Full Research Paper

Parallel Algorithm for GPU Processing; for use in High Speed Machine Vision Sensing of Cotton Lint Trash

Mathew G. Pelletier

United States Department of Agriculture, Agricultural Research Services, Lubbock Tx. USA
E-mail: mpelletier@lbk.ars.usda.gov

Abstract: One of the main hurdles standing in the way of optimal cleaning of cotton lint is the lack of sensing systems that can react fast enough to provide the control system with real-time information as to the level of trash contamination of the cotton lint. This research examines the use of programmable graphic processing units (GPU) as an alternative to the PC's traditional use of the central processing unit (CPU). The use of the GPU, as an alternative computation platform, allowed for the machine vision system to gain a significant improvement in processing time. By improving the processing time, this research seeks to address the lack of availability of rapid trash sensing systems and thus alleviate a situation in which the current systems view the cotton lint either well before, or after, the cotton is cleaned. This extended lag/lead time that is currently imposed on the cotton trash cleaning control systems, is what is responsible for system operators utilizing a very large dead-band safety buffer in order to ensure that the cotton lint is not under-cleaned. Unfortunately, the utilization of a large dead-band buffer results in the majority of the cotton lint being over-cleaned which in turn causes lint fiber-damage as well as significant losses of the valuable lint due to the excessive use of cleaning machinery. This research estimates that upwards of a 30% reduction in lint loss could be gained through the use of a tightly coupled trash sensor to the cleaning machinery control systems. This research seeks to improve processing times through the development of a new algorithm for cotton trash sensing that allows for implementation on a highly parallel architecture. Additionally, by moving the new parallel algorithm onto an alternative computing platform, the graphic processing unit "GPU", for processing of the cotton trash images, a speed up of over 6.5 times, over optimized code running on the PC's central processing unit "CPU", was gained. The new parallel algorithm operating on the GPU was able to process a 1024x1024 image in less than 17ms. At this improved speed, the image processing system's

performance should now be sufficient to provide a system that would be capable of real-time feed-back control that is in tight cooperation with the cleaning equipment.

1. Introduction

In the cotton ginning industry, knowledge of the cotton grade, during the processing stages, gives the cotton ginner a unique advantage and opportunity to reduce the amount of lint lost, during the lint-cleaning stage, while maintaining a quality product. This advantage is enhanced when the information, regarding the amount of trash in the cotton, is coupled with one of the new lint cleaning machines that have the ability to dynamically adjust how much cleaning they perform on the cotton, figure 1. By dynamically measuring the cotton trash content in real-time, the system could be automatically tuned to avoid wasting valuable lint as well as preventing fiber damage, from over-cleaning, which leads to a higher quality product and hence a more valuable commodity. This is beneficial to all aspects of the industry. New developments in the online classing of cotton, has recently seen the deployment of custom cotton classing systems that utilize computer vision systems to accomplish this task. There is however a lack of real-time sensors that are fast enough to allow for determination of the trash content immediately before the machine performs the cleaning. The lack of suitable rapid response sensing systems creates a disconnect between the cotton that's measured to the cotton that's being cleaned. Current state of the art sensors typically sense the product's trash levels after the cleaning has already taken place or so far before the cleaning takes place that it's impossible to synchronize the cleaning with the lint being cleaned. This disconnect, results in the users compensating through an increased dead-band range that in effect severely under-utilizes the dynamic cleaning technology. This research was undertaken with a view to developing new technology that is capable of analyzing the cotton fast enough to provide an immediate response to the cleaning equipment so the cotton that is being analyzed is the cotton that's being cleaned.
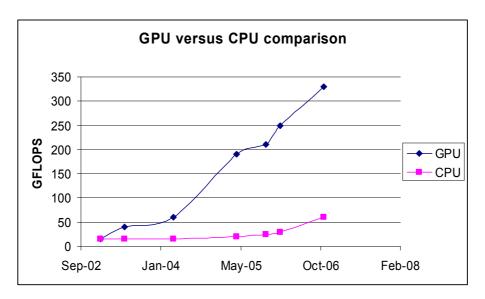
Figure 1. Lint cleaner developed by USDA-ARS that is dynamically adjustable. Photo's courtesy of Continental Eagle Corp.

2. Background

To date, a limited amount of computer vision applications have been accomplished on general purpose microcontrollers due to the lack of computing power and the necessary high data transfer rates. Typical microcontrollers run at 20MHz or less with a few exceptions running at 40MHz-80MHz. In addition to these reduced central processing unit (CPU) speeds, the microcontrollers are further limited by being designed as 8 or 16 bit processors with only a general purpose instruction set and limited arithmetic logic unit (ALU). Under these constraints a computer vision application would take minutes if not hours to complete a single complex image analysis. These low powered embedded processors simply don't have the computational capability that an image processing system requires to be effective in a dynamically changing control system. This fact has kept the bulk of computer vision applications tethered to the PC or a high end digital-signal-processing (DSP) platform that consists of a DSP with either a video frame-grabbing chipset or a customized processor embedded directly into hardware via implementation onto an application specific processor (ASIC) or a field-programmable-gate-array chip (FPGA) (Buck et al., 2004; DeCoro and Tatarchuk, 2007; Galoppo et al., 2005; Kipfer et al, 2004; Kolb, 2005; Krueger and Westermann, 2003; Larsen and McAllister, 2001; Scheuermann and Hensley, 2007; Strzodka et al, 2005). In recent years however, a new alternative has emerged from the computer graphics arena targeted primarily at the PC based gaming industry. These new graphic processors "GPUs" have now evolved to the point where they have numerous CPUs on board the chip that are designed to run in a massively parallel architecture. Of particular interest are the recent advances of the GPUs towards fully programmable vector stream processing systems that are inherently parallel in nature. The results of these developments is to provide the scientific processing community with GPU architectures that have now advanced to the point where they provide a much faster alternative to computing of highly parallel algorithms, (figure 2) than the traditional single-pass Von Neumann architecture (Bell and Newell, 1971) of the traditional PC CPU design. This paper presents the development and testing of a highly parallizable algorithm for cotton trash image processing for use on one of the latest GPUs.
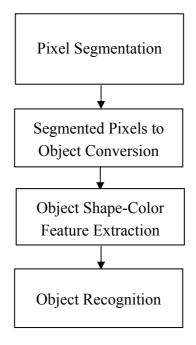
Figure 2. Graph detailing incredible performance increase over the last several years of the GPU over the traditional general purpose CPU on today's modern PC's. (Source: Nvidia 2006). Note: GFLOPS is defined as 1 trillion floating point operations per second.

3. Results and Discussion

In the interest of developing a rapid analysis system, a new machine vision processing algorithm has been developed that was designed to provide a highly parallel approach to the cotton trash identification problem. As the cotton trash system in practice is, in many instances, a retrofit system that is placed into existing systems, it's been found that controlled lighting is less controlled than would be considered optimal for alternative algorithms such as a pre-computed lookup table based on a Bayesian Classifier approach, (Pelletier 1999a,b, 2003). The main problem with the Baye's Classifier approach is the need for pre-computation of the Bayesian statistics, typically provided in the form of presorted classes obtained by an expert or human classer. Thus, it's not practical to dynamically adjust the Bayesian statistics as the expert is not available for periodic recalculation of the Bayesian statistics. This requirement forces the Bayesian approach to demand that the system provide a stable environment where the image statistics are unchanging. However, in trial deployments into several commercial installations, it was found that the static-image-statistic's criteria, was not valid. In practice, the changing lighting conditions and system placement as a retrofit onto various types of machines, typically creates a wide variation in the image statistics for each member of the feature set {trash, background, lint}; primarily due to the fact that each member of the feature set moves in and out of full or partially lit areas or becomes alternatively and repeatedly immersed in lighting and then later in shadows. To compensate for the widely changing lighting environments, encountered in typical commercial installations, required an alternative image processing algorithm to overcome the difficulties of the varying statistics. The new developments that have been brought about by this research were also coupled with the additional goal of increasing the processing speed of the algorithm to achieve a robust system that would also be capable of performing real-time trash feedback control. In an effort to obtain higher processing speeds, the research developed the algorithm with a goal to obtain a highly parallel algorithm suitable for use on highly parallel vector processors.

Figure 3. Image analysis to extract the quantities of the various trash constituents.

The basic overview of the image processing algorithm, figure 3, shows the steps required to process the image from raw color pixels into a set of statistics to inform the mechanical cleaning system of the quantity and type of trash; the basic information required by an optimal imaging/mechanical control system. The start of the image processing algorithm is to process each pixel, by analysis of the current target pixel against the target pixel's local neighboring pixels with the goal to determine or classify the target pixel into either lint or trash, figure 4. Noting that the bulk of the time required for the image processing algorithm is tied to this first step of pixel identification, the focus of the new development was to optimize the processing of this stage of the algorithm.
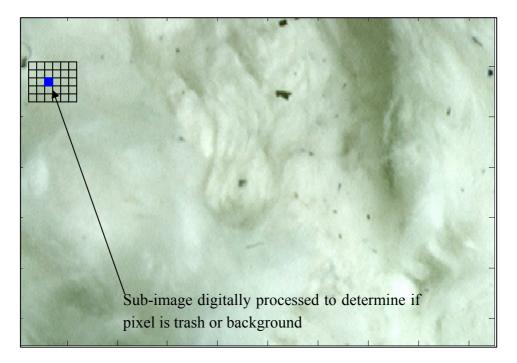
Figure 4. Sub-image is analyzed to determine if pixel is trash or lint.

Sub-image digitally processed to determine if pixel is trash or background

The new algorithm, under investigation, was developed around a rapid single-pass Gaussian band-pass convolution kernel, "GBPCK", that effectively partitions the color space such that a simple threshold operation following the GBPCK will allow for the generation of a binary image where each pixel is classed to be either a trash or lint pixel. In practice, the GBPCK was shown to be remarkably robust across a wide variety of lighting situations. The single-pass Gaussian band-pass convolution kernel, "GBPCK", is implemented on a 7x7 finite impulse response, "FIR" two-dimensional convolution kernel or filter. As an FIR filter operates solely on the existing data and utilizes no feedback, the filtering operations operate solely on the incoming image without modifying it during the processing of the image. All results from the FIR filtering process are then stored in a second processed image. This non-feedback requirement of the FIR filter is a key enabling feature that allows for the implementation of the algorithm to be performed simultaneously on each pixel in parallel.

The basis for the Gaussian-normal band-pass filter is derived from the Gaussian low-pass filter, which is illustrated in equation 1:

$$G_{bpf}(r,\sigma) = 1/(\sigma \sqrt{(2\pi)})\ e^{-r^2/(2\sigma)} \qquad\qquad (1)$$

Where:

      r := distance from center of non-causal filter

      $r = \sqrt{(x^2 + y^2)}$

      σ := Gaussian half-width

For illustration, we present a simplified form of the Gaussian band-pass filter that can be constructed from the difference of two Gaussian low-pass filters with differing extents, as shown in equation 2:

$$G_{bpf}(r) = 1/(\sigma \sqrt{(2\pi)})\ e^{-r^2/(2\sigma)} - 1/(\rho \sqrt{(2\pi)})\ e^{-r^2/(2\rho)} \qquad\qquad (2)$$

Where:

      r := distance from center of non-causal filter

      $r = \sqrt{(x^2 + y^2)}$

      ρ := spread of the Gaussian filter 1

      σ := spread of the Gaussian filter 2

In practice the Gaussian band-pass filter was comprised of the sum of several Gaussian filters. By utilization of multiple-cascaded Gaussian filters, the shape of the Gaussian curve can be highly tuned for both extent and fall-off, allowing for optimum processing for the specific application. In order to optimize the calculation of the filter in real-time, the filter coefficients were pre-calculated. For the research subject under investigation, the discrete two-dimensional Gaussian band-pass filter was implemented from the consolidated cascade of multiple Gaussian-normal filters as detailed in equation 3:

$$G_{bpf}(m,n) = \beta * e^{-(1/4*r)^2} + 2\beta\ e^{-(1/2*r)^2} + (4\beta - \alpha)*e^{-r^2} + 8\beta\ e^{-(2*r)^2} + 16\beta\ e^{-(4*r)^2} \qquad (3)$$

Where:

      $r = \sqrt{(x^2 + y^2)}$:= distance from the center of the convolution kernel

      $\beta = 0.0108$

      $\alpha = 0.3182$

To gain insight into how the Gaussian band-pass filter, hereafter known as the GBPCK, is affecting the image, the frequency response of the discrete two-dimensional filter of equation 3 was calculated using the discrete-Fourier-Transform to transfer from the discrete spatial position domain to the

discrete frequency domain (Strum and Kirk, 1988; Jain, 1989; Porat, 1998) where the two-dimensional discrete-Fourier-Transform is illustrated in equation 4:

$$H(k, g) = \sum \sum h(m,n) \, e^{-j(2\pi/N) \, k \, n} \, e^{-j(2\pi/M) \, g \, n} \tag{4}$$
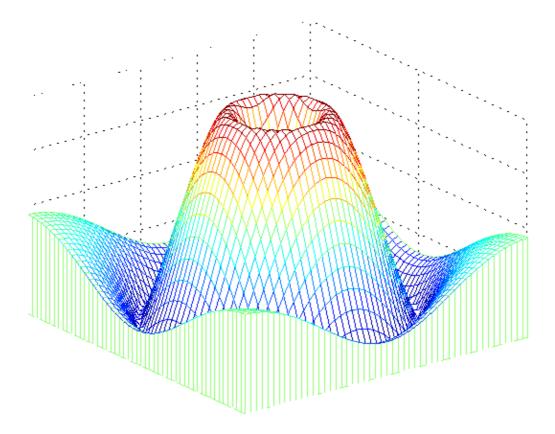
Summed over the interval $n = 0$ to $n = N-1$ & $m = 0$ to $m = M-1$

Where:

   N :- number image points in the sampled x-dimension
   M :- number image points in the sampled y-dimension
   $j := \sqrt{(-1)}$

The discrete frequency response of the filter, as calculated from equation 4, via the fast Fourier-Transform (FFT) algorithm, is shown in figure 5. For clarity, the one-dimensional cross-section of the filter is shown in figure 6.

Figure 5. Frequency response of the two-dimensional Gaussian band pass filter for highly parallel, lighting independent, rapid trash analysis.

We note here that one of the key criteria's for online classing is the ability of the algorithm to separate cotton from trash in both lit and shadowed areas. One example of the performance of the GBPCK suitability for performing this task is detailed in figure 7.

Figure 6. Frequency response of the one-dimensional cross-section of the 2D Gaussian band pass filter detailed in figure 5.
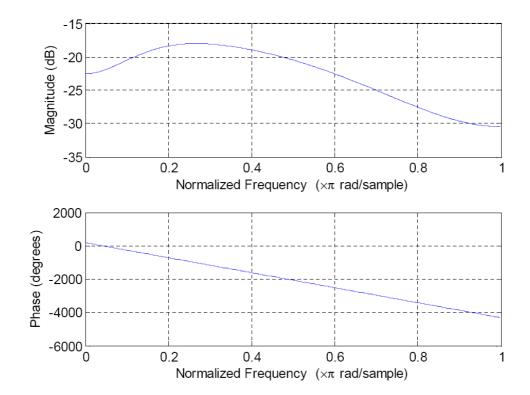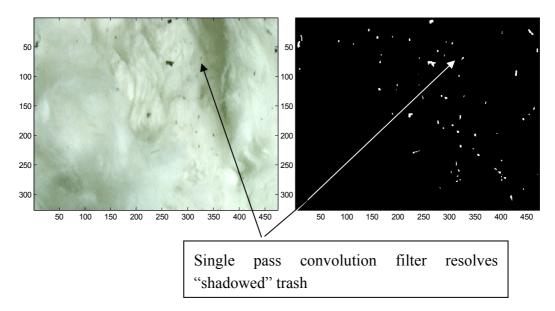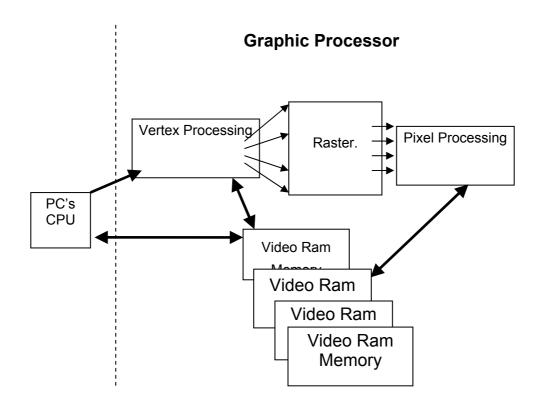

Figure 7. Performance of new single pass GCK algorithm.

Single pass convolution filter resolves "shadowed" trash

Once the development of a single pass filter was completed, the next task was to fine tune the implementation of the filter to effect the fastest processing on the given hardware. To gain insight into areas that would provide a meaningful speed-up, which was one of the primary goals of this research, it was also crucial to provide a baseline performance by which to judge the GPU approach. The algorithm was initially optimized for use on a Pentium 4 processor using the extended operation set for "Single Instruction for Multiple Data" or SIMD. The SIMD extension for the Pentium 4 provides a

single vector processor that is capable of multiplying 4 single-precision floating point numbers in parallel. Performance of the algorithm after adjustment to take advantage of the Pentium's SIMD CPU chipset extensions as well as inline expanded and optimized C code, resulted in a processing time of 7.5 frames per second. This performance increase represents a significant speedup over the previous algorithm implementation of 2.5 frames per second. The next step in the development was to compare the optimized SIMD performance to the same algorithm running on an NVIDIA GeForce 8800 Ultra GPU graphics processing unit, housed on a pci-express bus card, where the code would then have the opportunity to take advantage of the GPU's 132 vector processors. We note here that while the GPU has 132 vector processors, each capable of multiplying 4 single precision floating point numbers in parallel, the core is only running at 500 MHz versus the Pentium's core at 3.0GHz. Given the speed disparity between the GPU processor to the Pentium core, one cannot expect a speedup of 132 times for the GPU over the CPU. Other potential problem areas for implementation on the GPU platform lies in the bottle necks that are created by pushing large amounts of image data across the pci-express bus into the GPU's video ram, figure 8. In short, one may not even expect a 132-X times (500MHz/3000MHz) = 22-X gain from running on the GPU core versus normal operations that take place via computation on the CPU due to other hardware constraints.

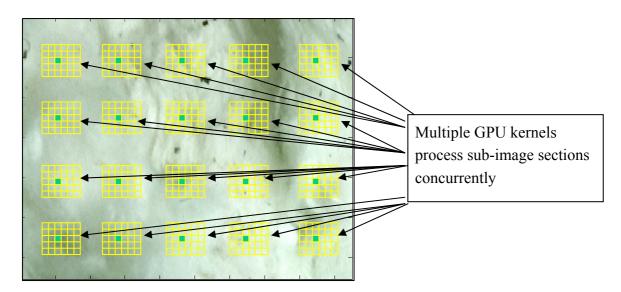**Figure 8.** Processing flow using a GPU processor

In moving the algorithm off of the PC's traditional computing platform, the CPU, to the GPU; the developers of the modern GPUs have designed the system such that the CPU passes both an algorithm as well as the image and geometry data to the GPU. To effect efficient general computation on the GPU, the algorithm must be transformed to fit the highly specialized stream processing the GPU was designed to perform, figure 8. As such only certain algorithms that are inherently parallel in nature,

can be converted, however for those suitable algorithms, once the transformation is made, the massively parallel architecture of the modern GPUs becomes available which enables dramatic increases in performance over the traditional CPU style computations. To ease the transformation process, the graphic card developers have developed an augmented C programming language that allows one to specify how the image is to be broken into numerous sub-images that will all be computed concurrently as well as the algorithm each vector processor should execute. By combining both the GBPCK algorithm with the threshold and non-linear median-filter shot-noise reduction operation, to form a combined GBPCK-TM analyzer, it allowed the algorithm to be moved completely off of the CPU and onto the GPU. The new cotton trash analyzer processing program now has the ability to break the large image into numerous sub-images that can all be analyzed concurrently as detailed in figure 9. Once the GPU analyzes the data per specification of the GBPCK-TM, it then transfers the fully analyzed binary image solution back to the CPU along with trash statistics such as trash content. In testing of the GBPCK-TM algorithm for cotton trash identification, we found the following:

- Transitioning from the Bayesian Classifier to a single pass GBPCK, was enabling technology as commercial use dictated retro-fitting the system into on-line environments which resulted in wide fluctuations in the positioning of the lint bat, which in turn created large fluctuations in both the light intensity as well as large variations depth and quality of the shadowed lint. These lighting inconsistencies resulted in large degradations in the quality of the performance of the Bayesian classifier. When tested against these same lighting fluctuations, the GBPCK performance was significantly improved over the Bayesian Classifier.

- In practice the Bayesian Classifier in practice required extensive on-site training. In contrast, the GBPCK provided a robust rapid startup as no on-site training was required and eliminated the need for ambient light shielding, thereby improving system's ability to be retrofit into a wider variety of machine designs as well as reducing manufacturing and installation system costs.

- We note herein for those not familiar with cotton classification, that the standard by which all cotton lint is graded, is the human visual system. An elaborate process has been developed over the last century that enables for a stable transition from year to year of the cotton classing grade. Each year in Memphis Tennessee, a standard set of boxes, holding lint samples that are representative of the color and trash grade. Each classing office throughout the US utilizes a set of these boxes as the defacto standard for any lint sample that is in question as to the correct grade. As such, the human visual system is the standard which is judged against a contiuously changing historical standard. Given this standard, the best test for performance of the system is a visual analysis of the quality of the recognition, which can be best judged by examination of the included figure 7. We also note that when the system was run against the official photograph prints of the cotton classing standards, used daily by USDA-AMS cotton classers, the system was able to predict the grade with a coefficient of determination of r^2=0.99. While it is also recognized that with the shadowing, the performance will be degraded, field tests indicated a level of performance comparable to the correlations achieved between human classers during periodic re-grading of the lint that occurs across the course of the season at the request of customers, typical r^2=0.80-0.85. In house testing suggests the performance with shadows is within ½ leaf grade, which is typically within 1 standard deviation of the natural variability of the lint in the ginning process.

- By transitioning from the Bayesian Classifier to a single pass GBPCK, along with optimizations of the algorithm, improvements over processing speed was gained.
    - Optimization of algorithm effected a speed up of 2.5 times (7.5 frames/sec).
- By moving the code from the CPU to the GPU and utilizing the combined GBPCK-TM algorithm, further improvements were gained:
    - When utilizing a single Nvidia GeForce 8800 GPU with 132 vector processors; a speed up of 20 times was gained (60 frames/sec) over the Bayesian Classifier and a speed up of eight was gained over running highly optimized code on the CPU.
- At 60 Frames/second:
    - 71 sq-m.of cotton can be imaged/bale with the new GBPCK-TM algorithm running on the GPU.
- versus 2.5 Frames/second, utilizing the previous Bayesian Classifier approach:
    - 3.5 sq-m.of cotton can be imaged/bale.

Figure 9. Massively parallel approach to image processing showing how numerous sub-images are all analyzed concurrently using 132 vector processors on a GPU.

4. Conclusions

As the cotton ginning industry moves toward machines that have the ability to dynamically adjust the amount of cleaning the machine performs, a great deal of valuable lint is saved and there is a significant reduction in fiber damage as well. The missing element at this time is the ability of the sensors to determine the required amount of cleaning for the cotton as it is feed into the machine. This research has demonstrated that through the use of massively parallel processing, that is now possible on today's programmable GPUs, a machine vision algorithm suitable for real-time classing of cotton, can be processed in a significantly reduced time that is sufficient to open the door for the possibility of processing the trash content of the incoming lint in time to set the machine so that it cleans the cotton that was just analyzed. This just-in-time analysis provides the enabling technology with the capacity to allow for a system that is optimized to clean the cotton that is being fed into the machine at the

precise optimal cleaning level. Once this transition from the current system, that looks at a sample of cotton taken either way before processing or after the cleaning has already taken place, moves toward one where the machine is cleaning the cotton that was analyzed as it is being fed into the machine; performance gains are expected to be upwards of 30% in reduction of valuable lint loss. This level of improvement can be expected due to the fact that today's systems use a very large dead-band to protect the users against both the inherent wide variability in the cotton lint's trash distribution, as well as the potential for changes to take place before the machine can react to the changing cotton trash levels. It is also expected that this technology will likely drive new machine designs that can not only optimize the cleaning across the entire width of the machine, which effectively cleans 100% of the cotton to remove the trash from the 4% of the cotton that actually have trash particles, towards a machine that only cleans the 4% of the cotton that actually contains the trash. Once this level of automation is reached, significant reductions in lint loss as well as subsequent reductions in fiber damage will also become possible as there should be an additional 96% reduction in lint loss and fiber damage. As the typical first stage lint cleaners generate upwards of 4.5-7 kg of lint loss, if 4.5 kg of lint per bale can be saved, this would represent a $100 million U.S. of added annual revenue to US cotton growers as well as a similar amount for the international growers.

References

1. Buck, I.; Foley, T.; Horn, D.; Sugerman, J.; Fatahalian, K.; Houston, M.; Hanrahan, P. Stream computing on graphics hardware. *ACM Proc. SIGGRAPH* **2004**, *23*, 777-786.
2. Jain, A.K. *Fundamentals of Digital Image Processing*; Prentice-Hall: New Jersey, 1989; pp. 15-18.
3. Kipfer, P.; Segal, M.; Westermann, R. Uberflow. A GPU-based particle engine. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*, Grenoble, France, 2004; ACM: New York, NY, USA, 2004; pp. 115-122.
4. Kolb, A.; Latta, L.; Rezk-Salama, C. Hardware-based simulation and collision detection for large particle systems. In *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*, Grenoble, France, 2004; ACM: New York, NY, USA, 2004; pp. 123-131.
5. Krueger, J.; Westermann, R. Linear algebra operators for GPU implementation of numerical algorithms. *ACM Transactions on Graphics* **2003**, *22*(3), 908-916.
6. Pelletier, M. G.; Barker, G. L.; Baker, R. V. Non-contact image processing for gin trash sensors in stripper harvested cotton. *ASAE* **1999**, Paper #991126, 9 p.
7. Pelletier, M. G.; Barker, G. L.; Baker, R. V. Image processing for stripper harvested cotton trash content measurement: a progress report. In *Proc. Beltwide Cotton Prod. Res. Conf.*, 1999; pp. 1412-1415.
8. Pelletier, M. G.: Real time measurement system for seed cotton or lint. U.S. Patent, No. 6,567,538; 2003.
9. Porat, B. *A course in digital signal processing*; John Wiley and Sons, Inc.: New York, 1997; pp. 93-159.
10. Strum, R. D., D. E. Kirk. *First Principles of Discrete Systems and Digital Signal Processing*. Addison-Wesley Publishing Co.: Reading, MA, 1988; pp. 363-527.

11. Strzodka, R.; Doggett, M.; Kolb. A. Scientific computation for simulations on programmable graphics hardware. *Simulation Modelling Practice and Theory* **2005**, *13*(8) 667-680 (Special Issue: Programmable Graphics Hardware).