一种BPNNs识别算法的医学检测泛实时性问题研究

刘玉成[1] 理查德·丁[2] 张颖超[3]
(南京财经大学国家级重点实验中心 南京 210003)[1] (美国波士顿克罗诺斯研究所 波士顿 02101-02117)[2]
(南京信息工程大学信息与控制学院 南京 210044)[3]

摘 要 尿沉渣空间环境的复杂性,导致采集的有形成分图像存在较多冗余信息,提取有效的图像信息变得较为困难,进而使得识别系统需要处理的数据量十分巨大。虽然BP神经网络算法的串行版本DJ8000系统平台解决了细胞等有形成分的识别准确率问题,但其不能满足尿沉渣图像医学检验的实时性要求。为此,提出了基于BP神经网络算法优化的并行处理GPU框架的系统平台。它采用并行优化框架,同步高效地对数据进行加速处理;同时,以GPU计算和测试平台为硬件系统支持,无论是在硬件指标、数据传输及总线技术还是软硬件的兼容性方面,都有助于解决算法中时常出现的负载不均衡的问题。实验数据表明,BP神经网络尿沉渣识别算法在优化并行框架的GPU系统处理平台上显示的加速比、时效比和运行时间等相关性能参数值都有所提升。相比于DJ8000系统平台,优化的AMD HD7970和NVIDIAGTX680两个并行处理GPU框架系统平台相应的加速比参数值分别是前者的10.82～21.35个和7.63～15.28个标准当量。实验数据充分说明,优化并行框架的GPU处理系统中相关的逻辑数据、地址数据和线性寻程的函数映射关系均能相互动态分配对接并优化算法架构,实现软件到硬件系统的最优比映射,最终解决由于线程间负载不均衡导致的性能瓶颈问题,从而有效地化解了医学领域实时检测中的时效性这一难题。

关键词 BP神经网络,GPU平台,负载不均衡,并行优化,线程协调
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.06.053

Research on Pan-real-time Problem of Medical Detection Based on BPNNs Recognition Algorithm

LIU Yu-cheng[1] Richard·DING[2] ZHANG Ying-chao[3]
(State Key Laboratory, Nanjing University of Finance and Economics, Nanjing 210003, China)[1]
(U.S.A. Kronos Research Institute of Boston, Boston 02101-02117, USA)[2]
(School of Information and Control, Nanjing University of Information Science and Technology, Nanjing 210044, China)[3]

Abstract Due to the complexity of the urine sediment space environment, there is much redundant information of the collected tangible component image, and it also becomes difficult to extract effective image information. Therefore, the amount of data that need to be deal with is huge. Although the serial version DJ8000 system platform of BP neural network algorithm solves the problem of recognition accuracy of tangible components such as cells, it can't meet the real-time requirement of urine sediment image medical examination. To solve this problem, this paper presented a system platform of parallel processing GPU framework based on BP neural network algorithm optimization. It uses parallel optimization framework to synchronize and accelerate processing of data efficiently. At the same time, it supports the hardware platform based on GPU computing and test platform. Whether from the hardware indicators, data transmission and bus technology or hardware and software compatibility, it will help solve the problems, which often occur in the uneven load irregularities. Experimental data show that BP neural network algorithm for urinary sediment identification improve the performance parameters such as speedup, aging ratio and running time on GPU platform processing platform. Compared with DJ8000 system platform, the parallel processing GPU framework system platforms of AMD HD7970 and NVIDAGTX680 are optimized, and their corresponding acceleration ratio parameter values are 10.82～21.35 and 7.63～15.28 standard equivalents respectively. The data show that optimizing the mapping relationship between logical data, address data and linear seeking function in the GPU processing system of parallel frame can dynamically allocate and optimize the algorithm structure and optimize the mapping between software and hardware system. Finally, it solves the

到稿日期:2017-01-12 返修日期:2017-04-16 本文受江苏省六大人才高峰(2106-A-027),江苏省高校自然科学基金项目(12016KJD520122)资助。

刘玉成(1980－),男,硕士,讲师,主要研究领域为系统分析与集成、计算机图像学、智能模式识别等,E-mail:lyc0871@163.com(通信作者);**理查德·丁**(1970－),男,博士,高级研究员,主要研究领域为数据库技术、情报技术等;**张颖超**(1960－),男,教授,博士生导师,主要研究领域为系统控制和仿真、网络控制技术等,E-mail:qpl@nuc.edu.cn(通信作者)。

problem of performance bottlenecks caused by load imbalance between threads. Thus, it effectively resolves the problem of real-time detection in urinary sediment environment.

Keywords BP neural networks, GPU platform, Load imbalance, Parallel optimization, Thread coordination

模式识别技术应用广泛，而 BP 神经网络算法[1]作为识别领域的重要方法，其应用已十分成熟。神经网络在识别尿沉渣中有形成分方面的应用也取得了一定的成果，在特征提取、有监督和无监督学习、训练、检测识别等环节也都比较稳定。由于尿沉渣环境提取有效图像的空间环境较为复杂，且图像处理技术及要求不断提高，导致 DJ8000 系统所提取的有形成分的信息图像所含信息规模、数据量巨大。为了确保识别的准确性，所采取的样本图片的分辨率大幅提高，这使得输入计算和测试平台的每个基本单位图片的数据量很大。标准的 CPU 或者 DJ8000 型的串型处理方法由于计算速率等专业性处理性能的限制，不能满足医学检测和分类的实时性要求，从而限制了它们在医学领域的应用和推广。因此，文中创新性地提出了 BP 神经网络尿沉渣图像算法集成 GPU 平台的并行优化技术。

神经网络(Neural Networks)的识别是一个科学而又严谨的方法和过程，其复杂性和精确性不言而喻。识别(Recognition)的方法[2-3]包括很多具体过程，但其主要部分由训练(Rraining)与学习构成，具体过程完成之后，可形成与之相对应的识别函数(Identification Function)及其关系，从而有利于成分识别。神经网络可被视为一个特殊的函数映射(Mapping)，适用于有输入输出的一一对应关系，而细胞图像识别将细胞成分的主分量(PCA)作为其输入值，图像输出的复杂函数映射问题表示为类别属性。因此，利用细胞识别器进行的智能识别(Intelligent Recognition)也能使用 BP 神经网络来完成。BP 类型的算法是当前应用于神经网络学习的有效算法之一，特别是基于多层次前馈型网络(Multiple Layer Feed Forward Network)的系统，即 MLFFN 网络，其算法的精髓也是误反型(即 EBP 算法)。该算法的优点是可以任意调控其精度以逼近任意的连续型函数，因此可能对网络中的各层权系数进行修正。

对于文中所架构的 GPU 计算平台 AMD HD7970 和 NVIDIAGTX680 系统，采用了并行优化设计。该提升主要针对标准 GPU 的静态对应模式进行类似于静态路由寻找逻辑地址发包的工作模式及其体现出的线性数据处理负载不均衡的弊端。所有处理模块的逻辑数据、地址数据和线性寻程的函数映射关系都在 GPU 的 kernel 运行前被程式化处理，这使得多个相对独立的处理单元(包括 CU-Compute Unit 的计算单元和 PE-Processing Unit 的处理部件)不能动态协助工作，无法实现动态调整而达到相对满负荷的运转。

而 BP 神经网络识别算法的并行优化框架能突破加速比等关键性指标的性能瓶颈，主要得益于其良好的并行性优化方法、改进的 BP 神经网络识别算法与高性能的 GPU 数据测试计算平台的无缝集成，它们使得处理系统中相关的逻辑数据、地址数据和线性寻程的函数映射关系能相互动态分配对接，优化算法架构，从而实现软件到硬件系统的最优比映射，解决由于线程间负载不均衡导致的性能瓶颈问题。

BP 神经网络细胞识别算法在 GPU 处理平台上表现出的高效性主要体现在识别的准确性、处理数据的加速比和速率等方面。同步、高效地对数据进行加速处理以及识别率和时效性对应的各项指标性能的提高，均可归因于 GPU 计算处理平台硬件的高性能和负载不均衡特性算法的性能优化架构的 GPU 优化并行处理。具体采用的并行优化方法包括 UBER-Kernel 、Persistent Thread、粗粒度并行、本地队列、动态映射和全局等 6 个基本功能性单元在内的若干个方法组合。根据不同特征选择不同的优化 GPU 并行组合，可显著增强 GPU 的计算能力和可编程等性能，有利于在更广泛的领域内应用。

1 实验背景和工作准备

本研究是在原有尿沉渣有形成分识别课题[4]前期研究工作的基础通过上，通过改进型 BP 神经网络识别算法和 DJ8000 采集系统检测和分类的集成方法来处理尿沉渣有形成分的进一步研究和拓展。针对医学检测对实时性和准确性要求的不断提高，本研究做了一些有效的推进工作，提出了 BP 神经网络识别算法的 GPU 并行优化处理框架。该框架中拥有的若干个基本功能性单元：粗粒度并行处理、UBERKERNEL、PERSISTENT THREAD 等，可以优化配置成多种组合方法，能解决信息处理过程中的线程负载不均衡而导致的时效性差的问题。从实验数据来看，系统获得了很好的性能，可以充分说明本研究针对尿沉渣有形成分特征处理提出的 BP 神经网络算法的 GPU 并行优化框架方法相对于前期研究工作中所采用的传统尿沉渣识别算法更具先进性和时效性。

2 并行算法 GPU 集成的可行性研究

大量的前期研究工作和实验数据表明，要解决 GPU 并行架构所暴露出来的不均衡分配问题(包括进程中线程间负载不均衡非规则特性现象在内的问题)及其产生的时效性等问题，系统平台必须具备以下几个基本条件：1)与算法相匹配的硬件要求，包括 GPU 计算和测试平台对系统所提出的处理器要求，以及文中应用的 DJ8000 数据图像库与 GPU 对接的现场总线技术，最后还包括连接源图像库与 GPU 存储设备的传输通道的带宽[5]，使其满足多通道数据信息计算和处理下的高速率传输；2)与 BP 神经网络识别算法的特点相适应的并行优化框架及处理方法。Sharma 等人[6]把 BP 神经网络识别算法融合到 GPU 平台，处理负载不均衡现象时采用了所谓的金字塔模式，把预处理后的图片进行规格上的标准化、统一化，使得在线程间处理的目标图像的数量相对均衡，且其基于系统的每个线程处理图像数据的信息量也会保持相对平衡，从而解决了图像负载不均衡的矛盾。实验表明，这一金字塔模式的处理思想和方法在解决问题时的效果有限，因此并没有得到广泛应用，自然也没有在图像识别领域得到推广。Hefenbrock 等人[7]提出的并行性优化 GPU 框架的思想

具备与 BP 尿沉渣神经网络算法相结合的前提条件。该方法的核心思想是将特征参数计算、学习分类模块、标准统一化图像等三级并行优化模式进行集成。从大的理论框架上来说，三级并行优化方案符合识别算法的大数据信息量计算的实时性要求，但具体分析框架中的模块细节时就不难发现，它针对数据信息量较大的学习分类模块，只是以机械性地增加处理线程数目的方式来提高本级处理数据的能力，从而使得系统的整体处理速度和性能与 FPGA 相当。但是实验证明，这种机械式地增加线程数量的方法在实际应用领域有很大的局限性，最主要的问题是它无法满足数据计算量大的要求，兼容性也存在一定的问题。

虽然上述各种不同方向的研究工作和成果能解决加速性能、兼容性能和时效性方面[8-9]的一些问题，但这种简单机械式的方法只能提升局部性能，并没有从整体上解决根本问题；任务处理过程中的线程之间缺乏相应的动态协调机制，且在与 BP 神经网络识别算法融合的有效性方面也很难令人满意。即便如此，这些思路和方案还是为实现 BP 神经网络识别算法与 GPU 硬件平台两者的完美融合提供了宝贵的实验参数和经验，为后期的研究工作指明了正确的方向，具有很好的指导作用。

依据前期针对 BP 神经网络识别算法的并行性优化 GPU 处理框架集成方法[10-11]方向的相关研究和所积累的丰富经验，以及处理两者融合时在兼容性方面所做的大量工作，结合文中 BP 神经网络识别算法的特点，本文提出了一种优化的并行处理 GPU 框架的系统平台，其在硬件性能指标、数据传输、总线技术及软、硬件的兼容性等方面都可以满足当下医学处理的实时性要求。计算信息处理的相应速度以及各种系统性能参数所体现出来的数值指标都体现出了系统平台整体的高性能，有效地解决了优化前并行 GPU 架构带来的负载不均衡问题，在满足了医学检测准确性的前提下也能满足时效性要求。

2.1　GPU 系统平台的适用性分析

差异化高性能的主流处理芯片是 GPU 集成化水平提高的核心条件之一。如此，GPU 系统平台就能够在处理擅长的专用数据的同时，兼具处理智能识别等通用复杂计算和数据依赖任务的能力。整体上看，GPU 架构具有良好的统一性[12-13]，即其具有大规模细粒度并行处理器，又具有层次式的架构特点。如图 1 所示。

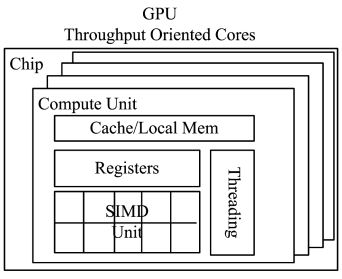


图 1　GPU 的整体架构

Fig. 1　Overall architecture of GPU

从局部细节上分析，它采用了数量众多的计算单元和超长的流水线，且内部结构布局合理，因此性能更加优越。

另外，在 GPU 架构的吞吐量设计中，它拥有很多的 ALU 和很少的 Cache。当 Cache 获取数据时，它会及时转发给对应的线程（有关线程间动态协调的运行机制将在并行可行性实现章节中详细论述），如图 2 所示。

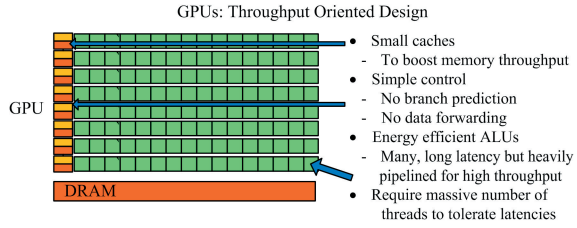


图 2　GPU 功能分布的设计

Fig. 2　Design of functional distribution of GPU

据此，本研究运用 GPU 并行优化框架的层次式处理器具有以下优点：1）在编程方式方面，GPU 采用的是 HOST＋DEVICE的统一模式，并使用具有 Microcontrollers 功能的核心中央处理器部件来统一分配调度计算任务和线程；同时，应用了 G-B-THREAD 的层次式组织方式来动态协调工作。2）在线程调度方面，应用硬件依次固化的前置函数映射对应链接，使得以 WORK GROUP 为基本单位的线程在 K-GPU 程序启动前就已经被对接完成，减少了内存占用，并缩短了进程的处理时间。3）在内存访问方式方面，访问的权限从传统的私有方式转变为共享方式。无论是 WORK GROUP 内部共享的局部内存空间，还是相对私有的访问空间，都统一为相通的权限，实现共存共访，以达到动态化协调、优化访问方式、加快处理速率、提高整体性能的目标。4）在 GPU 的内部结构组成方面，若干个基本处理部件 PE 构成一个 CU，而若干个 CU 又组成了一个独立的 GPU 。

上述特性表明，GPU 系统平台较满足医学检测实时性方面的要求。

2.2　BP 神经网络的针对性分析

2.2.1　BP 网络结构分析

BP 神经网络是基于 BP 算法的多层次前馈型（Feed Forward）神经网络[14]。前沿性 Back Propagation（反传型）网络及其算法思想被广泛应用于模式识别（Pattern Recognition）、数据压缩（Data Compression）、函数逼近（Function Approximation）和分类（Classification）等领域，即对输入的矢量值以预定义的数值进行适当分类。本文应用上述的一系列方法对细胞的各种成分进行学习、训练和识别。

2.2.2　BP 网络算法分析

根据尿沉渣细胞识别算法的特点，文中选用的 BP 型神经网络算法的运算的重要逻辑是自后到前的，以各个单元输出层的误差值（Error）为判断依据，主要方法是逐层地计算出其隐含层（Hidden Layer）的误差值。该 BP 神经网络的算法可分为两个主要阶段：1）正向过程，其各单元的输出值由输入信息值从输入层经过隐含层逐层计算得到；2）各单元的误差值由输出误差逐层向前计算出隐含层而得到，并以各单元的差值来修正前层的权值。

相对应地，反传型（Back-propagation Type）算法的概念和过程如下：目的和作用可视为权值修正，方法一般可采用梯度法（Gradient Method），输出函数的表达形式一般是 sigmoid 函数。于是，可微必定是其输出函数（Output Functions）的基

本要求之一。系统以 BP 神经网络 X 层的第 j 个计算单元(Unit)为例,各个单元中具体环节表示的含义是:网络输出层的第 k 个单元用小写字母 k 来表示,而其输出值用 O 表示;同理,网络输入层的第 i 个单元用小写字母 i 来表示,而在该神经网络中,上一输入层至下一层传递数值的权重(Weights)值[15]用 w_{ij} 来表示,如图 3 所示。

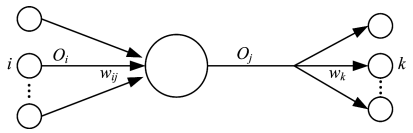


图 3 BP 算法的单元变量约定

Fig. 3 Unit variable convention of BP algorithm

对在 BP 神经网络每一层的各个单元中的元素,均按图 3 中箭头方向的顺序做正向算法(Forward Algorithm)的运算,其数值来源于采集的输入样本(Sample)值。具体运算公式如下:

$$net=\sum_i w_{ij}O_i \tag{1}$$

$$O_j=f(net_j) \tag{2}$$

用 Y_j 代表采集的样本数值计算出的理想化输出值(Idealized Output Value),实际值则用等价公式 $\hat{y}_j=O_j$ 来计算,而前后层级之间产生的误差 E 则利用以下计算公式即通过 Y_j 与其均值的差的平方得出:

$$E=\frac{1}{2}\sum_j(y_j-\hat{y}_j)^2 \tag{3}$$

简化起见,局部梯度(Local Gradient)的概念可用等价式(4)来表示:

$$\delta_j=\frac{\partial E}{\partial net_j} \tag{4}$$

鉴于权值与误差值之间的关系,其计算结果可能导致严重偏差(Deviation)。对其进行修正可得:

$$\frac{\partial E}{\partial w_{ij}}=\frac{\partial E}{\partial net_j}=\frac{\partial net_j}{\partial w_{ij}}=\delta_j O_i \tag{5}$$

权值修正应使误差减小得最快,修正量为:

$$\Delta w_{ij}=-\eta\delta_j O_i \tag{6}$$

$$w_{ij}(t+1)=w_{ij}(t)+\Delta w_{ij} \tag{7}$$

其中,η 为步长。

情况 1 如果节点(层级元素)J 是输出单元,则有计算式(8)、式(9):

$$O_j=\hat{y}_j \tag{8}$$

$$\delta_j=\frac{\partial E}{\partial \hat{y}_j}\frac{\partial \hat{y}_j}{\partial net_j}=-(y_j-\hat{y}_j)f'(net_j) \tag{9}$$

情况 2 若层级元素 J 为非输出单元,则依据图 3 可知,O 会严重干扰下一层级的计算结果。为此,由式(10)计算得出:

$$\delta_j=\frac{\partial E}{\partial net_j}=\sum_k\frac{\partial E}{\partial net_k}\cdot\frac{\partial net_k}{\partial O}\cdot\frac{\partial O_j}{\partial net_j}=\sum_k\delta_k w_{jk}f'(net_j) \tag{10}$$

依据 KOS. M-ROV 定理,在对应的反向传播神经网络中,其要素满足了以下基本要求:1)科学的层次结构;2)合理的权重值;3)自由逼近的可微函数(Continuous Function)。

鉴于此,再结合尿沉渣有形成分的提取特征,文中选取具有三层结构的 BP 神经网络作为尿沉渣[16-18]有形成分的识别分类器最为适合。

本研究以 BP 神经网络识别算法与 GPU 的集成作为新起点,对基于 GPU 并行性优化处理框架的尿沉渣有形成分识别算法特别是实时性方面的相关内容作深入研究和分析。

3 BP 神经网络识别算法的并行优化 GPU 实现

本研究中设计的 BP 神经网络识别算法的实现与 GPU 系统良好的集成受益于两大先天优势:1)文中采用的识别算法具有适合在并行的 GPU 系统平台上运行的特点;2)文中提出的 GPU 并行框架突破了处理可兼容算法的并行性运行所产生的效率性[19-20]问题,即优化前已解决了并行 GPU 架构的负载不均衡问题。

3.1 识别算法的并行性特征集成

在并行实验运行的 BP 神经网络识别算法的集成过程中,识别算法充分利用了本系统设计的 BP 神经网络算法具备的先天优势,其相关模拟性特点和兼容性优势明显满足整体实现的要求。

1)其利用了非常适合并行处理的运行程序结构。这在第 2 节对神经网络算法的介绍和可行性研究论证中已经得到了充分验证和说明。该算法展示出了明显的三级化并行架构,即计算特征参数、学习分类模块、标准统一化图像的三级并行优化模式。每个不同组级的处理任务的共性是它们相互独立,并行执行需要处理的数据,完全符合 GPU 并行处理的 SIMD 架构的特征和优势要点。

2)结合计算内核的优势,针对该程序代码的最终实现,其最重要的任务是读写内存中的数据以及计算数值和逻辑。一方面,其采用核心处理器和内存条的高性能来加快处理速度;另一方面,主要依靠内存来减少读写寄存数据的时钟周期,从而突破了受限于大量计算的性能瓶颈,并解决了所产生(即 Compute-intensive 特征)的数据量处理带来的时效性问题。当然,这需要与上述条件协同完成。

鉴于上述识别算法呈现出的大规模信息量处理计算的良好特性,以及该算法本身具有的兼容性并行计算和处理的优势,BP 神经网络识别算法已被成功地融合并运行于 GPU 并行处理系统平台上,有助于 GPU 突破负载不均衡特性。

3.2 并行 GPU 的优化性特征集成

在 GPU 系统平台部分的集成过程中运用上述另外一个重要的先决条件,使得并行框架完美地与识别算法相集成,主要解决了实现要求的重心问题,即解决了运行中出现的负载不均衡[21-22]难题。文中第 2 节(即并行算法 GPU 集成的可行性研究)中已详细阐述了相关理论和运行机制,本节主要实现 GPU 系统平台的并行性框架和关键性技术优化等方面的集成。

本研究根据 BP 神经网络识别算法满足的并行性处理架构的特征和优势,创造性地将该识别算法在 GPU 平台上进行了 NAÏVE 实现。

由于 GPU 平台并行处理的主要目标是学习分类模块中的数据,因此采用的 BP 神经网络识别算法的 NAÏVE 版本在 GPU 的处理平台上使用了针对性、高效性和简单化的并行策

略。识别算法自身信息数据计算的过程特性和功能特点，使得这一模块的信息处理量十分巨大。为更好地解决数据计算所带来的时效性瓶颈问题，采用优化框架的 GPU 系统来并行处理信息计算量巨大的学习分类模块。运行机理中的策略可被概括为：1）开启与分类窗口呈线性关系的线程；2）每个线程任务相对独立，但动态协调，呈多对多模式；3）降低数据的读取频率，用 L-D-S 方式实现信息的本地共享；4）给采集的特征参数降维，并在 W-G 工作组中尽量采用低维度结构数据。

虽然目前并行集成工作取得了一定的阶段性成果，但它并没有完全解决医学细胞检测中产生的时效性问题，而仅仅是解决了整个问题的并行性系统架构的总体设计和方向性问题。从相关实验中 GPU 运行反映出来的性能指标参数来看，想要最终解决相关的实时性问题，还必须突破 NAÏVE 版本运行中线程之间产生的负载不均衡问题，这一问题其实也是 GPU 系统平台集成其他各种算法时常会出现的问题。其中，后者更是解决医学细胞检测实时性问题的关键。

此外，相关实验中的数据对比也表明，一方面，在相应的级联分类器中，无论是 DJ8000 还是 CPU 的串行算法版本，它们所显示的加速比等相关性能参数值与 NAÏVE 版本在 GPU 平台上显示的性能指标值相比，都证明 GPU 系统平台的性能不但没有得到提高，反而还略低于前者，但除此之外的其他性能指标全都优于前者；另一方面，NAÏVE 版本在 GPU 平台上运行的性能参数没有达到预期的效果。机理中策略所对应的线程任务的动态协调机制还需进一步优化和调整。

相关性能参数不理想的主要原因可被概括为：在并行架构的学习分类模块中，其数据计算量巨大，而 GPU 所采集的有效的尿沉渣有形成分图像在进行数据计算处理时，进程之间是一一对应来负责处理的，并且每个样本图像上有效的有形成分差异很大。一些采集图片的有形成分的数量大、种类繁多，而有些图片中的尿沉渣有形成分几乎又检测不到有效目标，从而使得 BP 神经网络识别算法的 GPU 在进行系统实现时，运行处理的程序中的线程之间的任务量呈现级数化差距，直接导致时钟周期严重不统一，有限的系统资源被严重浪费，线程之间的运程不均匀，最终未能充分发挥 GPU 的并行处理能力，导致 NAÏVE 的 GPU 并行系统架构的优势无法得到体现，GPU 强大的计算处理性能未能发挥作用。

指标参数集中反映出的焦点问题是，由 Thread 线程间负载不均衡带来的一系列系统性能及其相关次生问题。解决这些问题的关键是，结合 BP 神经网络识别算法的特点，优化调整 GPU 系统平台的并行处理架构。整体 GPU 并行优化框架中各模块的功能分布及运行情况如图 4 所示。

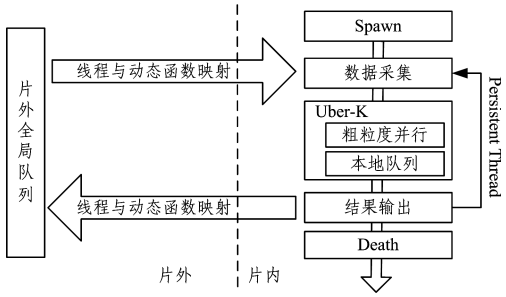


图 4　并行优化框架及模块

Fig. 4　Parallel optimization framework and modules

综合分析上述 GPU 并行优化框架结构，从整体性能和局部性能结合的角度来看，并行模块的局部构成之一的学习分类级模块的巨大数据量的运算所带来的时效性问题，是本研究提出的 GPU 集成 BP 神经网络识别算法的优化并行框架的性能瓶颈。为提高 GPU 并行系统的加速比等相关性能，从整体的角度来考虑，本文系统着重优化和提升了学习分类级模块的运行处理速度。优化后的 GPU 并行框架简单分成了 6 个基本功能性单元，对应的并行优化策略分别表现如下。

1）UBER-Kernel。开启与分类处理窗口呈线性关系的线程统一化，集本单元中若干个基本功能的 kernel 为一体的全局 UBER-K。其具体空间规格统一化的流程如图 5 所示。

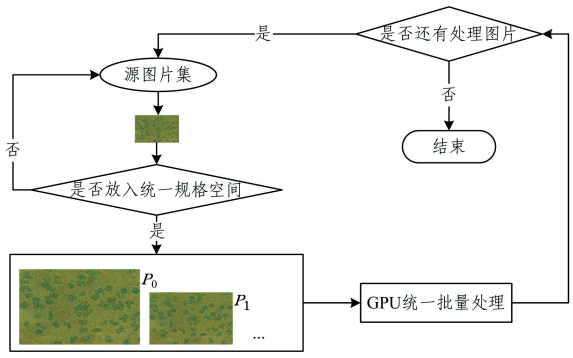


图 5　空间规格统一化流程

Fig. 5　Uniformization of space specification

2）Persistent Thread 和粗粒度并行。两者结合的目的是综合赋予 UBER-K 新的运作模式，改变原有 GPU 中 Thread 与 kernel 的时钟周期，使之统一化，进而增大优化框架 GPU 的粒度。

3）动态映射和全局。转变原有 GPU 静态函数对应的映射方法，呈现全局性动态协作、映射协调的多对多模式。

4）本地队列。减少特征参数值的数据维度，降低信息数据读取的频率，利用 L-D-S 的方式实现信息数据的本地化共享。

据此，本研究工作实现了 BP 神经网络识别算法在 GPU 系统平台上的优化并行架构的搭建，并最终解决了 Thread 线程间负载不均衡及其相关次生问题，实现了并行型优化框架 GPU 的 BP 神经网络识别算法。相关实验数据所显示出来的加速比等具有代表性的性能参数也充分表明，以 NAÏVE 版本为标准构架的 GPU 系统平台及其并行框架的优化工作是成功的。

4　实验性能数据及其分析

性能评价以并行 GPU 优化架构的 AMD HD7970 和 NVIDIA GTX680 计算平台与 DJ8000 或 CPU 的串行算法版本计算平台等反向性证明的相应实验数据为依据。

在实际对比实验中，选取了更具典型性与针对性的 DJ8000 系统平台、基于 NAÏVE 标准构架版本优化的 GPU 并行框架对应的 AMD HD7970 与 NVIDIA GTX680 这 3 种不同计算平台为目标，随机抽取 2000 张 DJ8000 数据采集系统采集的识别源图像作为测试样本，并将其随机分为 4 组来

进行相同的实验，其中每个组都包括了若干个有形成分。对尿沉渣中有形成分的图像做分类识别实验，对所得出的相关性能指标作比较分析。由于各种采集图像样本中的源图像的有形成分的规格、数量和背景等都不尽相同，有些甚至大相径庭，因此优化的 GPU 并行框架系统平台会对识别图像进行前期的统一化处理。对有代表性的若干个有形成分进行预处理，得到的统一化识别图像（效果图）如图 6 所示。

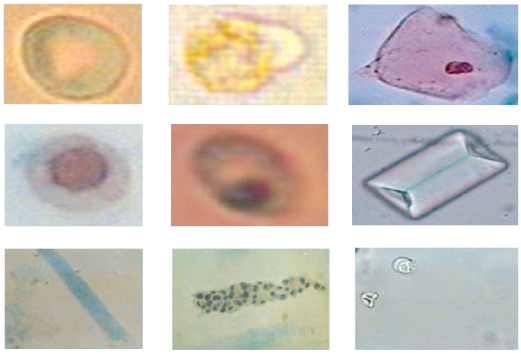


图 6 统一化识别图像

Fig. 6 Standardization of images

相关对比实验数据表明：在尿沉渣有形成分识别率（见表 1）基本稳定的情况下，不论是线程间的任务的运行时间还是动态协调加速比等核心性能参数，并行算法 GPU 优化架构的 2 个处理平台运算出的指标数值均优于串行 DJ8000 算法架构处理平台得出的指标数值。基于 3 个不同系统，4 个随机采集组的运算时间分别为：78.28 min，224.58 min，374.51 min，124.53 min（DJ8000）；10.26 min，18.92 min，24.51 min，13.76 min（NVIDIA GTX680）；7.23 min，15.68 min，17.54 min，10.08 min（AMD HD7970）。于是，NVIDIA GTX680 与 AMD HD7 970 对应于 OJ8000 系统的加速比率分别为：7.63，11.87，15.28，9.05 和 10.82，14.32，21.35，12.36（标准当量）。具体实验数值所对应的折线图和直方图如图 7 所示。

实验数据图直观地显示出基于 DJ8000 平台所采集的尿沉渣有形成分图片的数据，集成 NAÏVE 标准构架版本的 GPU，形成并行优化后的识别算法。相较于前期典型的标准版本 DJ8000 系统上的加速比等时效性性能，优化的 AMD HD7970 和 NVIDIA GTX680 两个并行处理 GPU 框架系统平台上相应的加速比参数值分别是前者的 10.82～21.35 和 7.63～15.28个标准当量，且线程间任务的运行时间也大大缩短。

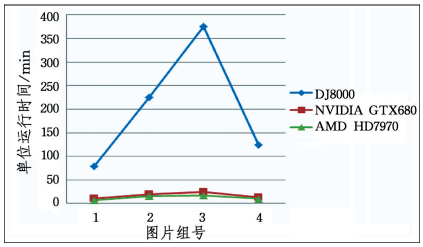
表 1 识别准确率

Table 1 Recognition accuracy results

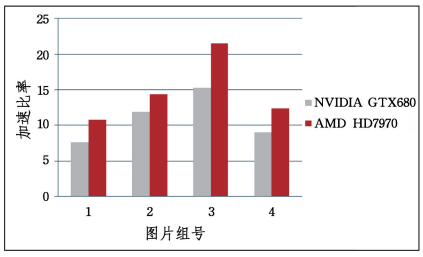
（单位：%）

Accuracy	Red Cell	White cell	Lympho-cyte	Crystal-lization	Epithelial Cells	Yeast-like fungus
DJ8000	92.2	91.1	87.3	91.0	88.5	79.5
NVIDIA GTX 680	91.8	90.4	85.98	91.1	91.3	76.7
AMD HD 7970	92.4	89.5	86.7	89.8	89.9	78.1

注：上述识别率数值为 4 组随机抽样总的平均值



(a)单位运行时间



(b)加速比

图 7 性能参数值对比

Fig. 7 Comparison of performance parameters

本研究所做的相关优化工作在基于尿沉渣有形成分识别准确率大体不变的前提下，不管是在 AMD HD7970 还是在 NVIDIA GTX680 等不同的系统平台下，相应的性能参数均达到 DJ8000 系统上加速比等时效性性能的数倍甚至是数十倍，完成了预期的高性能目标，自然也满足了医学检测的实时性要求。以上结论再次验证了本次实验工作相对已往工作推进的有效性和研究方向的正确性，它为今后深入研究该问题打下了坚实的基础，并提供了良好的工作平台。

结束语 本研究提出的基于 BP 神经网络识别算法优化的并行处理 GPU 框架的系统平台，突破了线程间负载不均衡导致的处理性能瓶颈，有效地解决了医学领域实时检测的时效性问题。由于本研究中的 AMD HD7970 和 NVIDIA GTX680 2 个 GPU 计算平台采用 DJ8000 采集处理系统采集的识别源图片，它们在数据信息对接时可能会产生时钟上的不一致，或是会因为软、硬件系统而出现进程延迟，这会使得整体的数据处理周期出现加长的假象，进而影响对系统时效性等性能指标的判断，这也是今后工作中需要进一步研究和细化的方面，它需要更科学而又精确的实验来进行定性和定量的分析。

致 谢 感谢 U. S. A. Kronos Research Institute of Boston、九州电子、南谷科技等对本科研项目的大力支持。

参 考 文 献

[1] JIAO L, YANG S Y, LIU F, et al. Seventy Years Beyond Neural Networks: Retrospect and Prospect[J]. Chinese Journal of Computers, 2016, 39(8): 1697-1716. (in Chinese)
焦李成，杨淑媛，刘芳，等. 神经网络七十年：回顾与展望[J]. 计算机学报，2016，39(8)：1697-1716.

[2] RAFAEL C G, RICHARD E W, EDDINS S L. 数字图像处理的 MATLAB 实现[M]. 北京：清华大学出版社，2013.

[3] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Image Net

classification with deep convolutional neural networks[C]// Proceedings of the Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1097-1105

[4] LIU Y C, ZHANG Y C. Image Colorimetry Segmentation Based on White Balance Algorithm[J]. Computer Engineering, 2012, 38(20): 195-196. (in Chinese)
刘玉成，张颖超. 基于白平衡算法的图像色度学分割. [J]. 计算机工程，2012，38(20)：195-196.

[5] TZENG S, ATNEY A, OWENS J D. Task management for irregular-parallel workloads on the GPU[C]// Proceedings of the Conference on High Performance Graphics. Ville, Switzerland, 2010: 29-37.

[6] SHARMA B, THOTA R, VYDYANATHAN N, et al. Towards a Robust, Real-time Face Processing System Using CUDA-enabled GPUs[C]// Proceedings of the 2009 IEEE International Conference on High Performance Computing. Kochi, India, 2009: 368-377.

[7] GHORAYEB H, STEUX B, LAURGEAU C. Boosted Algorithms for Visual Object Detection on Graphics Processing Units[C]// Proceedings of the 7th Asian Conference on Computer Vision. Hyderabad, India, 2006: 254-263.

[8] MERRILL D, GARIAND M, GRIMSHAW A. Scalable GPU Graph Traversal[C]// Proceedings of the 17th ACM SIGPLAN symposium on Principles and Parallel Pragramming. New York, USA, 2012: 117-128.

[9] AILA T, LAINE S. Understanding the Efficiency of Ray Traversal on GPUs[C]// Proceedings of the Conference on High Performance Graphics. New York, USA, 2009: 145-150.

[10] CEDERMAN D, TSIGAS P. On Dynamic Load Balancing on Graphics Processors[C]// Proceedings of the 23rd ACM Symposium on Graphic Hardware. Ville, Switzerland, 2008: 57-64.

[11] CHATTERJEE S, GROSSMAN M, SBIRLEA A, et al. Dynamic Task Parallelism with a GPU Work-Stealing Runtime System [C]// Proceedings of the 24th International Workshop on Languages and Compilers for Parallel Computing. Fort Collins, USA, 2011: 203-217.

[12] NVIDIA Corporation. NVIDIA GeForce GTX 750 Ti: Featuring First-Generation Maxwell GPU Technology. Designed for Extreme Performance per Watt[R]. IEEE, 2014.

[13] AMD Corporation. Accelerated Parallel Processing OpenCLTM [R]. IEEE, 2014.

[14] CIRESAN D, MEIER U, MASCI J, et al. A committee of neural networks for traffic sign classification[C]// Proceedings of the International Joint Conference on Neural Networks. San Jose, USA, 2011: 1918-1921.

[15] JIA Y Q, SHELHAMER E, DONAHUE J, et al. Caffe: convolutional architecture for fast feature embedding[C]// Proceedings of the ACM International Conference on Multimedia. Orlando, FL, USA, 2014, 675-678.

[16] LIU M, QUAN T, LUAN S. An attribute recognition system based on rough-set theory-fuzzy neural network and fuzzy expert system[C]// Fifth World Congress on Intelligent Control and Automation, 2004(WCICA 2004). IEEE, 2004: 2355-2359.

[17] PEI Y K. Study on Urine Sediment Recognition System Based on Neural Network and Fuzzy Reasoning[D]. Nanjing: Nanjing University of Science and Technology, 2008. (in Chinese)
裴元焜. 基于神经网络与模糊推理的尿沉渣识别系统研究[D]. 南京：南京信息工程大学，2008.

[18] SHEN M L. Study on Urinary Sediments Visible Component Automation Clssfication System[D]. Changchun: Changchun University of Science and Technology, 2006. (in Chinese)
沈美丽. 尿沉渣有形成分自动分类系统研究[D]. 长春：长春理工大学，2006.

[19] BURTSCHER M, NASRE R, PINGALI K. A Quantitative Study of Irregular Programs on GPUs[C]// Proceedings of the 2012 IEEE International Symposium on Workload Characterization. La Jolla, USA, 2012: 141-151.

[20] NASRE R, BURTSCHER M, PINGALI K. Atomic-free Irregular Computations on GPUs[C]// Proceedings of the 6th Workshop on General Purpose Processor Using Graphics Processing Units. New York, USA, 2013: 96-107.

[21] NASRE R, BURTSCHER M, PINGALI K. Data-Driven Versus Topology- driven Irregular Computations on GPUs[C]// Proceedings of IEEE 27th International Symposium on Parallel & Distributed Processing. Boston, USA, 2013: 463-474.

[22] YAN S G, LONG G P, ZHANG Y Q. StreamScan: Fast Scan Algorithms for GPUs Without Global Barrier Synchronization[C]// Proceedings of the 18th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. Guangzhou, China, 2013.

(上接第 274 页)

[11] LIN T L, CHANG T, HUANG G X, et al. Improved interview video error concealment on whole frame packet loss [J]. Journal of Visual Communication and Image Representation, 2014, 25(8): 1811-1822.

[12] YAN K, YU M, PENG Z J, et al. A new low-complexity error concealment method for stereo video communication[J]. Journal of Optoelectronices Laser, 2015, 26(11): 2200-2208.

[13] GONZALES C, WOODS R E. Digital Image Processing(3rd Ed) [M]. NewJersey, Addison-Wesley, 2007: 741-742.

[14] XIA Y Q, YANG J Y. Two level stereo match approach based on maximum-window [J]. Computer Science, 2006, 33(3): 208-211. (in Chinese)
夏永泉，杨静宇. 基于最大窗口的二次立体匹配方法[J]. 计算机科学，2006，33(3)：208-211.

[15] JVT reference software home page [OL]. http://iphome.hhi.de/suehring/tml.

[16] ZHOU Y, XIANG W, WANG G K. Frame loss concealment for multiview video transmission over wireless multimedia sensor networks [J]. IEEE Sensors Journal, 2015, 15(3): 1892-1901.