

Universidade de Aveiro

DEPARTAMENTO DE ELECTRÓNICA, TELECOMUNICAÇÕES E INFORMÁTICA

47022- ARQUITECTURA DE COMPUTADORES AVANÇADA

Home group assignment 2

Semi-Global Matching stereo processing using CUDA

8240 - MESTRADO INTEGRADO EM ENGENHARIA DE COMPUTADORES E TELEMÁTICA

António Rafael da Costa Ferreira NMec: 67405 Rodrigo Lopes da Cunha NMec: 67800

Docentes: Nuno Lau e José Luís Azevedo

Janeiro de 2016 2015-2016

Conteúdos

1	Introdução
2	Exercício 1
	2.1 Cuda Kernel da função "determine_costs()" 3
3	Exercício 2
	3.1 Cuda Kernel(s) da função "iterate_direction_dirxpos_dev()
	e das funções correspondestes a outras direcções 5
4	Exercício 3
	4.1 Cuda Kernel da função "inplace_sum_views()" 16
5	Conclusão

1 Introdução

O trabalho proposto para a unidade curricular de Arquitetura de Computadores Avançada foi a implementação em CUDA para o processamento de um Semi-Global Matching.

Este programa tem como objetivo determinar a imagem de disparidade entre duas imagens idênticas mas de posições diferentes, como se de dois olhos se tratasse, uma vista com o olho da esquerda e outra com o olho da direita.

O relatório reflete todas as geometrias de kernel implementadas, formas de pensamento, métodos de como foram implementados os algoritmos, resultados, tutorial para correr o código elaborado, e por último a conclusão deste mesmo trabalho.

2 Exercício 1

2.1 Cuda Kernel da função "determine costs()"

Neste primeiro exercício, era pedido que se desenvolvesse um kernel em CUDA que substituísse a função determine costs().

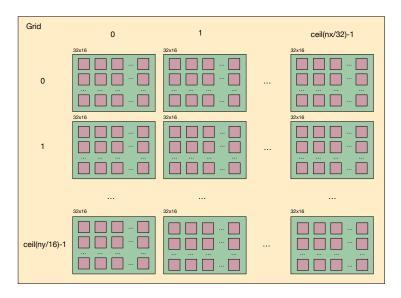


Figura 1: Geometria do Kernel para a função determine costs()

Neste kernel optou-se por uma geometria (Figura 1 constituída por uma grid de tamanho (ceil(nx32) x ceil(ny16)) com blocos de 32 x 16 threads cada. Nesta função, cada thread corresponde a um pixel da imagem, e cada um calcula o valor de custo, sendo este a diferença entre as imagens num determinado pixel.

Este exercicío foi ainda realizado de duas maneira, uma utilizando a *global memory*, e outra onde se coloca as imagens e o valor de COSTS na *texture memory*.

Para a *global memory* utilizou-se o seguinte algoritmo para desenvolver o kernel:

```
__global__ void determine_costs_device(const int *left_image, const int *right_image, int *costs,
const int nx, const int ny, const int disp_range)
{
   int i = blockIdx.x * blockDim.x + threadIdx.x;
   int j = blockIdx.y * blockDim.y + threadIdx.y;

   if (i < nx && j < ny)
   {
```

```
for ( int d = 0; d < disp_range; d++ ) {
    if(i >= d){
        COSTS(i,j,d) = abs( LEFT_IMAGE(i,j) - RIGHT_IMAGE(i-d,j));
    }
}
```

Com esta implementação obtiveram-se os seguintes resultados:

```
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ ./sgm
Host processing time: 5160.187500 (ms)
Device processing time: 5048.691406 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ ./sgm -p 64
Host processing time: 19739.484375 (ms)
Device processing time: 19562.093750 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
```

Figura 2: Resultados obtidos utilizando global memory

TEXTURE MEMORY CODIGO E RESULTADOS

3 Exercício 2

3.1 Cuda Kernel(s) da função "iterate_direction_dirxpos_dev()" e das funções correspondestes a outras direcções

Para este exercicío foram implementadas duas versões para a utilização de *global memory*, sendo a versão 2 (otimizada) utilizada na utilização da *shared memory*.

Global Memory - Versão 1

Nesta versão, foram criadas duas geometrias apenas, sendo que uma diz respeito às iterações nas direções em x, e outra em y, visto que tanto para o lado positivo como para o negativo a geometria era idêntica.

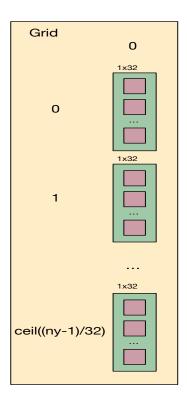


Figura 3: Geometria do Kernel para as funções iterate_direction_dirxpos() e iterate_direction_dirxneg()

Como podemos ver na figura 3, a grid é composta por ceil(ny/32) blocos, cada bloco composto por 32 threads, sendo cada uma responsável pela linha em x onde está inserida para cálculo dos respetivos paths.

Esta operação tem de ser efetuada sequencialmente pois o pixel seguinte depende sempre do anterior, pelo que se recorreu à seguinte implementação para o kernel iterate direction dirxpos() e para o kernel iterate direction dirxneg():

```
__global__ void iterate_direction_dirxpos_dev(const int dirx, const int *left_image,
                                const int * costs , int *accumulated costs ,
                                const int nx, const int ny, const int disp_range ){
        int i = 0;
        int j = blockIdx.y * blockDim.y + threadIdx.y;
        if(j < ny)
           \begin{array}{lll} \textbf{for ( int } d = 0; \ d < disp\_range; \ d++\ ) \ \{\\ ACCUMULATED\_COSTS(0\,,j\,,d) \ += \ COSTS(0\,,j\,,d\,); \end{array}
           for (i = 1; i < nx; i++){
             evaluate\_path\_dev\left(\ \&ACCUMULATED\_COSTS(\,i-dirx\;,j\;,0\,)\;,\right.
                                   &COSTS(i, j, 0),
                                    abs(LEFT IMAGE(i,j)-LEFT IMAGE(i-dirx,j)),
                                   &ACCUMULATED_COSTS(i,j,0), nx, ny, disp_range);
}
const int nx, const int ny, const int disp range )
{
        int i = nx-1;
        int j = blockIdx.y * blockDim.y + threadIdx.y;
        if(j < ny){
           \label{eq:for_disp_range} \mbox{for } (\mbox{ int } d = 0; \mbox{ } d < \mbox{ disp\_range}; \mbox{ } d \!\!\!\!\! + \!\!\!\!\! + ) \mbox{ } \{
               ACCUMULATED COSTS(nx-1,j,d) += COSTS(nx-1,j,d);
          \begin{array}{ll} \mbox{for} \, (\, i \, = \, nx - 2; \ i \, > = \, 0; \ i - - ) \{ \\ \mbox{evaluate\_path\_dev( \&ACCUMULATED\_COSTS($i$-dirx $, j $, 0$) }, \end{array}
                                   &COSTS(i,j,0),
abs(LEFT_IMAGE(i,j)-LEFT_IMAGE(i-dirx,j)),
                                   &ACCUMULATED_COSTS(i,j,0), nx, ny, disp_range);
       }
}
```

No caso da direção ser em y, então seguiu-se o mesmo pensamento que em x, obtendo a seguinte geometria:

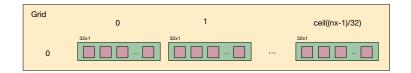


Figura 4: Geometria do Kernel para as funções iterate_direction_dirypos() e iterate_direction_diryneg()

Tal como apresentado na figura, neste caso a geometria é composta por uma grid de tamanho $\operatorname{ceil}(\operatorname{ny}/32)$ blocos, cada um composto por 32 threads, onde cada uma volta a ser responsável pelo cálculo do respetivo caminho de todos os pixeis daquela coluna.

Esta geometria volta a aplicar-se às direções positivas e negativa da mesma maneira tal como em x.

Foi então desenvolvido o seguinte código para os kernels iterate_direction_dirypos() e iterate_direction_diryneg():

```
__global__ void iterate_direction_dirypos_dev(const int diry, const int *left_image,
                       const int* costs , int *accumulated costs ,
                       const int nx, const int ny, const int disp_range )
{
   int i = blockIdx.x * blockDim.x + threadIdx.x;
   \quad \textbf{int} \quad j \; = \; 0 \, ; \quad
   if(i < nx){
       for (j = 1; j < ny; j++){
         evaluate\_path\_dev(\&ACCUMULATED\_COSTS(i,j-diry,0),
                       &ACCUMULATED_COSTS(i, j, 0), nx, ny, disp_range);
}
__global__ void iterate_direction_diryneg_dev(const int diry, const int *left_image, const int * costs , int *accumulated_costs ,
                       const int nx, const int ny, const int disp range )
{
     int i = blockIdx.x * blockDim.x + threadIdx.x;
     int j = ny-1;
     if(i < nx){
       for ( int d = 0; d < disp range; d \mapsto ) {
```

Nesta primeira versão, os resultados obtidos foram os seguintes:

```
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_global$ ./sgm
Host processing time: 5044.548340 (ms)
Device processing time: 4754.518066 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_global$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_global$ ./sgm -p 64
Host processing time: 19286.228516 (ms)
Device processing time: 17583.138672 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_global$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
```

Figura 5: Resultados obtidos utilizando a versão 1 com global memory

Notaram-se algumas melhorias, contudo é possível melhorar o speedup, e para isso recorreu-se a uma segunda versão, desenvolvida com o apoio da leitura do artigo ¹ fornecido pelos professores.

 $^{^1}$ Real-time Stereo Vision: Optimizing Semi-Global Matching, Matthias Michael, Jan Salmen, Johannes Stallkamp, and Marc Schlipsing, IEEE Intelligent Vehicles Symposium pp 1197-1202, 2013

Global Memory - Versão 2

Nesta segunda versão, decidiu-se alterar a geometria do kernel, de forma a que agora cada thread fosse responsável por um único valor de disparidade num path. Para isso a geometria criada para x foi a seguinte:

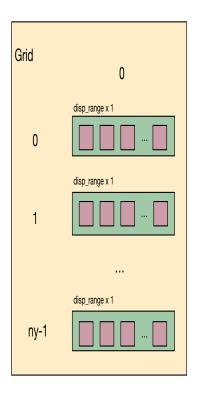


Figura 6: Geometria do Kernel para as funções iterate_direction_dirxpos() e iterate_direction_dirxneg()

A grid passa a ser composta por ny blocos, cada um com um número de threads igual ao disparity range. Passa então a existir um bloco para cada linha em x, composto por threads, onde cada uma corresponde a um valor de disparidade diferente.

A implementação destes dois kernels foi efetuada através do seguinte algoritmo:

```
__syncthreads();
       for (int l = 1; l < nx; l++){
         evaluate_path_dev( &ACCUMULATED_COSTS(l-dirx,j,0),
                             &COSTS(1,j,0),
abs(LEFT_IMAGE(1,j)-LEFT_IMAGE(1-dirx,j)),
                             &ACCUMULATED_COSTS(l, j, 0), nx, ny, disp_range, i);
         \_\_syncthreads();
      }
    }
}
const int nx, const int ny, const int disp_range )
{
       \begin{array}{lll} \textbf{int} & i \ = \ threadIdx.x; \\ \textbf{int} & j \ = \ blockIdx.y \ * \ blockDim.y \ + \ threadIdx.y; \end{array}
       if(i < disp_range \&\& j < ny){
         ACCUMULATED\_COSTS(nx-1,j\ ,i\ )\ +=\ COSTS(nx-1,j\ ,i\ )\ ;
         __syncthreads();
         for (int l = nx-2; l >= 0; l--)
              evaluate\_path\_dev\left(\ \&ACCUMULATED\_COSTS(\,l-dirx\;,j\;,0\,)\;,\right.
                               &COSTS(l,j,0),
                               abs(LEFT_IMAGE(l,j)-LEFT_IMAGE(l-dirx,j)),
                               \label{eq:costs} \& ACCUMULATED\_COSTS(l,j,0)\,,\ nx\,,\ ny\,,\ disp\_range\,,\ i\,)\,;
             __syncthreads();
      }
}
```

No caso da direção ser em y, a geometria utilizada foi a seguinte:

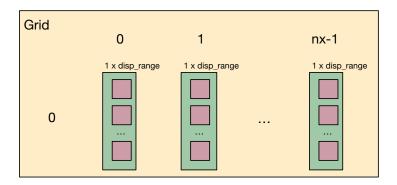


Figura 7: Geometria do Kernel para as funções iterate_direction_dirypos() e iterate_direction_diryneg()

Nesta situação, a grid é composta por nx blocos, cada um um número de threads igual ao disparity range, onde cada thread, tal como em x, é responsável por um valor de disparidade diferente.

A implementação dos kernels correspondestes a esta geometria é a seguinte:

```
__global__ void iterate_direction_dirypos_dev(const int diry, const int *left_image,
                         const int* costs , int *accumulated_costs ,
                         const int nx, const int ny, const int disp_range )
    int i = blockIdx.x * blockDim.x + threadIdx.x;
    int j = threadIdx.y;
    if(j < disp_range \&\& i < nx)
        ACCUMULATED\_COSTS(i, 0, j) += COSTS(i, 0, j);
        \_\_syncthreads();
        for (int l = 1; l < ny; l++){
          evaluate path dev( &ACCUMULATED COSTS(i,l-diry,0),
                          &COSTS(i,1,0),
                          abs(LEFT\_IMAGE(i\ ,l)-LEFT\_IMAGE(i\ ,l-diry\ ))\ ,
                          &ACCUMULATED_COSTS(i, l, 0), nx, ny, disp_range, j);
           __syncthreads();
    }
}
__global__ void iterate_direction_diryneg_dev(const int diry, const int *left_image,
                         const int* costs , int *accumulated_costs ;
                         const int nx, const int ny, const int disp_range )
{
      int i = blockIdx.x * blockDim.x + threadIdx.x;
      int j = threadIdx.y;
      if(j < disp_range \&\& i < nx)
```

Para que a implementação desta segunda versão funcionasse foi necessário recorrer ao comando $__syncthreads()$, de forma a que todas as threads esperassem umas pelas outras quando chegavam ao ponto onde este comando se encontra colocado, garantindo assim que tudo era feito sequencialmente, e posteriormente utilizado de maneira correta quando se recorresse à shared memory. Foi ainda necessário efetuar alterações no código da função $evaluate_path()$ para que agora dentro deste apenas calculasse o valor necessário para aquele valor de disparidade, ficando assim:

```
__device__ void evaluate_path_dev(const int *prior, const int *local,
                              int path_intensity_gradient, int *curr_cost ,
                              const int nx, const int ny, const int disp_range, const int d)
     memcpy(curr_cost, local, sizeof(int)*disp_range);
     int e_smooth = NPP_MAX_16U;
     \label{eq:formula} \textbf{for} \ ( \ \textbf{int} \ \textbf{d}\_\textbf{p} = \ \textbf{0}; \ \textbf{d}\_\textbf{p} < \ \textbf{disp\_range}; \ \textbf{d}\_\textbf{p} \!\!+\!\!\!+ \ ) \ \{
        if'(d_p - d = 0) {
           //No penality
           e_smooth = MMIN(e_smooth, prior[d_p]);
        else\ if\ (abs(d_p - d) == 1) 
           // Small penality
           e smooth = MMIN(e smooth, prior[d p]+PENALTY1);
        } else {
           // Large penality
           e\_smooth =
             MMIN(e_smooth, prior[d_p] +
                           MMAX(PENALTY1,
                           path_intensity_gradient ? PENALTY2/path_intensity_gradient : PENALTY2));
     curr cost[d] += e smooth;
     int min = NPP MAX 16U;
      \begin{array}{lll} \textbf{for} & (& \textbf{int} \ d\_s = 0; \ d\_s < & \text{disp\_range}; \ d\_s + + \ ) & \{ & \textbf{if} \ (& \text{prior} \left[ d\_s \right] < \text{min} \right) & \text{min=prior} \left[ d\_s \right]; \end{array}
```

```
curr_cost [d]-=min;
}
```

Com esta nova implementação, a melhoria no speedup foi brutal, melhorando bastante os resultados:

```
aca0203@nikola:-/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_globalv2$ ./sgm
Host processing time: 5412.062988 (ms)
Device processing time: 597.425415 (ms)
aca0203@nikola:-/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_globalv2$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
aca0203@nikola:-/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_globalv2$ ./sgm -p 64
Host processing time: 20342.712891 (ms)
Device processing time: 1224.891968 (ms)
aca0203@nikola:-/acomputadoresavancada/Trabalho 2/aca_sgm/ex2_p2_67405_67800_globalv2$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
```

Figura 8: Resultados obtidos utilizando a versão 2 com global memory

Para melhorar ainda mais estes resultados passou-se a utilizar a *shared memory* para o cálculo do path.

Shared Memory

Visto que as threads na versão 2 eram executadas em paralelo, e que existiam valores que todas partilhavam e necessitavam umas das outras para o cálculo dos mínimos, então decidiu-se que seria mais produtivo que as pesquisas fossem feitas na *shared memory*. Para isso, foi necessário criar um array de *shared memory* com shared memory size igual a disparity_range*sizeof(int), onde cada índice do array corresponde a um valor de disparidade.

A maioria das alterações foi feita na função evaluate_path(), visto ser nesta que se efetuam todas as pesquisas necessárias para determinar o current cost e o minimo. Com isto a função ficou da seguinte forma:

```
__device__ void evaluate_path_dev(const int *prior, const int *local,
                          \mathbf{int} \hspace{0.2cm} \mathtt{path\_intensity\_gradient} \hspace{0.1cm}, \hspace{0.1cm} \mathbf{int} \hspace{0.2cm} \ast \mathtt{curr\_cost} \hspace{0.2cm},
                          const int nx, const int ny, const int disp_range, const int d, int shmem[])
     memcpy(curr cost, local, sizeof(int)*disp range);
     int e smooth = NPP MAX 16U;
     for ( int d_p = 0; d_p < disp_range; d_p ++ ) {
       if \ (\ d_p - d == 0\ )\ \{
          //No penality
          e smooth = MMIN(e smooth, shmem[d p]);
       else if (abs(d_p - d) == 1) 
          // Small penali\overline{t}y
          {\tt e\_smooth} \ = \ MMIN(\, {\tt e\_smooth} \, , shmem \, [\, {\tt d\_p}] + PENALTY1) \, ;
       } else {
          // Large penality
          e \quad smooth =
            \overline{\text{MMIN}}(e\_\text{smooth}, \text{shmem}[d\_p] +
                       MMAX(PENALTY1,
                        path intensity gradient ? PENALTY2/path intensity gradient : PENALTY2));
     curr_cost[d] += e_smooth;
     int min = NPP MAX 16U;
     for ( int d s = 0; d s < disp range; d s++ ) {
       if (shmem [d_s]<min) min=shmem [d_s];
     curr_cost[d]-=min;
     __syncthreads();
    shmem[d] = curr cost[d];
}
```

Aqui foi também necessário colocar outra vez o comando __syncthreads(), para que todas as threads apenas escrevessem na memória partilhada quando todas tivessem calculado o mínimo valor nesta.

Mais uma vez com esta nova implementação, os resultados voltaram a melhorar muito em relação à memória global:

```
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ ./sgm
Host processing time: 5176.508789 (ms)
Device processing time: 414.550415 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ ./sgm -p 64
Host processing time: 20721.224609 (ms)
Device processing time: 1049.743286 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
```

Figura 9: Resultados obtidos utilizando shared memory

4 Exercício 3

4.1 Cuda Kernel da função "inplace sum views()"

A função *inplace_sum_views*, tem como objectivo a soma de pixeis de duas imagens. Para esta a geometria utilizada foi idêntica à da determinação de custos (Exercício 1), contudo possui uma pequena alteração, pois o número de colunas é agora correspondente a ceil((nx*disp_range)/32):

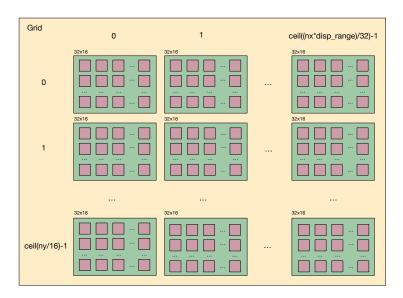


Figura 10: Geometria do Kernel para a função inplace_sum_views()

Para a implementação deste kernel criou-se o seguinte algoritmo:

Posteriormente a esta implementação obtiveram-se os seguintes resultados com a utilização do exercício 2 em memória global e em memória partilhada:

```
aca02030nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex3_p2_67405_67800_globalv2$ ./sgm
Host processing time: 5133.879395 (ms)
Device processing time: 494.530609 (ms)
aca02030nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex3_p2_67405_67800_globalv2$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
inages are identical
aca02030nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex3_p2_67405_67800_globalv2$ ./sgm -p 64
Host processing time: 19267.109516 (ms)
Device processing time: 1923.468994 (ms)
aca02030nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm/ex3_p2_67405_67800_globalv2$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
inages are identical
```

Figura 11: Resultados obtidos utilizando global memory

```
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ ./sgm
Host processing time: 5173.232910 (ms)
Device processing time: 317.796356 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ ./sgm -p 64
Host processing time: 20000.353516 (ms)
Device processing time: 862.332642 (ms)
aca0203@nikola:~/acomputadoresavancada/Trabalho 2/aca_sgm$ !./te
./testDiffs h_dbull.pgm d_dbull.pgm
images are identical
```

Figura 12: Resultados obtidos utilizando shared memory

5 Conclusão

Este trabalho foi útil para assentar todos os conhecimentos que se foi obtendo ao longo destes anos, tanto em Arquitectura de Computadores Avançada como em Arquitetura de Computadores I e II. Ambas as entregas foram primeiro planeadas em papel antes de se avançar para a implementação o que resultou em bons resultados a nível de tempo despendido, para ambos as entregas fez-se uso de dois dias de trabalho.

Na primeira entrega houve uma falha na implementação, esqueceu-se de realizar o shift left 2 do program counter para calcular o endereço para onde é suposto saltar a instrução jump.

Já na segunda entrega teve-se os máximos cuidados para que tudo funcionasse perfeitamente.