Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Compute-unified device architecture implementation of a block-matching algorithm for multiple graphical processing unit cards

Francesc Massanes
Marie Cadennes
Jovan G. Brankov

Compute-unified device architecture implementation of a block-matching algorithm for multiple graphical processing unit cards

Francesc Massanes
Marie Cadennes
Jovan G. Brankov
Illinois Institute of Technology
Medical Imaging Research Center
Chicago, Illinois 60616
E-mail: brankov@iit.edu

Abstract. *We describe and evaluate a fast implementation of a classical block-matching motion estimation algorithm for multiple graphical processing units (GPUs) using the compute unified device architecture computing engine. The implemented block-matching algorithm uses summed absolute difference error criterion and full grid search (FS) for finding optimal block displacement. In this evaluation, we compared the execution time of a GPU and CPU implementation for images of various sizes, using integer and noninteger search grids. The results show that use of a GPU card can shorten computation time by a factor of 200 times for integer and 1000 times for a noninteger search grid. The additional speedup for a noninteger search grid comes from the fact that GPU has built-in hardware for image interpolation. Further, when using multiple GPU cards, the presented evaluation shows the importance of the data splitting method across multiple cards, but an almost linear speedup with a number of cards is achievable. In addition, we compared the execution time of the proposed FS GPU implementation with two existing, highly optimized nonfull grid search CPU-based motion estimations methods, namely implementation of the Pyramidal Lucas Kanade Optical flow algorithm in OpenCV and simplified unsymmetrical multi-hexagon search in H.264/AVC standard. In these comparisons, FS GPU implementation still showed modest improvement even though the computational complexity of FS GPU implementation is substantially higher than non-FS CPU implementation. We also demonstrated that for an image sequence of 720 × 480 pixels in resolution commonly used in video surveillance, the proposed GPU implementation is sufficiently fast for real-time motion estimation at 30 frames-per-second using two NVIDIA C1060 Tesla GPU cards.*
 [DOI: 10.1117/1.3606588]

1 Introduction

Motion estimation in an image sequence has many potential uses such as detecting and tracking moving objects in video surveillance,[1] removal of temporal redundancy in video coding,[2–4] motion compensated filtering applied along a motion trajectory in medical imaging,[5] or motion compensated digital subtraction in angiography.[6] In all of these applications, a potential drawback is computation time needed for motion estimation.

Paper 10080R received May 20, 2010; revised manuscript received Feb. 4, 2011; accepted for publication Jun. 13, 2011; published online Aug. 12, 2011.

1017-9909/2011/20(3)/033004/10/$25.00 © 2011 SPIE and IS&T

Current applications requiring real-time motion estimation often use parallel designs for very-large-scale integration (VLSI) devices. For example, in Refs. 7 and 8, the block-matching algorithm (BMA) was implemented on a VLSI device. It is well known that these implementations are usually costly, difficult, and time consuming to develop. Some alternative non-GPU (graphical processing unit) configurable architecture approaches, targeting video coding application, are reviewed in Ref. 9.

There have been a number of reported efforts to use GPU cards for motion estimation as a part of a video coding scheme, see Ref. 3 for a review. Most of the GPU implementation attempts originated before the introduction of the compute unified device architecture (CUDA),[10,11] a computing engine developed by NVIDIA to facilitate easy use of the GPU. The early GPU implementations without CUDA are often complex and hard to understand. For some recent implementation of variable block size motion estimation as a part of H.264/AVC coding using CUDA, see Refs. 4 and 12.

In this work we present an easy to understand, general purpose BMA with full grid search (FS) using CUDA computing engine and multiple NVIDIA GPU cards. Our intention is twofold: to develop a fast and accurate motion estimation for use in real-time video sequence processing and to develop a good case example for understanding the CUDA environment with use of single or multiple GPU cards.

There are many other relevant motion estimation models (see Ref. 13 for a review), such as pixel-[14] or region-based,[15] even with variable region size.[13] In this work, we chose a block-matching model with the blocks to be equal and rectangular, however, implementing different estimation models will not significantly change the presented CUDA implementation. As such, the speedup should not be diminished. In addition, we chose the BMA because it has a very high computational cost; also, it is commonly used in video coding such as in MPEG and H.264/AVC,[16] and as such can benefit from parallel implementation on single or multiple GPU cards. Therefore, the presented evaluation is not only a case

study but a relevant implementation that one may consider using in video coding applications.

CUDA allows easy and straightforward implementation of motion estimation algorithms using standard C, with NVIDIA extensions, making program development fast. In addition, the GPU computation hardware like NVIDIA Tesla (C1060-CUDA compute capability: 1.3, released Q2 2008) delivers a staggering 933 GFLOPS in single precision at a cost of less than $1000 per single unit where six core Intel Core i7 980 XE delivers theoretical 40.0 GFLOPS (as of March 2010).

It is therefore possible to have great achievements with CUDA technology such as the computation of a shortest path in a 10 million vertex graph in less than 2 s (Ref. 17) or an implementation of a simple color electroholography reconstruction system, which are 1000× faster than the traditional computation platforms.[18]

In Sec. 2, we will explain the basics of motion estimation using BMA followed by a description of GPU hardware. The algorithm design and implementation are given in Sec. 3. Section 4 contains experimental results.

2 Background

In this section we will briefly explain the basics of the BMA motion estimation method and introduce the CUDA computational model. Knowledge in both areas is needed to fully appreciate the algorithm design described in Sec. 3.

2.1 *Block-matching Algorithm for Motion Estimation*

The block-matching algorithm[19] is the most popular method for the motion estimation[13] of local motion in an image sequence. This method essentially splits an image, of $I \times J$ pixels in size, into $K \times L$ blocks, and estimates each block displacement vector $\mathbf{v}$ (also called the motion vector). For each block $B_{k,l}, k = 1, \ldots, K, l = 1, \ldots, L$ this is achieved by minimizing the matching criterion between the reference and target image over all possible candidate displacement $\mathbf{v}$ within the search window S, as illustrated in Fig. 1.

There are several choices for matching criterion.[19] In this work, we adopted block summed absolute difference (SAD) between the reference and target image pixel defined as

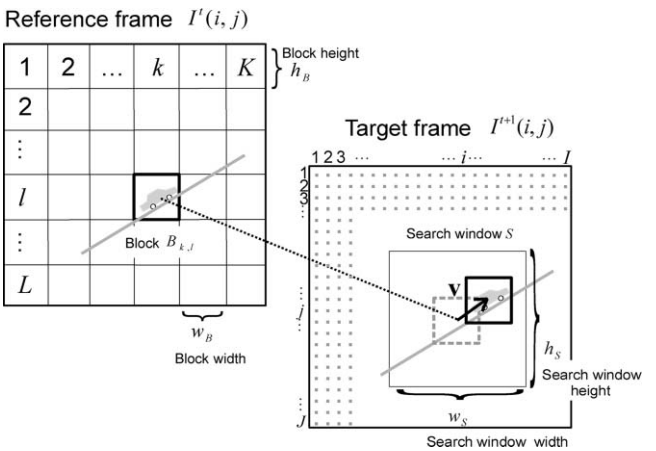


Fig. 1 Block-matching motion estimation; reference, target frame, and block displacement vector $\mathbf{v}$.

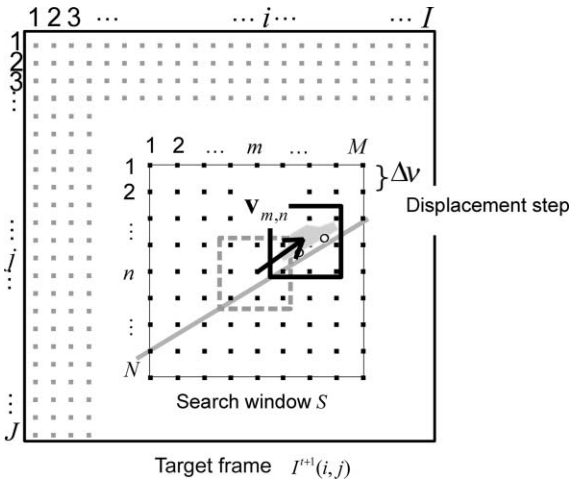


Fig. 2 Block-matching methods search window and grid.

$$J_{k,l}(\mathbf{v}) = \sum_{(i,j)\in B_{k,l}} |I^t(i,j) - I^{t+1}(i+v_1, j+v_2)|. \quad (1)$$

Here $\mathbf{v} = [v_1, v_2]^T$ is the $B_{k,l}$ block displacement, $I^t(i,j)$, $i = 1, 2, \ldots, I, j = 1, 2, \ldots, J$ represents the reference image intensity of the (i,j) pixel at a time frame t, $I^{t+1}(i,j)$ represents the target image intensity at a time frame $t+1$, and w_B, h_B are the block width and height.

Now we can define optimal displacement for block $B_{k,l}$ as

$$\mathbf{v}^* = \begin{array}{l} \arg\min_{\mathbf{v}} J_{k,l}(\mathbf{v}) \\ subject\, \mathbf{v} \in S \end{array}, \quad (2)$$

where S denotes the search window of w_S, h_S pixels in width and height.

In order to make this problem computationally feasible, we will restrict the possible displacement values of $\mathbf{v}$ to a discrete regular grid (see Fig. 2)

$$\mathbf{v}_{m,n} = \begin{bmatrix} \left(m - \frac{M-1}{2}\right)\Delta v \\ \left(n - \frac{N-1}{2}\right)\Delta v \end{bmatrix}, \; m = 1, 2, \ldots, M,$$
$$n = 1, 2, \ldots, N, \quad (3)$$

where Δv is the grid step size and $M = {}^{w_S}/_{\Delta v}$ and $N = {}^{h_S}/_{\Delta v}$ are the number of grid points.

In addition, if the search grid locations are restricted to integer values, e.g., $\Delta v = 1$, then the $J_{k,l}(\mathbf{v})$ calculation will not require interpolations of $I^{t+1}(i+v_1, j+v_2)$ image values. In this work, we will examine both integer and noninteger valued search grids.

Finally, in order to avoid influence of noise and false displacement vectors, e.g., in uniform regions, Eq. (2) is modified to suppress motion of blocks where the SAD is too small. This is mathematically described as

$$\tilde{\mathbf{v}} = \begin{cases} \mathbf{v}^* & J_{k,l}(\mathbf{v}^*) > w_B h_B C \\ 0 & otherwise \end{cases}. \quad (4)$$

Here, $\tilde{\mathbf{v}}$ is the final $B_{k,l}$ block displacement estimate and C can be empirically optimized for a given type of image sequence.

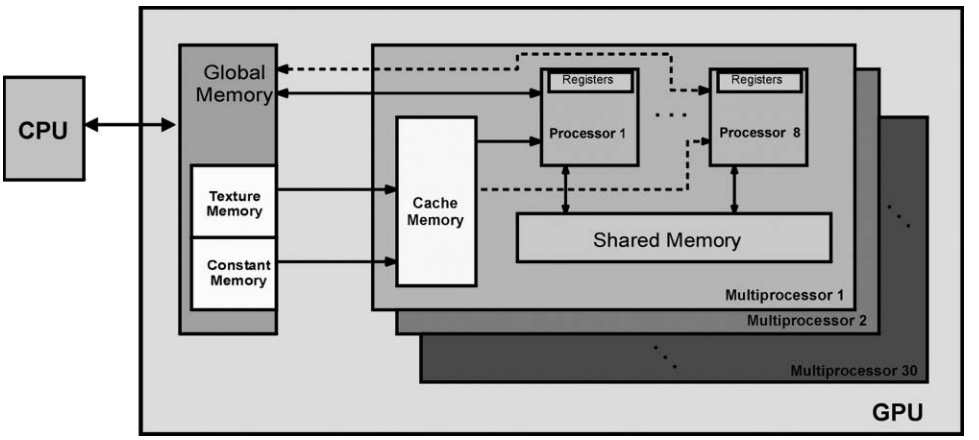


Fig. 3 Schematic of NVIDIA C1060 Tesla GPU card; memory and processors organization.

Note that for the BMA, a proper selection of the number of blocks and the search window size is needed. This is usually empirically adjusted, visually or quantitatively, until satisfying results are obtained. Alternatively, for the motion estimation of a known object the search window size can be estimated using expected object velocity.

2.2 *Computational Complexity of the Block-Matching Algorithm*

The block-matching algorithm's (BMA) computational complexity directly increases with the size of the search window and how the search is performed. The full search (FS) used in this work is a method that gives the best results and lowest matching error but is also the most computationally consuming implementation. Other searches such as the cross- and three-step search[19] can considerably reduce computational time, but potentially provide less accurate motion estimation and are slightly more complicated to implement. In addition, the FS method is not image content sensitive but only image size sensitive, whereas the cross- and three-step search are sensitive to both.

The computational complexity (CC) of the BMA with FS is

$$CC \sim O\left(K \cdot L \cdot w_B h_B \cdot \frac{w_S}{\Delta v}\frac{h_S}{\Delta v}\right) \tag{5}$$

for an image with $K \cdot L$ square blocks of w_B, h_B pixels in width and height and with $\frac{w_S}{\Delta v}\frac{h_S}{\Delta v} = M \times N$ possible candidate displacement vectors, where w_S, h_S defines the search window size, $\Delta v$is the grid step size, and $M \times N$ is the total number of possible displacement vectors per block. Note that $K \cdot w_B \times L \cdot h_B = I \times J$, so that CC can also be expressed as $\sim O\,(I \cdot J \cdot M \cdot N)$. Furthermore, the CC of the methods considerably increases when the displacement candidates are considered to be nonintegers due to the needed interpolation of the image values.

2.3 *Compute Unified Device Architecture Capable Graphical Processing Unit*

The CUDA computing engine from manufacturer NVIDIA exposes powerful GPU hardware to C, C++, Fortran and other programming interfaces.[10,11] GPUs are capable of simultaneously executing a high number of threads. Furthermore, GPUs have specific hardware for floating point arithmetic, 2D and 3D matrix cached access.[11] To a programmer, a CUDA capable card is a collection of multiprocessors (30 for Tesla C1060) where each multiprocessor has a number of processors (8 for Tesla C1060), see Fig. 3. Each multiprocessor has its own fast shared memory (16KB for the C1060) that is common to all the processors within it. In addition, every processor has its own fast memory registers (16384 for the C1060). Every multiprocessor shares the GPU card's global memory (4GB for the C1060) that includes texture and constant memory. In addition, each processor within multiprocessor performs cached access to texture and constant memory. The use of cache reduces the average memory access time since the cache is a smaller and faster memory, which stores copies of the data from the most frequently used memory locations. In addition, by using the attached texture hardware to cached memory, one can perform linear interpolation (1D, 2D, and 3D) – when this memory is accessed on noninteger location – at no added computation time.

From the program developer point of view, the CUDA model allows for a collection of functions (or "kernels" in CUDA-speak) running in parallel threads. The program developer decides the number of threads to be executed in a thread-block, and then the device will schedule the execution of the thread-blocks. This execution will start with joining thread-blocks into a grid followed by scheduling execution of a grid on the collection of multiprocessors. See Fig. 4 for a visual explanation. The developer can define a thread, level 0, and a number of threads in a thread-block, level 1. Further decisions on the execution are left to the GPU hardware, which attempts to group contiguous thread-blocks together, but this is not guaranteed.

In addition, in the CUDA model the threads in thread-blocks are sub-grouped in warps (a group of 32 threads). Each processor in the multiprocessor can sequentially perform the same operation on each thread of a warp, which makes each of them a single instruction multiple data (SIMD) processor. Therefore, for optimal performance, the programmer should minimize thread branching so that all the threads in a warp exactly execute the same instruction to fully utilize the SIMD technology.

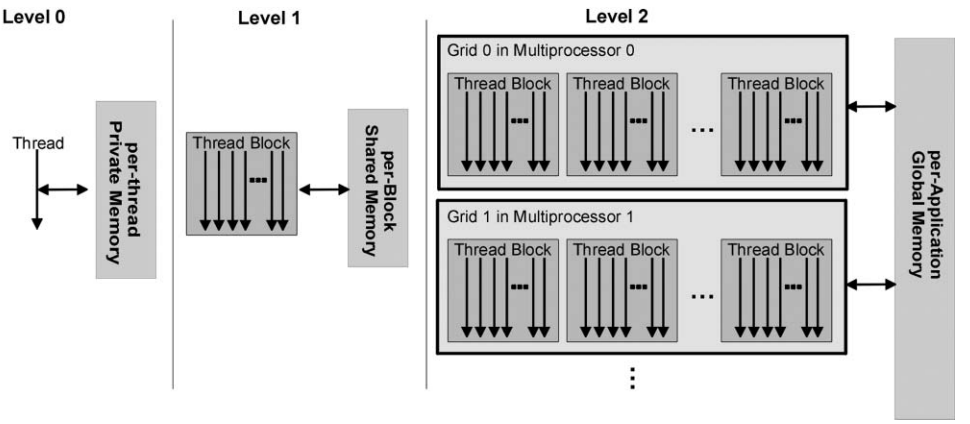


Fig. 4 CUDA hierarchy of threads, thread-blocks, grids, and memory space.

Ideally, for the Tesla C1060 card, if a warp uses the fastest operation, like integer addition, eight threads from the warp are processed in one GPU cycle. In reality, an average GPU processor processes three threads per cycle (this is if a warp uses a floating point operation), making the top performance of 3 × 8 processors × 30 multi-processors × 1.296 GHz - 933 GFLOPS.

Each processor warp scheduler can quickly switch content and put a warp on hold during time consuming operations like global memory fetching (400 to 600 cycles) or cached texture memory fetching (200 to 300 cycles). During this hold time, while the memory is being accessed, it will attempt to execute other warps (up to three additional). For this reason, to fully utilize CUDA capabilities,

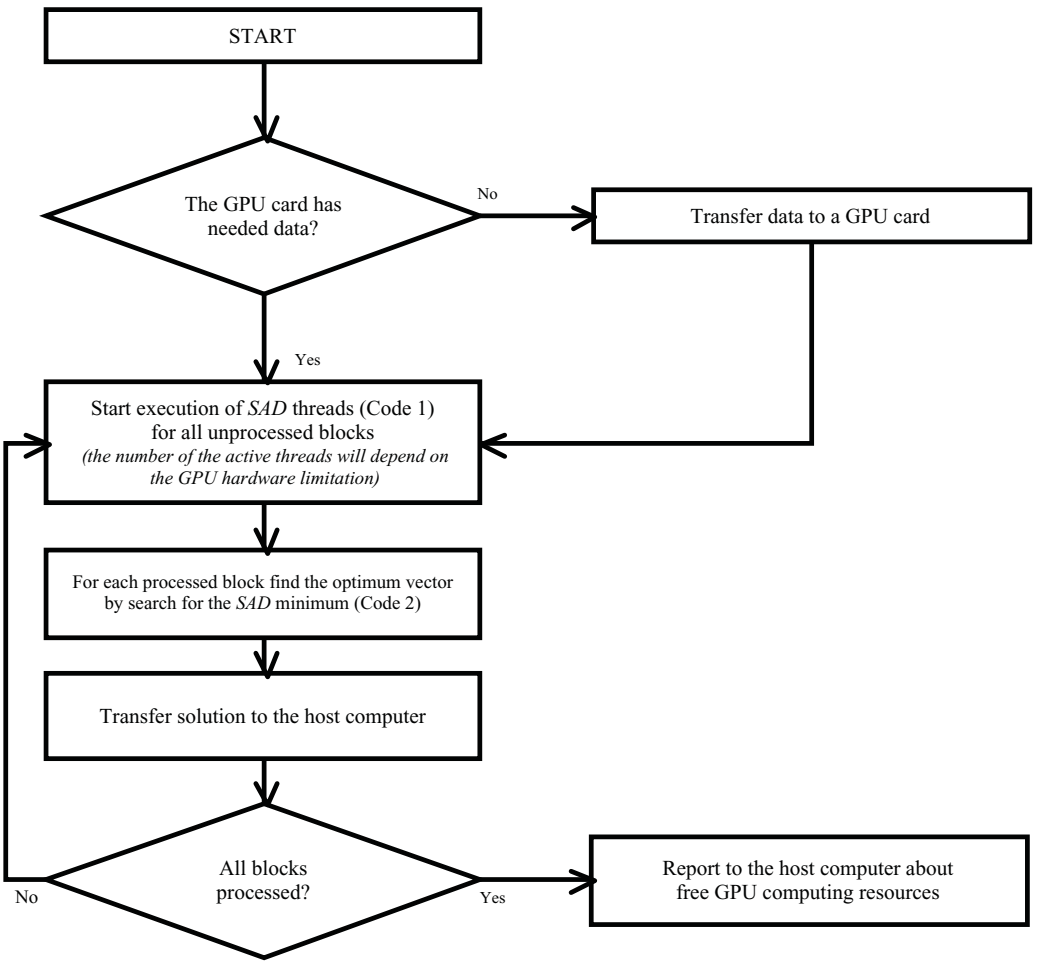


Fig. 5 Flowchart of the proposed BMA algorithm.

it is important to submit a larger number of threads then the number of processors. C1060 can simultaneously execute 240 (30 × 8) warps and have 30720 (30 × 1024) active threads.

3 Algorithm Design and Implementation

3.1 *Kernel Function*

In this work, we implemented a BMA[13] with an FS over all possible candidate vectors on a regular grid. The classical, serial, algorithm is very straightforward: for each block within the reference image, calculate the SAD for every candidate displacement vector and choose the best displacement as the one that minimizes the SAD.

The multicore GPU implementation has two relevant stages.

1. Code 1: Start a thread to work with quadruplet (k, l, m, n), where k and l are image block identifiers and m and n are identifiers of one candidate displacement vectors. Each thread will compute the $J_{k,l}(\mathbf{v}_{m,n})$, defined as SAD, for the $B_{k,l}$ block and the displacement candidate, $\mathbf{v}_{m,n}$, and store it to a global memory. So for each block $B_{k,l}$, we will have a total of $M \times N$ threads computing all possible values of $J_{k,l}(\mathbf{v}_{m,n})$ (see code 1).
2. Code 2: Next, a trivial thread is launched to find the minimum value over $m = 1, 2, \ldots, M$,$n = 1, 2, \ldots, N$, $(M \times N)$, stored values of $J_{k,l}(\mathbf{v}_{m,n})$.

Global memory access is one of the main GPU bottlenecks. To minimize this, in code 1 we use two mechanisms: 1. the reference, $I^t(i, j)$, and target, $I^{t+1}(i, j)$, images are stored in 2D cached texture memory; 2. all other variables are stored in fast register memory associated with the processor, and only one write to global memory is done at the thread end in order to store the calculated value of $J_{k,l}(\mathbf{v}_{m,n})$.

A flowchart, which schematically describes the proposed implementation, is shown in Fig. 5.

The number of threads per thread-block was optimized using the CUDA occupancy calculator that is provided with the software developer kit (SDK) from NVIDIA.[10] From the device code below, it is easy to estimate that each thread will use 10 to 11 registers and zero shared memory. By entering these numbers into the occupancy calculator, we can obtain thread-block sizes of 128, 256, and 512 threads that will provide full (100%) GPU utilization.

3.2 *Multi-Graphical Processing Unit Image Sequence Processing*

In addition to evaluation of a single GPU implementation, we examine two approaches for image sequence processing using multiple GPU cards. Ideally, one would hope for a linear reduction in execution time with the increased number of GPU cards in use. However, due to the overhead in image data transfer and GPU cards scheduling efficiency, this will not be the case. Therefore, we test two possible scenarios:

1. Parallel splitting. In this scenario, schematically shown in Fig. 6(a), the motion estimation threads between two consecutive frames, are grouped into super-blocks and then each super-block is submitted to be processed by a different GPU, all at the same time. Here, since each GPU processes only part of the original image pair, the total processing time is reduced.
2. Serial splitting. In this scenario, each GPU processes one image pair, Fig. 6(b), and since more than one image pair is processed at a time, the total processing time is reduced.

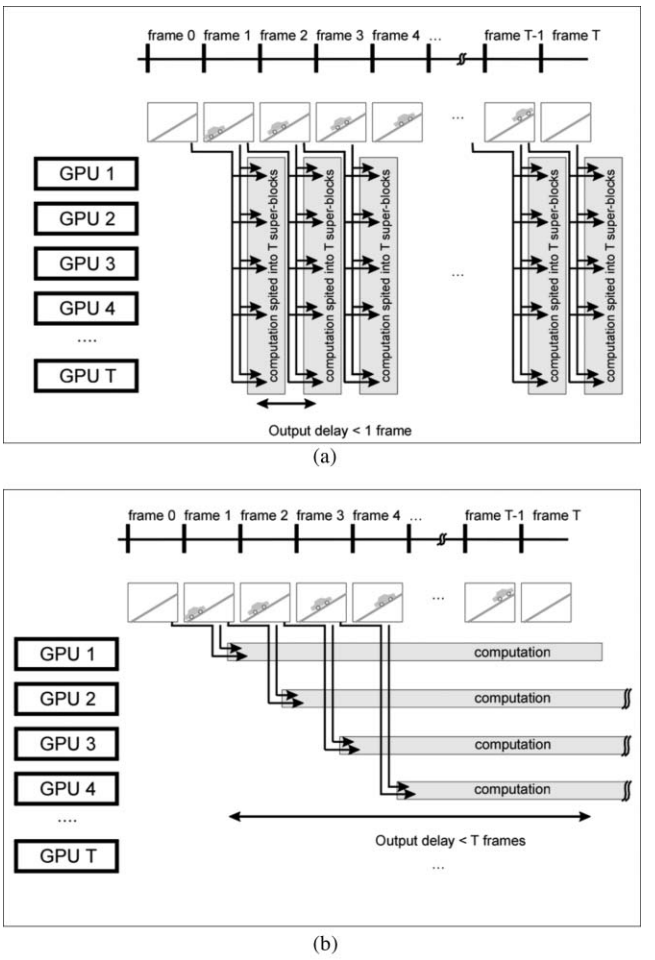


Fig. 6 (a) Parallel and (b) serial splitting of an image sequence.

It is worth noting that in serial splitting the memory requirement to store the reference and target images as well as values of $J_{k,l}(\mathbf{v}_{m,n})$for all possible displacement vectors $\mathbf{v}_{m,n}$ should not exceed GPU memory limits, where in parallel splitting this is not the case. In the experiments to follow, even when using images at a resolution of high definition television, we did not exceed GPU memory limitations.

4 Algorithm Performance

4.1 *Motion Estimation Accuracy Evaluation*

First, we visually and quantitatively compared correctness of the implemented parallel design for the Tesla C1060 GPU in respect to a serial single CPU core design using optimization flags for Xeon E5520 @ 2.27GHz CPU (-O3 and –fno-strict-aliasing). We compared the two methods using several images at a resolution of high definition television (HDTV) (1920 × 1080 pixels) of which a pair is shown in Fig. 7. In this testing, images were split into 400 blocks (20

Code 1: Parallel Kernel

```
//host code
int I, int J;                          // image width and height
int wB, int hB;                        // image-block width and height
int K = I/ wB;                         // number of blocks in i direction
int L = J/ hB;                         // number of image-blocks in j direction
int n_image_Blocks = K*L;              // total number of image-blocks
float dw = 0.5                         // search grid step size
int wS, int hS;                        // search window width and height
int M = wS /dw;                        // number of search grid points in horizontal direction
int N = hS /dw;                        // number of search grid points in vertical direction
int nVectors = M*N;                    // total number of search grid points
int* J_kl;                             //vector containing J_{k,l}(v_{m,n}) values allocated
                                       //in global memory by cudaMallocArray
texture<float,2> Reference_image;      //allocate reference image in global memory as a texture memory
texture<float,2> Target_image;         //allocate target image in global memory as a texture memory

//device code
__global__ void exhaustiveSearchKernel (int* J_kl, int I, int J, int M, int N, int wB, int hB, int n_image_Blocks, int wS, int hS, int nVectors)
{
// allocate variables in the register memory
volatile int idx = blockIdx.x * blockDim.x + threadIdx.x;        // idx = (K*L)* idBlock + idVector;
volatile int id_Block = idx/nVectors + offset;                   // id_Block = L*k + l
volatile int id_Vector = idx%nVectors;                           // id_Vector = M*n+m;
if (id_Block > n_image_Blocks ) return;                          // check if this is the last block
volatile float xB = (id_Block /L)*wB + 0.5f;                     // calculates block location
volatile float yB = (id_Block %L)*hB + 0.5f;                     //
volatile float v1 = ( ( (id_Vector/M)-M/2) * wS/M ) * dw;        // calculates displacement - v(m,n)
volatile float v2 = ( ( (id_Vector %M)-N/2) * hS/N ) * dw;       //

for ( int w = 0; w < wB; ++w )
    for ( int h = 0; h < hB; ++h )
        value += abs( tex2D(Rerence_image,xB+w,yB+h) -
        tex2D(Target_image,xB+w+v1,yB+h+v2) );
        // accessing 2D cached textured memory with interpolation

J_kl[idx] = (int)value;                          // return J_{k,l}(v_{m,n}) value to a global memory}
```

$\times$ 20), each of 96 $\times$ 54 pixels in size, and the search window was set to double the image-block size (192 $\times$ 108 pixels) with a 0.5 pixel (noninteger) displacement step so that for each block we evaluated 82944 candidate vectors. The parameter C described in Eq. (3) was empirically chosen to be four.

An example of estimated motion using the serial algorithm (CPU implementation) is shown in Fig. 8(a) and the output of the parallel algorithm (GPU implementation) can be found in Fig. 8(b).

In the entire test set, no visually significant difference in estimated motion was observed. A quantitatively small difference, on the order of the numerical precision, was found. This can be explained by the difference in numerical precision between the CPU and GPU hardware.[10] At present, the arithmetic operations in the

(a) (b)

Fig. 7 HDTV image of 1920$\times$1080 pixels; (a) reference and (b) target image, respectively.

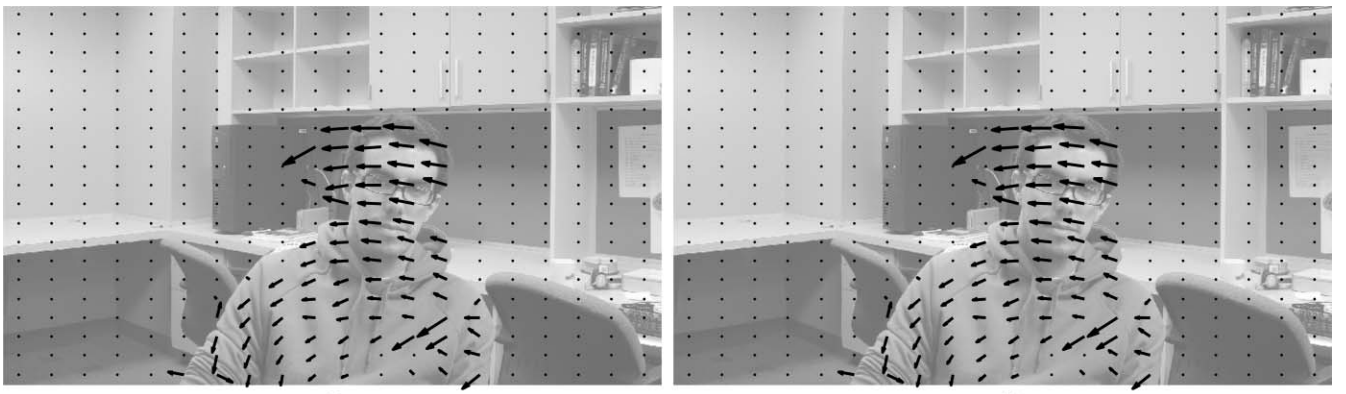


Fig. 8 Motion field estimated using (a) CPU and (b) GPU implementation, respectively.

NVIDIA GPU cards do not follow the floating-point IEEE-74 standard commonly supported by CPUs. New releases of NVIDIA GPUs (like Fermi) will use the IEEE-74 standard.

The GPU execution time for HDTV images was 7.23 s, whereas for CPU implementation it was 8025.00 s or 2 h 13 min 45 s, so the speedup for processing images in HDTV resolution was almost 1110 fold.

4.2 *Multi-Graphical Processing Unit versus CPU Execution Time Comparison*

4.2.1 *Noninteger search grid*

Next, we tested the proposed implementation using image sequences of various sizes with a fixed aspect ratio of 3:2 and 300 images in a sequence. In all experiments, the blocks are set to be 5%, the search window 10% of the image size, and the search step size of 0.5 pixels.

In Fig. 9, we reported the average frames-per-second (fps) achievable for GPU and CPU implementations. First, one can see that the maximum number of fps for CPU implementation is about 4.7 fps at 130 × 74 pixels image, where at the same resolution, one GPU can achieve around 600 fps (out of the range in Fig. 9). At a more common frame rate of 30 fps, one GPU can process images of up to 420 × 280 pixels in size. Next, it is interesting to see the change in image size that GPU implementation in parallel splitting achieves at a 30 fps rate: two GPUs can process 510 × 340 pixel images, three GPUs 549 × 366, and four GPUs 600 × 400 pixel images. In serial splitting, these numbers are even better at a rate of 30 fps: two GPUs can process 516 × 344 pixels, three GPUs 570 × 380, and four GPUs 621 × 414 pixel images.

Next, in Fig. 10, the GPU versus CPU speedup curves are shown. It is interesting to note that for each GPU configuration in use, these curves flatten at some point. We postulate that this is the point at which the number of scheduled threads allowed by the thread scheduler reaches the maximum capacity of ∼30 k active threads with the use of content switching and, therefore, offsets memory fetching and other delays. Two other observations can be made. The sequential splitting (graph on the right) has a faster increase in performance (for images smaller than 800 pixels in horizontal size), but its performance declines after image size reaches 2500 pixels.

In addition, we also tested if the CPU execution time follows the predicted model complexity of $O(I \cdot J \cdot M \cdot N)$, and we find an excellent agreement with a small difference for small images where the data transfer time overhead is more significant than the actual computation time.

4.2.2 *Integer search grid*

For completeness of the presented analysis, and since the CPU does not have dedicated interpolation hardware, we

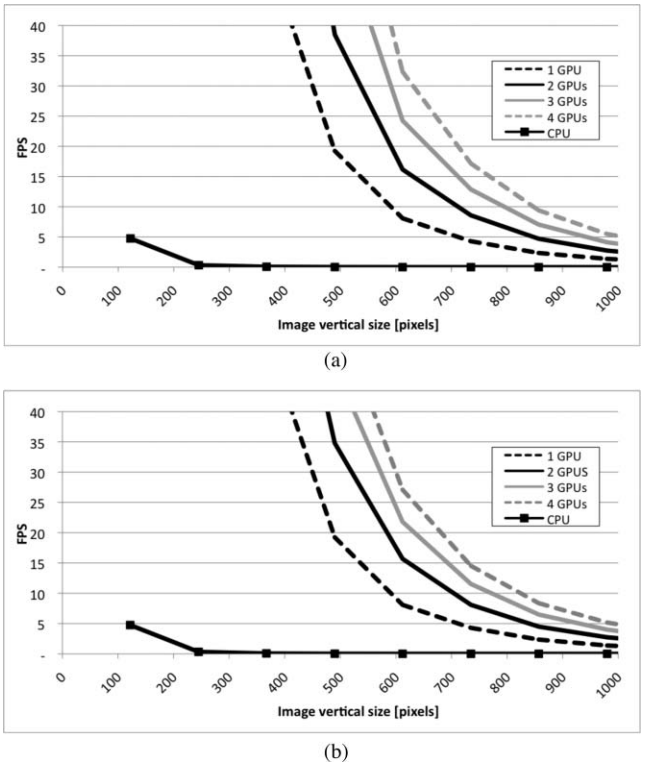
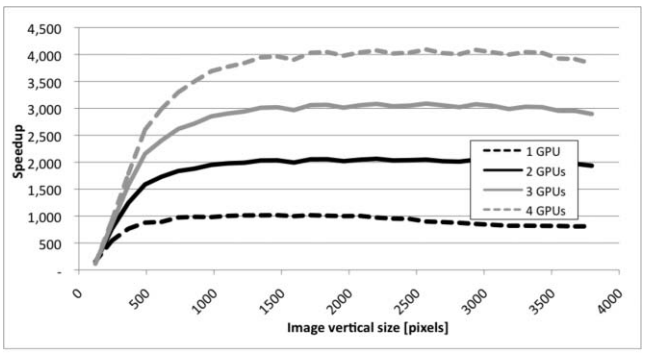
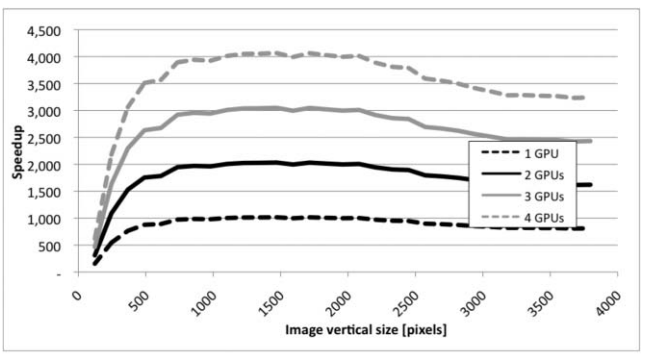


Fig. 9 Maximum achievable GPU implementation frame-rate-per-second (fps) as a function of the frame size for: (a) parallel and (b) sequential splitting.



(a)



(b)

Fig. 10 Achieved GPU implementation speedup, in comparison to CPU, as a function of the frame size for: (a) parallel and (b) sequential splitting.

also performed a test using an integer search grid. The GPU execution time did not change at all due to cache memory interpolation hardware, whereas the CPU computation time was approximately reduced by a factor of five. For this comparison we resized the search window so that the number of vectors in each test case would be the same as for a non-integer grid. This will make blocks be 10% and the search window 20% of the image size.

In all of the test examples, no visually or quantitatively measurable difference was found for the integer search grid. Two speedup curves, one each for integer and noninteger search grid, are shown Fig. 11. One can observe that the omission of the interpolation reduces speedup by a factor of 5 to about 200.

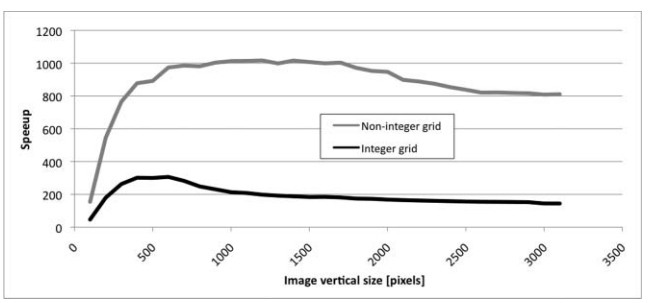


Fig. 11 Achieved GPU implementation speedup, in comparison to CPU, as a function of the frame size for integer and noninteger grid search.

4.3 *Comparison with Other Implementations of Block-Matching Algorithm for Motion Estimation*

4.3.1 *OpenCV implementation*

In these tests we used a video sequence (10 s in duration with 30 frames-per-second) of a moving car while the camera is panning in the opposite direction.[20] This allowed for direct comparison with OpenCV,[21–23] a library developed for Intel CPUs. We consider this to be an interesting comparison since OpenCV was recently used for real-time video processing.[24] The OpenCV implements the Pyramidal Lucas Kanade Optical flow algorithm[25] over a selected number of feature points.

In our implementation we used 400 blocks to match OpenCV, which uses 400 feature points. For a given sequence with 720 × 480 pixels per image, this yields a block size of 36 × 24 pixels (5% of image size). Next, we modified the search window size until satisfying results were obtained in all frames, resulting in a 72 × 48 pixel search window (10% of image size).

For OpenCV it takes 17.7 s to process a 10 s video sequence producing effective 16.9 fps. For the same sequence, serial splitting GPU implementation using two Tesla C1060 cards, it takes less than 10.1 s giving a possibility to achieve 30 fps (15 fps for one GPU card), i.e., real-time data processing. Therefore, the proposed multi-GPU implementation delivers a 1.75× speedup over OpenCV. Moreover, if one inspects Fig. 12, it is evident that the proposed multi-GPU implementation provides estimated motion on a denser grid and of a higher quality.

4.3.2 *H.264/AVC-simplified unsymmetrical multi-hexagon implementation*

Another relevant comparison to assess the performance of the proposed GPU implementation is to use the CPU-based simplified unsymmetrical multi-hexagon search (SmpUHex)[26] implementation used in H.264/AVC standard, which we will denote CPU-H.264/AVC-SmpUHex. We used a highly optimized implementation of H.264/AVC-SmpUHex for Intel CPUs, which can be downloaded at Ref. 16. For fairness of the comparison, we use 16 × 16 pixels blocks and a search area of 32 × 32 pixels in both codes to process the sequence of the car[20] of 720 × 480 pixels.

Using CPU-H.264/AVC-SmpUHex, it takes 33.69 s to compute the motion field for a 10-s video sequence effectively producing 8.9 fps, while proposed single-GPU implementation takes 23.88 s effectively producing 12.6 fps. Therefore, if considered as it is, the proposed singe-GPU implementation achieves a 1.41× speedup over CPU-H.264/AVC-SmpUHex. However, SmpUHex, as the name suggests, does not utilize a full grid search, though in most cases it produces results similar to FS. It was also measured that the proposed GPU implementation achieves speedup of 28× over the FS CPU implementation denoted as CPU-H.264/AVC-FS. Note that CPU-H.264/AVC-FS implementation is fully optimized and as such it is about 10× faster than our CPU FS implementation. The H.264/AVC-FS speedup is mainly achieved by pre-calculated block locations (see code 1), where, in our CPU and GPU implementation this is not done, it may be considered in the future along with implementing SmpUHex on GPUs.

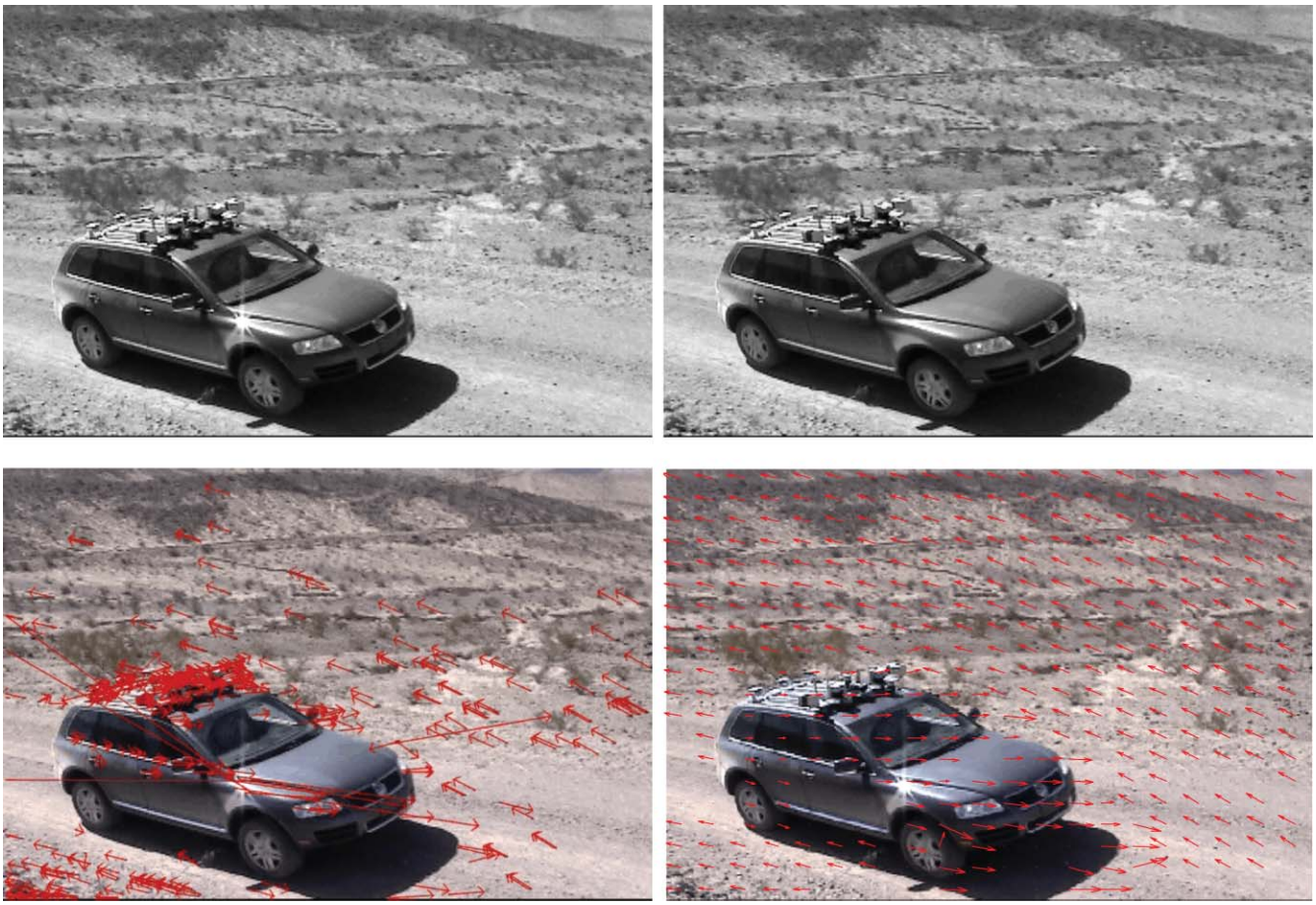


Fig. 12 Comparison with OpenCV implementation of the pyramidal lucas kanade optical flow algorithm.

4.3.3 *Tesla C2070*

Last, we have performed initial testing on a new Tesla C2070 (aka Fermi) (CUDA compute capability: 2.0, released Q2 2010). Interestingly, even though C2070 has a peak performance of 1.03 Tflops (C1060 has 0.933Tflops), we observed decreased speedup of only 665× and 50% GPU occupancy in contrast to C1060s 1000× and 100% occupancy. This indicates that in transferring code from C1060 to C2070, one needs to carefully re-optimze implementation so as to fully utilize C2070 hardware.

4.4 *Demo Program*

For the purpose of testing the proposed GPU implementation by a wider audience, we created an executable demo for Windows OS using OpenCV library. This demo program will perform video capturing and streaming, at 360 × 240 pixels resolution and 30 fps, through the Internet from the users camera to our GPU server. Our server will perform real-time motion estimation and return estimated motion field, which will be displayed at the user's screen. A demo can be downloaded from http://image.mirc.iit.edu/GPUdemo/.

5 Conclusion

In this paper, we presented and evaluated an implementation of the block-matching algorithm for motion estimation with FS using multiple GPU cards. At this time, our implementation is suitable for processing a surveillance video at 720 × 480 pixel resolution at 30 fps (real-time) using two C1060 Tesla GPU cards, outperforming the same CPU implementations by several orders of magnitude.

Further, we performed a comparison of proposed FS GPU implementation with two existing, CPU optimized, implementations which do not utilize FS, namely OpenCV implementation of the Pyramidal Lucas Kanade Optical flow algorithm[25] and SmpUHex,[26] implementation available in H.264/AVC standard. The presented results show a moderate speedup of the proposed GPU implementation, indicating that both CPU methods could be reimplemented on GPUs, and one should expect a significant reduction in computation time. This remains to be explored in future work.

In addition, the work presented here provides a good case example of how to use CUDA technology to increase the performance of video and image processing methods. It is not always easy to implement methods in a highly parallel architecture; for this reason, examples like this can provide some guidance while developing other applications.

Acknowledgments

This work was supported by NIH/NHLBI Grant Nos. HL091017 and HL065425. Massanes was supported through the la Caixa fellowship grant for post-graduate studies, Caixa d'Estalvis i Pensions de Barcelona, Spain. The authors would

like to acknowledge David M. Stavens for his help and generousity in providing data used in the comparison with OpenCV.

References

1. I. Cohen and G. G. Medioni, "Detecting and tracking moving objects for video surveillance," presented at *Conf. on Computer Vision and Pattern Recognition* CVPR, pp. 2319–2325, IEEE Computer Society, Washington, DC (1999).
2. M. H. Chan, Y. B. Yu, and A. G. Constantinides, "Variable size block matching motion compensation with applications to video coding," in *IEE Proc. Communications, Speech and Vision*, **137**(4), 205–212 (1990).
3. N. Cheung, X. Fan, O. C. Au, and M. Kung, "Video coding on multicore graphics processors," *IEEE Signal Processing Magazine*, Vol. 27, no. 2, pp. 79–89, (March 2010).
4. D. Lin, H. Xiaohuang, N. Quang, J. Blackburn, C. Rodrigues, T. Huang, M. N. Do, S. J Patel, and W. M.-W. Hwu, "The parallelization of video processing," *IEEE Signal Process. Mag.* **26**(6), 103–112 (2009).
5. T. Marin and J. G. Brankov, "Deformable left-ventricle mesh model for motion-compensated filtering in cardiac gated SPECT," *Med. Phys.* **37**(10), 5471–5481 (2010).
6. Y. Deuerling-Zheng, M. Lell, A. Galant, and J. Hornegger, "Motion compensation in digital subtraction angiography using graphics hardware," *Comput. Med. Imaging Graph.* **30**(5), 279–289 (2006).
7. P. Baglietto, M. Maresca, and M. Migliardi, "Parallel implementation of the full search block matching algorithm for motion estimation," in *Proc. IEEE Int. Conf. on Application Specific Array Processors*, pp. 182–192, IEEE, Piscataway, NJ (1995).
8. S. Dutta and W. Wolf, "A flexible parallel architecture adapted to block-matching motion-estimation algorithms," *IEEE Trans. Circuits Syst. Video Technol.* **6**(74), 74–86 (1996).
9. J. Vanne, E. Aho, K. Kuusilinna, and T. D. Hamalainen, "A configurable motion estimation architecture for block-matching algorithms," *IEEE Trans. Circuits Syst. Video Technol.* **19**(4), 466–477 (2009).
10. NVIDIA, wepage http://www.nvidia.com/cuda, February 2010.
11. NVIDIA CUDA, "Cuda programming guide, version 2.3," February 2010.
12. C. Wei-Nien and H. Hsueh-Ming, "H.264/AVC motion estimation implementation on compute unified device architecture (CUDA)," presented at *IEEE Int. Conf. on Multimedia and Expo*, pp. 697–700, 23 June 2008–26 April 2008.
13. A. Bovik, *The Essential Guide to Video Processing* (Academic, New York, 2009).
14. B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.* **17**, 185–203 (1981).
15. N. G. Bourbakis, "Visual target tracking, extraction and recognition from a sequence of image using the LG graph approach," *Int. J. Artificial Intell. Tools* **11**(4), 513–529 (2004).
16. H.264/AVC implementation by the International Telecommunications Union available at http://iphome.hhi.de/suehring/tml/.
17. P. Harish and P. J. Narayanan, "Accelerating large graph algorithms on the GPU using CUDA" in *HiPC, Lecture Notes in Computer Science* Vol. 4873 (2007).
18. A. Shiraki, N. Takada, M. Niwa, Y. Ichihashi, T. Shimobaba, N. Masuda, and T. Ito, "Simplified electroholographic color reconstruction system using graphics processing unit and liquid crystal display projector," *Opt. Express* **17**, 16038 (2009).
19. A. Murat Tekalp, *Digital video processing*, (Prentice-Hall, Englewood Cliffs, New Jersey, 1995).
20. D. Stavens, webpage http://ai.stanford.edu/~dstavens/.
21. B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proc. of the 1981 DARPA Imaging Understanding Workshop*, pp. 121–130 (1981).
22. OpenCV, webpage http://opencv.willowgarage.com, February 2010.
23. G. Bradski and A. Kaehler, *Learning OpenCV: Computer Vision with the OpenCV Library*, O'Reilly Media, Inc., Cambridge, MA (2008).
24. L. Y.-P. Zhang Lei and X.-F. Zhang, "Research of the real-time detection of traffic flow based on OpenCV," in *Proc. IEEE Int. Conf. on Computer Science and Software Engineering*, pp. 870–873 (2008).
25. J. Y. Bouguet, "Pyramidal implementation of the lucas kanade feature tracker: description of the algorithm," *Intel Corporation Microprocessor Research Labs* (2000).
26. X. Yi, J. Zhang, N. Ling, and W. Shang, "Improved and simplified fast motion estimation for JM," in *Proc. JVT Meeting*, Tech. Report JVT-P021, July 2005.

Francesc Massanes was accepted at the Center for Interdisciplinary Studies (CFIS) at UPC-Barcelona Tech, Spain, in 2005 to pursue simultaneous degrees in computer science and mathematics. In 2008, he was awarded with a scholarship from the Agency for Administration of University and Research to work in the department of Languages and Computer Systems (LSI). In January 2009, he joined the department of Computer Architecture as a student collaborator. In June 2009, he finished a degree in mathematics at UPC-Barcelona Tech. In June 2010, he finished his degree in computer science at UPC-Barcelona Tech with the Final Thesis titled: Emulation of Human Motion Perception. He is now a Master of Science student in electrical engineering at the Illinois Institute of Technology with a fellowship grant for post-graduate studies from "La Caixa," Barcelona, Spain.

Marie Cadennes has been pursuing a degree in electrical engineering at ENSEA, Cergy-Pontosie Cedex, France, since 2007. She is currently a research scholar at the Medical Imaging Research Center at the Illinois Institute of Technology, Chicago.

Jovan G. Brankov received his diploma of electrical engineering from the University of Belgrade, Serbia in 1996. He received MSEE and PhD degrees from the Electrical and Computer Engineering Department of the Illinois Institute of Technology in 1999 and 2002, respectively. Brankov joined the Electrical and Computer Engineering Department at the Illinois Institute of Technology in 2002 as a researcher, was promoted to research assistant professor in 2004, and currently he is assistant professor in the same department. His current research topics include 4D and 5D tomographic image reconstruction methods for medical image sequences, multiple-image radiography (a new phase-sensitive imaging method), and image quality assessment based on a human-observer model. He is author/co-author of more than 90 publications.