户机与服务器的处理时间,以及并行计算加速比。**结果** ①并行计算组与串行计算组比较,客户机重采样时间均显著减少($P<0.01$)。②16台客户机组与8台客户机组、24台客户机组与16台客户机组比较,客户机重采样时间均显著减少($P<0.01$),但32台客户机组与24台客户机组、40台客户机组与32台客户机组比较,客户机重采样时间无显著差异($P>0.05$)。**结论** 分布式并行计算能够大幅度减少数字化人体数据集重采样生成冠状面与矢状面图像的时间,但串行处理和通信开销制约了进一步并行化的规模。

关键词:并行计算;可视化人体;重采样

中图法分类号:R319;R322 **文献标识码**:A

Effect of distributed parallel-computing on the time of resampling of digitized human

RAN Xu[1], SONG Wen-qiang[1], TAN Li-wen[2] ([1]Department of Computer, [2]Department of Anatomy, College of Medicine, Third Military Medical University, Chongqing 400038, China)

Abstract: Objective To explore the effect of distributed parallel-computing on the time of resampling of digitized human with different amounts of client. Methods Computers were randomly divided into one group of serial-computing and five groups of parallel-computing, and the same data of digitized human was resampled by each group. The resampling time of clients and server in each group, speedup of parallel-computing were compared. Results Compared with the group of serial-computing, the time of resampling of the groups of parallel-computing decreased greatly ($P<0.01$). Compared with the group of 8 clients, the time of resampling of the group of 16 clients decreased greatly ($P<0.01$). Compared with the group of 16 clients, the time of resampling of the group of 24 clients decreased greatly ($P<0.01$). Conclusion With distributed parallel-computing for resampling transverse sectional anatomical data from digitized human, the resampling time decreases greatly, but serial processing and communication restrict further parallel processing.

Key words: parallel-computing; visible human; resampling

数字化人体研究是计算机图形图像技术与临床解剖学相结合,通过研究人体信息多模式集成,从而实现人体结构和功能的数字化、可视化,最终达到对人体功能的精确模拟,随着20世纪90年代美国第1例可视化人体的诞生,我们从此进入了数字化人体医学的新领域[1]。数字化人体结构数据集是连续的横断面图像数据,由于医学研究对于图像精度的要求较高,数字化人体数据集的数据量十分庞大[2](100 GB以上)。迄今为止,所有的商业软件都只能在单机上实现二维图像的重采样,在不降低图像精度的前提下需要较长的采样时间(100 s以上),难以满足实时处理与显示的需求,严重阻碍了数字化人体数据集在教学与临床上的应用。因此,本研究将分布式并行计算思想应用于数字化人体重采样,探讨不同客户机规模的分布式并行计算对数字化人体图像重采样时间的影响,为数字化人体数据集的海量数据快速处理打下基础。

1 材料与方法

1.1 断层图像数据的获取

选取首例中国女性数字化人体图像数据集的头部连续横断面图像数据共861张(即1 000~1 860层),连续横断面层厚为0.25 mm,其分辨率为3 072×2 048像素,像素间距为0.167 mm,每张横断面图像大小为36 MB(TIFF格式)[3,4]。为

作者简介:冉 旭(1977-),男,重庆市人,硕士研究生,讲师,主要从事医学图像处理方面的研究。电话:(023)68752219, E-mail:hc_ranxu@tom.com

通讯作者:宋文强,电话:(023)68752220

收稿日期:2006-12-05;修回日期:2006-12-30

了更好地表现重采样过程的实时性，首先对原始数据集图像进行了去噪、背景去除、格式转换等前期处理[5]，处理后的图像分辨率和像素间距保持不变，每张横断面图像大小为 53 ~ 2 393 kB 不等（无损压缩的 PNG 格式），总数据量为 1.14 GB。

1.2　方法

1.2.1　分布式并行计算环境　　使用配置为 Pentium 4 2.93 GHz，512 MB 内存，80 GB 硬盘的 40 台计算机，金浪 100 MbPS 交换机搭建一个高速计算机网络运行平台，平台上应用自己在 Microsoft Visual Studio. NET 开发平台下，以 C#语言进行 Socket 编程开发的 CVH-SC 软件对各计算机进行任务分配与调度。

1.2.2　CVH-SC 软件的实现方法

1.2.2.1　客户机连接服务器　　各客户机启动客户端程序侦听 UDP8000 端口，等待加入服务器。服务器启动服务器端程序通过 UDP8000 端口在子网内进行广播，请求客户机加入，然后侦听 TCP8000 端口。各客户机接收到服务器广播后分析广播报文数据，如果是加入命令，便从报文中取出服务器的 IP 地址，启动 TCP 连接进程，连接服务器 TCP8000 端口。

1.2.2.2　服务器分配任务　　等待客户机加入完毕，服务器端程序根据加入的客户机数量进行任务的动态分配，目前本实验采用的是根据图片文件数量平均分配任务的方式。服务器端程序将重采样命令、重采样位置坐标、重采样方式、各客户机 IP 地址及其对应的任务分配组合成一个广播报文，通过 UDP8000 端口在子网内进行广播。

1.2.2.3　客户机重采样　　各客户机接收到广播后分析报文数据，如果是重采样命令，则取出重采样位置坐标和方式、自己 IP 地址对应的任务分配等信息开始重采样工作。客户机重采样完成自己所分配的部分图像后启动发送进程，向服务器 TCP8000 端口发送处理完成后的图像[6]。

1.2.2.4　服务器合成图像　　服务器侦听到各客户机的数据发送请求后，启动多个接收进程接收客户机发送的数据存储到指定的目录，同时服务器启动一个检测进程检测是否所有的客户机都完成数据传送的工作。待所有客户机数据上传完毕，服务器启动图像合并进程，将客户机上传的图像排序后进行拼接工作。

1.2.2.5　服务器修正图像　　拼接后的图像即为重采样完成的冠状面或矢状面图像，但由于原数据集在 X、Y 轴上的像素点距为 0.167 mm，而 Z 轴上切片的层厚为 0.25 mm，故拼接后的图像发生了一定的形变，因此还应在 Y 轴方向进行插值处理，插值后新图像的高度 = 拼接后图像的高度 ×（0.25 ÷ 0.167）。

1.2.3　重采样实验　　应用 CVH-SC 软件，分别选取不同的客户机数量（1、8、16、24、32、40 台）进行分组，以 1 台客户机分组作为串行计算组，其他各分组作为并行计算组，分别进行数字化人体图像重采样生成冠状面和矢状面图像，每次重采样在相同的计算环境下重复 10 次，共 120 次实验，记录每次重采样客户机与服务器分别所用时间。

1.2.4　重采样加速比的比较　　一个并行算法的主要优势在于它相对于串行算法所能取得的增益，加速比 S_p 是反映这方面性能的最常见评价指标。

$$S_p = T_1 / T_p$$

式中，T_1 为单机上串行算法的处理时间，T_p 为使用 n 台处理机运行并行算法所用的时间。

实验中，以 1 台客户机、1 台服务器完成重采样所需总时间的均值 $\bar{T}_1$ 作为串行算法的处理时间，n 台客户机、1 台服务器在 CVH-SC 软件的调度下完成重采样所需总时间的均值 $\bar{T}_p$ 作为并行算法的处理时间进行加速比 S_p 的运算与比较。

1.3　统计学处理

结果以 $\bar{x} \pm s$ 表示，用 SPSS 10.0 统计软件进行 t 检验及单因素方差分析。

2　结果

2.1　重采样生成的冠状面与矢状面图像

应用 CVH-SC 软件成功重采样生成头部任意选取位置的冠状面、矢状面图像，图像可以放大缩小进行观察也可以按照选定的大小进行裁切（图 1、2）。

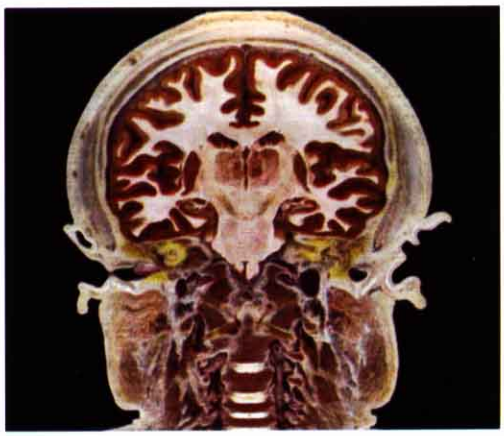


图 1　重采样生成的冠状面图像

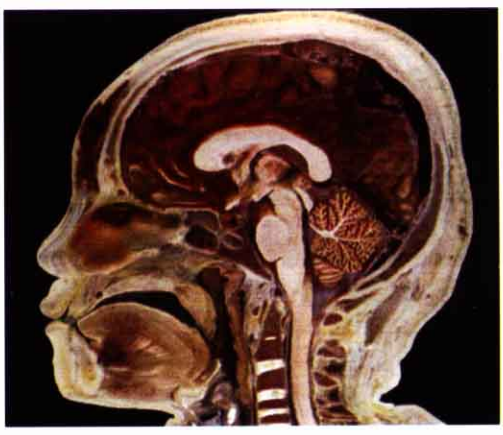


图 2　重采样生成的矢状面图像

2.2　重采样时间的变化

与串行计算组比较，并行计算组客户机重采样时间均显著减少（$P < 0.01$）。随着参与重采样客户机数量的增加，客户机重采样所需时间持续下降，在 24、32、40 台客户机组，重采样时间达到并保持在较低水平（10 s 量级）。但 32 台客户机组与 24 台客户机组，40 台客户机组与 32 台客户机组比较，重采样时间并无显著差异（$P > 0.05$）。服务器重采样时间各组相互比较均无显著差异（$P > 0.05$）（表 1、2）。

表1 不同客户机数量对重采样生成冠状面图像所需时间的影响(s, $\bar{x} \pm s$)

重采样时间	客户机数量(台)					
	1	8	16	24	32	40
客户机重采样时间	283.31 ±5.09	38.59 ±2.17[a]	19.62 ±1.89[ab]	13.19 ±1.53[ac]	11.35 ±2.05[a]	9.05 ±1.67[a]
服务器重采样时间	8.97 ±0.52	8.71 ±0.59	9.28 ±0.63	9.07 ±0.45	9.38 ±0.71	9.31 ±0.68

a: $P<0.01$, 与1台客户机组比较; b: $P<0.01$, 与8台客户机组比较; c: $P<0.01$, 与16台客户机组比较

表2 不同客户机数量对重采样生成矢状面图像所需时间的影响(s, $\bar{x} \pm s$)

重采样时间	客户机数量(台)					
	1	8	16	24	32	40
客户机重采样时间	273.58 ±4.86	35.84 ±1.56[a]	17.75 ±1.98[ab]	12.53 ±2.01[ac]	10.27 ±1.89[a]	9.23 ±1.57[a]
服务器重采样时间	5.26 ±0.34	5.67 ±0.57	5.38 ±0.48	5.16 ±0.47	5.49 ±0.41	5.31 ±0.65

a: $P<0.01$, 与1台客户机组比较; b: $P<0.01$, 与8台客户机组比较; c: $P<0.01$, 与16台客户机组比较

2.3 重采样加速比的变化

随着参与重采样的客户机数量的增加,加速比大体上呈现线性增长的趋势,但是加速比增长的趋势逐渐减弱并最终趋于稳定。在整个重采样实验中,冠状面重采样的加速比增幅始终小于矢状面重采样,并有逐渐扩大的趋势。

3 讨论

分布式并行计算是将一个计算任务分解成若干片段,分配到网络中的多台计算机同时计算,从而提高问题求解速度,或者求解单机无法求解的大规模问题[7]。数字化人体数据集的数据量庞大,在个人计算机上处理需要大量的计算机时间,但对于重采样生成冠状面与矢状面图像而言,局部数据之间的相关性较小,较少地涉及学科知识和人工干预,完全满足分布式并行计算的特征。本实验结果显示利用校园网内连接的大量个人计算机,只要进行合理的任务分配与调度,完全能够以较低的成本搭建高性能的海量信息计算平台,完成数字化人体数据集数据的高速处理,大幅度减少重采样的时间,提高重采样加速比,实现图像数据的实时处理与显示。

本研究对串行计算组与并行计算组以及不同客户机规模的并行计算组之间进行了处理时间、加速比的比较,结果显示随着参与重采样客户机数量的增加,数字化人体图像重采样所需时间大幅度减少,但当客户机数量达到一定规模时,总的处理时间达到并保持在较低水平,加速比也趋于平稳,这主要是算法中服务器上串行处理部分和处理机之间通信开销的瓶颈逐渐凸现,加速比是无法单纯通过增加处理机的数目来提高的,这也正符合了固定规模的加速比模型[8]。

实验结果显示,在处理数据量不变的情况下,服务器上用于合成图像的串行处理时间各组之间无显著差异,并且与客户机的数量无明显关系($P>0.05$)。由于重采样生成的冠状面图像分辨率为3 072×681像素,而重采样生成的矢状面图像分辨率为2 048×681像素,因此冠状面重采样后图像在服务器上需要更多的合成处理时间,故在冠状面重采样中串行处理部分对加速比的影响更加明显,因此冠状面重采样加速比增幅始终小于矢状面重采样。

综上所述,分布式并行计算能够显著减少数字化人体图像重采样生成冠状面与矢状面图像的时间,提高重采样加速比,但随着并行规模的扩大,串行处理部分和通信开销的瓶颈逐渐凸现,制约了进一步并行化的程度。如何采用更好的策略进一步优化算法,解决瓶颈问题,以更好的解决数字化人体数据集海量数据的处理是我们需要进一步研究的内容。

参考文献:

[1] Spitzer V M. The visible human: a new language for communication in health care education[J]. Caduceus, 1997, 13(2): 42-48.

[2] 张绍祥,刘正津. 数字化可视人体研究[J]. 第三军医大学学报, 2003, 25(7): 561-562.

[3] 张绍祥,刘正津,谭立文,等. 首例中国女性数字化可视人体数据集采集与可视化研究[J]. 第三军医大学学报, 2003, 25(5): 394-396.

[4] 李 恺,张绍祥,谭立文,等. 游离肝脏可视化数据获取技术研究[J]. 第三军医大学学报, 2005, 27(23): 2388-2390.

[5] 谭立文,王忠贵,张绍祥,等. 数字化可视人体连续断层标本的参数测量[J]. 解剖学杂志, 2005, 28(2): 195-197.

[6] 李 军,李德仁. 分布式遥感图像处理中的若干关键技术[J]. 武汉测绘科技大学学报, 1999, 24(1): 15-19.

[7] McHugh J, Abiteboul S, Goldman R, *et al*. Lore: A database management system for semistructured data[J]. SIGMOD Record, 1997, 26(3): 54-66.

[8] 卢丽君,廖明生,张 路. 分布式并行计算技术在遥感数据处理中的应用[J]. 测绘信息与工程, 2005, 30(3): 1-3.

(编辑 栾 嘉)