GRAYSCALE - EDGE DETECTION Labo Geavanceerde Computertechniek $_{(JLIZNM)}$

Jona Cappelle & Jonas Bolle 17 mei 2020



Sessie Datum: 4 Mei, 2020

Partners: Jona Cappelle

Jonas Bolle

Klas: MELICTEES

Begeleider: Stijn Crul

Inhoudsopgave

1	Inle	eiding	2
2	Pro 2.1	Schleemstelling en analyse performantie Grayscale	2 2 2
	2.2	2.1.2 Tijd op CPU Edge detection 2.2.1 Tijd op GPU 2.2.2 Tijd op CPU	3 4 4 5
3	Bes	luit	7
A	CO	DE GRAYSCALE	8
В	CO	DE EDGE DETECTION	13
${f L}^{rac{1}{2}}$	ijst	van figuren	
	1 2 3 4 5	Opbouw PNG afbeelding Resultaat grayscale	3 3 5
\mathbf{L}^{i}	ijst	van tabellen	
	1	Samenvattende tabel tijden CPU en GPU	7
$\mathbf{L}_{\mathbf{i}}$	ist o	of Listings	
	1	Keuze Blocksize en nBlock	5

1 Inleiding

In dit labo van geavanceerde computertechniek gaan we een grayscale van een afbeelding maken en edge detection toepassen.

Een grayscale operatie zet alle kleurenwaarden om naar zwart-wit waarden met een equivalente luminantie. Een edge detection operatie detecteert grote verschillen tussen naburige pixels.

2 Probleemstelling en analyse performantie

2.1 Grayscale

De bedoeling van het eerste deel van het labo is het omzetten van een kleurenafbeelding in een grayscale afbeelding. Het resultaat hiervan is terug te vinden in figuur 2. Bij de grayscale zijn er verschillende mogelijkheden om deze te implementeren. Men kan het gemiddelde nemen van de RGB waarden en dit gemiddelde toekennen aan elke R, G en B waarde. Er kan ook gewerkt worden met verschillende verhoudingen voor RGB waarden. In dit labo hebben we voor deze eerste optie gekozen.

De ingelezen afbeelding zetten we met 'lodepng' om naar een ééndimensionale array met structuur zoals weergegeven in figuur 1.

image[i]	image[i+1]	image[i+2]	image[i+3]
Red	Green	Blue	Alpha

Figuur 1: Opbouw PNG afbeelding

Hier moeten we er rekening mee houden dat het 'alpha' kanaal, dat de opacity bepaalt, altijd de waarde 255 moet hebben om de afbeelding niet doorzichtig te maken.

De code van grayscale is terug te vinden in deel A op pagina 8 en volgende.

2.1.1 Tijd op GPU

We meten de uitvoertijd op de GPU bij verschillende blocksizes. Hier werd het kopiëren van de data van en naar de GPU niet mee in rekening gebracht. De laagste tijden worden bekomen bij blocksizes met veelvouden van 32. Een grafiek hiervan is terug te vinden in figuur 3.

De tijd die het duurt om de grayscale uit te voeren is:

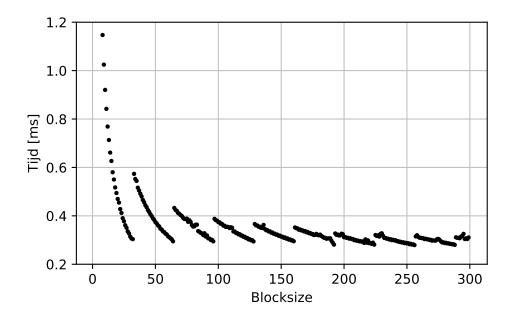
$$t_{GPU}=304\,\mu\mathrm{s}$$





(a) Input (b) Output

Figuur 2: Resultaat grayscale



Figuur 3: Meting tijd op GPU bij verschillende blocksizes

2.1.2 Tijd op CPU

Wanneer we een gray scale implementatie op de CPU schrijven, duurt het veel langer om deze uit te voeren namelijk:

$$t_{CPU} = 656 \,\mu\text{s}$$
 3 of 20

Dit is meer dan dubbel zo lang dan de uitvoering op de GPU (zonder mee kopiëren van de data). Uit vorig labo konden we al besluiten dat het langer zal duren met het mee kopiëren van de data maar dat bij een groot aantal fotos die omgezet moeten worden de GPU met zijn parallelisatie toch veel sneller zal zijn. Om deze reden werd deze meting hier niet opnieuw gedaan.

2.2 Edge detection

In het tweede deel van het labo gaan we edge detection toepassen op een afbeelding. Het resultaat van deze operatie is terug te vinden in figuur 4. Edge detection is een mooi vervolg op grayscale, daar we de grayscale ook nodig hebben om aan edge detection te doen. Voor de edge detection werd gewerkt met de gekende sobel filter. Het werkt op basis van een 3x3 matrix vermenigvuldiging met elke pixel van het beeld. Wanneer er veel verschil is naburige waarden van pixels, zal deze bewerking ofwel een zeer groot, of een zeer klein resultaat opleveren. We maken gebruik van twee 3x3 matrices, één voor de veranderingen in de x-richting te detecteren (G_x) en één voor de veranderingen in de y-richting te detecteren (G_y) . Om deze waarden samen te voegen (G) en negatieve getallen te vermijden, wordt hier ook de wortel van de kwadraten van beide x en y resultaten genomen.

$$\mathbf{G}_x = egin{bmatrix} +1 & 0 & -1 \ +2 & 0 & -2 \ +1 & 0 & -1 \end{bmatrix} * \mathbf{A} \quad ext{and} \quad \mathbf{G}_y = egin{bmatrix} +1 & +2 & +1 \ 0 & 0 & 0 \ -1 & -2 & -1 \end{bmatrix} * \mathbf{A}$$

$$\mathbf{G} = \sqrt{{\mathbf{G}_x}^2 + {\mathbf{G}_y}^2}$$
[1]

De code van edge detection is terug te vinden in deel B op pagina 13 en volgende.

2.2.1 Tijd op GPU

Als we de edge detectie op de GPU uitvoeren, duurt het:

$$t_{GPU} = 336 \, \mu s$$

Hierbij werd rekening gehouden met het optimaal aantal threads per block een veelvoud te kiezen van 32. Zie listing 2.2.1. Er werd gekozen voor een Blocksize van 64 op 16 = 1024. nBlocks volgt hieruit: 64 (4 op 16). Door hier in 2D de threads te organiseren, laat dit ons toe de edge detection op de gpu intuïtiever te coderen. De threads die naast





(a) Input

(b) Output

Figuur 4: Resultaat edge detection GPU

en onder elkaar staan, zullen ook de waarden van de pixels berekenen die naast en onder elkaar staan.

Listing 1: Keuze Blocksize en nBlock

```
// Choose Blocksize & nBlock in 2D
dim3 BLOCKSIZE(64,16);
dim3 nBlocks(ceil(width/64),ceil(height/16));
```

2.2.2 Tijd op CPU

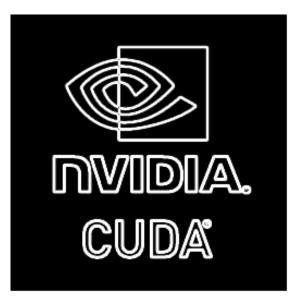
Als we de edge detectie op de CPU uitvoeren, duurt het:

$$t_{CPU} = 3496 \, \mu \mathrm{s}$$

Dit is 10 keer langer dan de bewerking op de GPU! Hoe intensiever de taken worden, hoe groter de tijdswinst als we te werk gaan met een GPU.

Het resultaat van de edge detection op de CPU is terug te vinden in figuur 5.





(a) Input

(b) Output

Figuur 5: Resultaat edge detection CPU

3 Besluit

Grayscale		
	Tijd op GPU	$304\mu s$
	Tijd op CPU	656 μs
Edge detection		
	Tijd op GPU	$336\mu\mathrm{s}$
	Tijd op CPU	3496 μs

Tabel 1: Samenvattende tabel tijden CPU en GPU

Uit tabel 1 kunnen we besluiten dat het uitvoeren van bewerkingen op de GPU veel sneller kan gebeuren dan op een CPU.

Ook hier zien we terugkomen dat 32 threads per block optimaal is voor performantie. Dit komt aangezien CUDA GPU's kernels runnen die blokken van threads gebruiken met een veelvoud van 32. Indien de gebruiker ook dit veelvoud van 32 hanteert, zal er zoveel mogelijk parallelisatie optreden. Wanneer niet voor dit veelvoud gekozen wordt, zullen enkele threads niet gebruikt worden en zal het programma bijgevolg trager uitgevoerd worden.

Een andere conclusie die we kunnen maken is dat de edge detection langer duurt dan de grayscale. Bij de grayscale wordt enkel het gemiddelde van de pixels genomen: twee optellingen en één vermenigvuldiging. Bij de edge detection wordt er vermenigvuldigd met twee maal een 3x3 matrix, wat meer tijd vergt (meer vermenigvuldigingen en optellingen).

Conclusie: Bij zeer kleine workloads heeft het kopiëren van data van en naar de GPU een grote invloed op de uitvoeringstijd. Bij grotere parallele workloads is het veel sneller om een GPU te gebruiken dan een CPU. Dit kunnen we ook zien aan de speedup factor van de grayscale functie en de egde detection functie: de GPU voert de grayscale functie 2x sneller uit dan de CPU, bij de egde detection is dit x10!

A CODE GRAYSCALE

```
// GRAYSCALE -- Jona Cappelle -- Jonas Bolle
  // includes, system
6 #include <stdlib.h>
7 #include <stdio.h>
  #include <string.h>
  #include <math.h>
11 // includes CUDA
12 #include <cuda_runtime.h>
13
14 // includes, project
  #include <helper_cuda.h>
15
 #include <helper_functions.h> // helper functions for SDK examples
17
18 // own includes
 #include "iostream"
  #include "cstdlib"
20
21 #include "time.h"
                       // timing on cpu
22 #include "lodepng.h" // PNG image read
23
 extern "C"
24
25
26
  void decodeOneStep(const char* filename) {
27
         unsigned error;
         unsigned char* image = 0;
29
         unsigned width, height;
30
31
         error = lodepng_decode32_file(&image, &width, &height,
32

    filename);
         if(error) printf("error %u: %s\n", error,
33

→ lodepng_error_text(error));
34
         /*use image here*/
35
36
         free(image);
37
  }
38
39
```

```
void encodeOneStep(const char* filename, const unsigned char* image,
40
      unsigned width, unsigned height) {
          /*Encode the image*/
41
          unsigned error = lodepng_encode32_file(filename, image, width,
42
           → height);
          /*if there's an error, display it*/
44
          if(error) printf("error %u: %s\n", error,
45
              lodepng_error_text(error));
   }
46
   48
   // KERNEL GRAYSCALE
49
   50
   int BLOCKSIZE;
51
52
   // GPU
   __global__ void grayscale(unsigned char* image, unsigned char*
      grayImage,unsigned width,unsigned height)
   {
55
          int j = (blockIdx.x*blockDim.x + threadIdx.x)*4;
56
57
          if(j < width*height*4)</pre>
          {
59
                  grayImage[j] = (image[j]+image[j+1]+image[j+2])/3;
60
                  grayImage[j+1] = (image[j]+image[j+1]+image[j+2])/3;
61
                  grayImage[j+2] = (image[j]+image[j+1]+image[j+2])/3;
62
                  grayImage[j+3] = 255;
63
          }
64
65
   }
66
67
   // CPU
68
   void grayscale_cpu(unsigned char* image, unsigned width, unsigned
69
      height)
   {
70
          for(int j=0; j < (width*height*4); j+=4)</pre>
71
          {
72
                  image[j] = (image[j]+image[j+1]+image[j+2])/3;
73
                  image[j+1] = (image[j]+image[j+1]+image[j+2])/3;
74
                  image[j+2] = (image[j]+image[j+1]+image[j+2])/3;
75
          }
76
   }
77
```

```
78
   79
   // Program main
80
   81
   int main()
   {
83
          84
          // Load PNG file
85
          86
         float millis = 0;
87
         unsigned char *image_in, *image_out, *image_in_dev,
          → *image_out_dev;
         unsigned width, height, width_dev, height_dev;
89
90
         const char* filename = "test.png";
91
92
         unsigned error;
         unsigned char* image = 0;
95
         error = lodepng_decode32_file(&image, &width, &height,
96
            filename);
          if(error) printf("error %u: %s\n", error,
97
            lodepng_error_text(error));
98
          // allocate arrays on host
99
          image_in = (unsigned char *)malloc(width*height*4 *
100

    sizeof(char));
          image_out = (unsigned char *)malloc(width*height*4 *
101

    sizeof(char));
102
          // File to store measured time data
103
         FILE *f = fopen("data.csv", "w");
104
105
   for (int BLOCKSIZE = 1; BLOCKSIZE < 300; BLOCKSIZE++)</pre>
106
107
          int nBlocks = (width*height*4) / BLOCKSIZE + ((width*height*4) %
108
          \rightarrow BLOCKSIZE == 0 ? 0 : 1);
         printf("nBlocks: %d", nBlocks);
109
110
   //
           StopWatchInterface *timer = 0;
111
112 //
           sdkCreateTimer(&timer);
113 //
           sdkResetTimer(&timer);
  //
           sdkStartTimer(&timer);
114
```

```
115
             grayscale_cpu(image, width, height);
   //
116
117
             sdkStopTimer(&timer);
118
   //
             printf("Tijd: %f\n", sdkGetTimerValue(&timer));
             sdkDeleteTimer(&timer);
   //
120
121
                     // allocate arrays on device
122
           cudaMalloc((void **)&image_in_dev, width*height*4 *
123

    sizeof(char));
           cudaMalloc((void **)&image_out_dev, width*height*4 *
124
               sizeof(char));
125
           cudaMemcpy(image_in_dev, image, width*height*4*sizeof(char),
126
               cudaMemcpyHostToDevice);
           cudaMemcpy(image_out_dev, image_out,
127

→ width*height*4*sizeof(char), cudaMemcpyHostToDevice);
128
             unsigned *width_1 = &width;
   //
129
             cudaMemcpy(width_dev, width_1, sizeof(unsigned),
   //
130
       cudaMemcpyHostToDevice);
             cudaMemcpy(height_dev, &height, sizeof(unsigned),
   //
131
       cudaMemcpyHostToDevice);
132
           // Record time on GPU with cuda events
133
           cudaEvent_t start, stop;
134
           cudaEventCreate(&start);
135
           cudaEventCreate(&stop);
136
137
   138
           cudaEventRecord(start);
139
           grayscale <<< nBlocks, BLOCKSIZE >>> ( image_in_dev,
140
           → image_out_dev, width, height );
           cudaEventRecord(stop);
141
   142
143
           cudaMemcpy(image_in, image_in_dev, width*height*4*sizeof(char),
144
           \rightarrow cudaMemcpyDeviceToHost);
           cudaMemcpy(image_out, image_out_dev,
145

→ width*height*4*sizeof(char), cudaMemcpyDeviceToHost);
146
           cudaEventSynchronize(stop);
147
           cudaEventElapsedTime(&millis, start, stop);
148
```

```
149
             printf("Tijd op GPU: %f\n", millis);
150
             fprintf(f, "%d,%f\n", BLOCKSIZE, millis);
151
    }
152
153
             // Close the file
             fclose(f);
155
156
             // Save the result image
157
             const char* output_filename = "output.png";
158
             encodeOneStep(output_filename, image_out, width, height);
159
160
             // Free memory
161
             free(image_in);
162
             free(image_out);
163
164
             cudaFree(image_in_dev);
165
             cudaFree(image_out_dev);
166
167
             printf("Done!");
168
169
             return 0;
170
    }
171
```

B CODE EDGE DETECTION

```
// EDGE DETECTION -- Jona Cappelle -- Jonas Bolle
  // includes, system
  #include <stdlib.h>
6
  #include <stdio.h>
  #include <string.h>
  #include <math.h>
  // includes CUDA
11
  #include <cuda_runtime.h>
12
13
  // includes, project
14
  #include <helper_cuda.h>
15
  #include <helper_functions.h> // helper functions for SDK examples
16
17
  // own includes
18
  #include "iostream"
19
  #include "cstdlib"
20
  #include "time.h"
                        // timing on cpu
21
  #include "lodepng.h" // PNG afbeelding inlezen
22
23
  extern "C"
24
25
  void decodeOneStep(const char* filename) {
26
         unsigned error;
27
         unsigned char* image = 0;
         unsigned width, height;
29
30
         error = lodepng_decode32_file(&image, &width, &height,
31

    filename);

         if(error) printf("error %u: %s\n", error,
32
          → lodepng_error_text(error));
33
         /*use image here*/
34
35
         free(image);
36
  }
37
38
```

```
void encodeOneStep(const char* filename, const unsigned char* image,
     unsigned width, unsigned height) {
          /*Encode the image*/
40
          unsigned error = lodepng_encode32_file(filename, image, width,
41
          → height);
42
          /*if there's an error, display it*/
43
          if(error) printf("error %u: %s\n", error,
44
           → lodepng_error_text(error));
   }
45
   47
   // KERNEL EDGE DETECTION
48
   49
   int BLOCKSIZE;
50
51
   // GPU
52
   __global__ void edge(unsigned char* orig, unsigned char* result,unsigned

    width,unsigned height)

   {
54
55
          int x = (threadIdx.x + blockIdx.x * blockDim.x)*4;
56
      int y = (threadIdx.y + blockIdx.y * blockDim.y);
      float dx, dy;
58
      width=4*width;
59
      if( x > 0 && y > 0 && x < (width-1) && y < (height-1)) {
60
          dx = (-1* orig[(y-1)*width + (x-4)]) + (-2*orig[y*width+(x-4)])
61
           \rightarrow + (-1*orig[(y+1)*width+(x-4)]) +
                   orig[(y-1)*width + (x+4)]) + (2*orig[y*width+(x+4)])
               (
                        orig[(y+1)*width+(x+4)]);

→ + (
          dy = (
                  orig[(y-1)*width + (x-4)]) + (2*orig[(y-1)*width+x])
63

→ + (
                   orig[(y-1)*width+(x+4)]) +
               (-1* orig[(y+1)*width + (x-4)]) + (-2*orig[(y+1)*width+x])
64
               \rightarrow + (-1*orig[(y+1)*width+(x+4)]);
          result[y*width + x] = sqrt((dx*dx) + (dy*dy));
          result[y*width + x + 1] = sqrt((dx*dx) + (dy*dy));
66
          result[y*width + x + 2] = sqrt((dx*dx) + (dy*dy));
67
          result[y*width + x + 3] = 255;
68
      }
69
70
   }
71
72
   // CPU
```

```
void edge_cpu(unsigned char* image,unsigned char* image_out, unsigned
       width, unsigned height)
    {
75
76
             for(int j=0; j < (width*height*4); j+=4)</pre>
77
         {
             image[j] = (image[j]+image[j+1]+image[j+2])/3;
79
             image[j+1] = (image[j]+image[j+1]+image[j+2])/3;
80
             image[j+2] = (image[j]+image[j+1]+image[j+2])/3;
81
        }
82
         int dx,dy,val;
              int sobel_x[3][3] =
             \{ \{ -1, 0, 1 \},
85
               \{ -2, 0, 2 \},
86
               \{-1, 0, 1\};
87
88
         int sobel_y[3][3] =
             \{ \{ -1, -2, -1 \}, 
               { 0, 0, 0 },
91
               { 1, 2, 1 } };
92
93
         for (int x=1; x < (width-1); x++)
94
           {
              for (int y=1; y < (height-1); y++)</pre>
96
                 {
97
98
                      dx = (sobel_x[0][0] * image[width*4 * (y-1) + (x-1)*4])
99
                               + (sobel_x[0][1] * image[width*4 * (y-1) + x*4]
100
                               → ])
                               + (sobel_x[0][2] * image[width*4 * (y-1) +
101
                               \hookrightarrow (x+1)*4])
                               + (sobel_x[1][0] * image[width*4 * y
102
                                   (x-1)*4])
                               + (sobel_x[1][1] * image[width*4 * y
                                                                            + x *4
103
                               → ])
                               + (sobel_x[1][2] * image[width*4 * y
104
                               \rightarrow (x+1)*4])
                               + (sobel_x[2][0] * image[width*4 * (y+1) +
105
                               \rightarrow (x-1)*4])
                               + (sobel_x[2][1] * image[width*4 * (y+1) + x *4
106
                               + (sobel_x[2][2] * image[width*4 * (y+1) +
107
                               \hookrightarrow (x+1)*4]);
```

```
108
                dy = (sobel_y[0][0] * image[width*4 * (y-1) + (x-1)*4])
109
                       + (sobel_y[0][1] * image[width*4 * (y-1) + x *4]
110
                       + (sobel_y[0][2] * image[width*4 * (y-1) +
111
                       \rightarrow (x+1)*4])
                       + (sobel_y[1][0] * image[width*4 *
112
                       \rightarrow (x-1)*4])
                       + (sobel_y[1][1] * image[width*4 * y
113
                       + (sobel_y[1][2] * image[width*4 * y
114
                       \hookrightarrow (x+1)*4])
                       + (sobel_y[2][0] * image[width*4 * (y+1) +
115
                       \hookrightarrow (x-1)*4])
                       + (sobel_y[2][1] * image[width*4 * (y+1) + x *4]
116
                       + (sobel_y[2][2] * image[width*4 * (y+1) +
117
                        \rightarrow (x+1)*4]);
118
                int val = (int)sqrt((dx * dx) + (dy * dy));
119
120
                if(val < 0) val = 0;
121
                if(val > 255) val = 255;
122
123
                image_out[width*4 * (y-1) + (x-1)*4] = val;
124
                image_out[width*4 * (y-1) + (x-1)*4+1] = val;
125
                image_out[width*4 * (y-1) + (x-1)*4+2] = val;
126
                image_out[width*4 * (y-1) + (x-1)*4+3] = 255;
127
             }
128
        }
129
   }
130
131
132
   133
   // Program main
   135
   int main()
136
   {
137
          138
          // Load PNG file
139
          140
          float millis = 0;
141
```

```
unsigned char *image_in, *image_out, *image_in_dev,
142
             → *image_out_dev;
             unsigned width, height, width_dev, height_dev;
143
144
             const char* filename = "test.png";
145
             unsigned error;
147
             unsigned char* image = 0;
148
149
             error = lodepng_decode32_file(&image, &width, &height,
150

    filename);
             if(error) printf("error %u: %s\n", error,
151
                 lodepng_error_text(error));
152
             // allocate arrays on host
153
             image_in = (unsigned char *)malloc(width*height*4 *
154

    sizeof(char));

             image_out = (unsigned char *)malloc(width*height*4 *
155

    sizeof(char));

156
157
             FILE *f = fopen("data.csv", "w");
158
    //
               StopWatchInterface *timer = 0;
160
               sdkCreateTimer(&timer);
    //
161
    //
               sdkResetTimer(&timer);
162
    //
               sdkStartTimer(&timer);
163
164
               edge_cpu(image, image_out, width, height);
    //
165
166
    //
               sdkStopTimer(&timer);
167
    //
               printf("Tijd: %f\n", sdkGetTimerValue(&timer));
168
    //
               sdkDeleteTimer(&timer);
169
170
             // Grayscale on CPU
171
             for(int j=0; j < (width*height*4); j+=4)</pre>
172
             {
173
                     image[j] = (image[j]+image[j+1]+image[j+2])/3;
174
                     image[j+1] = (image[j]+image[j+1]+image[j+2])/3;
175
                     image[j+2] = (image[j]+image[j+1]+image[j+2])/3;
176
             }
177
178
             // Allocate arrays on device
179
```

```
cudaMalloc((void **)&image_in_dev, width*height*4 *
180
               sizeof(char));
            cudaMalloc((void **)&image_out_dev, width*height*4 *
181

    sizeof(char));
182
            cudaMemcpy(image_in_dev, image, width*height*4*sizeof(char),
               cudaMemcpyHostToDevice);
            cudaMemcpy(image_out_dev, image_out,
184
               width*height*4*sizeof(char), cudaMemcpyHostToDevice);
185
            // Record time on GPU with cuda events
            cudaEvent_t start, stop;
187
            cudaEventCreate(&start);
188
            cudaEventCreate(&stop);
189
190
            191
            // Choose Blocksize & nBlock in 2D
192
            dim3 BLOCKSIZE(64,16);
193
            dim3 nBlocks(ceil(width/64),ceil(height/16));
194
            195
196
            cudaEventRecord(start);
197
            edge <<< nBlocks, BLOCKSIZE >>> ( image_in_dev, image_out_dev,
            → width, height );
            cudaEventRecord(stop);
199
200
            cudaMemcpy(image_in, image_in_dev, width*height*4*sizeof(char),
201
            cudaMemcpy(image_out, image_out_dev,
202
              width*height*4*sizeof(char), cudaMemcpyDeviceToHost);
203
            cudaEventSynchronize(stop);
204
            cudaEventElapsedTime(&millis, start, stop);
205
206
           printf("Tijd op GPU: %f\n", millis);
207
208
              fprintf(f, "%d,%f\n", BLOCKSIZE, millis);
209
210
           fclose(f);
211
            const char* output_filename = "output.png";
213
            encodeOneStep(output_filename, image_out, width, height);
214
215
```

```
free(image_in);
216
             free(image_out);
217
218
             cudaFree(image_in_dev);
219
             cudaFree(image_out_dev);
220
             printf("Done!");
222
223
             return 0;
224
    }
225
```

Referenties

[1] "Sobel operator - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Sobel{_}operator