FPGA acceleration of Markov Random Field TRW-S Inference for Stereo Matching

MEMOCODE 2013 Design Contest

Jungwook Choi and Rob A. Rutenbar
{jchoi67,rutenbar}@illinois.edu

Oct. 19 2013
Outline

• Sequential tree reweighted message passing (TRW-S) for better MRF inference

• FPGA implementation of streaming arch. for high throughput TRW-S inference

• Performance results (+ Video stereo matching DEMO)
MRF inference for Stereo Matching

Maximum a posteriori (MAP)

\[
\arg\max_x P(x|y) = \arg\max_x P(y|x)P(x)
\]

Energy minimization on Markov random fields (MRF)

\[
\arg\min_x \text{Energy}(x)
\]

Label assignments
Observations

Posterior
Likelihood
Prior

3D depth map based on per-pixel likelihood of depth

3D depth map by MRF MAP inference

Data cost
Prior
Smoothness cost

Likelihood
Data cost
Prior
Smoothness cost
Loopy Belief Propagation

• “Based on what I hear, what label should I choose?”
  – Each node propagates “Belief” to neighbors via “message”
  – So called, “Message Passing” algorithm

• BP on **Tree**: Guaranteed to find optimal labeling
• BP on **Loopy graph**: No guarantee... (sometimes bad)
Better Alternative: Tree-Reweighted Message Passing

- Idea: decompose a loopy graph to a set of trees

Energy is the weighted sum of tree energy

\[ \text{Energy}(x) = \sum_{T \in \text{trees}} \omega_T \cdot \text{Energy}_T(x_T) \]

- Decompose

\[ d_p = \omega_{T3} \cdot d_{T3[p]} + \omega_{T4} \cdot d_{T4[p]} \]

New goal: maximize this lower bound

Equality \(\rightarrow\) optimum energy!

\[ \min_{x} \text{Energy}(x) \leq \sum_{T \in \text{trees}} \omega_T \cdot (\min_{x_T} \text{Energy}_T(x_T)) \]

Goal: minimize overall energy

BP on a tree \(\Rightarrow\) No loop!!
Sequential TRW (TRW-S)

- New goal: max LB by *two step message passing*
  - Averaging belief & message update
- Sequential message passing ⇒ *Convergence property*
  - *Lower bound* is guaranteed not to decrease
  - ⇒ More chance to find the optimum energy!!

- Averaging belief
\[ \hat{\theta}_p(x_p) = \gamma_{pq} \cdot \left\{ d_p(x_p) + \sum_{s \in Nb(p)} M_{sp}(x_p) \right\} \]

- Message update
\[ M_{pq}(x_q) = \min_{x_p} \{(\hat{\theta}_p(x_p) - M_{qp}(x_p)) + V_{pq}(x_p, x_q)\} \]

- Challenge: *parallelize* “sequential message passing”
Comparison: BP-M and TRW-S

- Benchmark: Flower stereo images* (360x262x16 label)
  - Run message passing for 80 iterations
Streaming TRW-S HW Architecture

- Key: *diagonal ordering* of message passing for *parallelism*
- Decoupled, streaming arch.
- Launch/retire 1 pixel/clock
  - Complete label-set likelihood updates for all labels
- Deep pixel-proc pipeline
  - 14 stages deep
  - So: 14 pixels “in flight” / clock
Hybrid CPU+FPGA Platform

- Our platform: Convey HC-1
  - CPU-FPGA cache-coherent virtual memory system
  - Max memory BW: 1Kbit/cycle(\(~20\text{GB/sec}\)/FPGA (runs @150MHz)
  - Non-blocking FPGA function call

* Image from Convey computer
Stereo Matching Running on CPU+FPGA Platform

RD_Datacost
MRF_INFER
GET_DEPTH
WR_DEPTH

Stages:
- **RD_Datacost**
- **MRF_INFER**
- **GET_DEPTH**
- **WR_DEPTH**

**Host CPU**

**FPGA cores**

1024 b/cy

**FPGA Memory (DRAM)**

**Depth Map**

[Data cost : Messages]
Performance Results

- TRWS vs. BP: Min energy

TRW-S achieves lower minimum energy in both cases.
Performance Results

- TRWS vs. BP: Number of errors

**Tsukuba**: TRW-S takes two iterations to achieve lower #Error

**Judging**: TRW-S achieves comparable #Error, despite limit of MRF model
Performance Results

- Adjusted run time
  - Acc_BP: accuracy of reference BP, 40 iter.
  - Acc_TRWS: accuracy of TRW-S, 2 iter.

<table>
<thead>
<tr>
<th></th>
<th>Run time</th>
<th>Acc_TRWS</th>
<th>Acc_BP</th>
<th>Adjusted run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tsukuba</td>
<td>6.40 msec</td>
<td>82.1%</td>
<td>79.8%</td>
<td>6.40 msec</td>
</tr>
<tr>
<td>Judging</td>
<td>6.40 msec</td>
<td>84.3%</td>
<td>86.5%</td>
<td>7.43 msec</td>
</tr>
</tbody>
</table>
Performance Results

TRWS energy minimization over time (Tsukuba)

- It takes ~3.2msec to process 1 iteration
- Memory bandwidth utilization:
  - Data size: 64b x16(node data) x110,592(#nodes/iter) x2(fw/bw) x2(ld/st) = 56.6MB
  - Memory bandwidth: 1Kbit/cycle = 19.2 GB / sec
  - Memory bandwidth utilization: 56.6MB / (19.2GB/sec x 3.2msec) = 92.2%

⇒ Speed is bounded by memory bandwidth
Performance Result: Video Demo

- Function level pipelining + frame level parallelization
  - 3-frame parallel stereo matching, 80 iterations for each inference
  - Speed: Flower 12.3 frame/sec
Summary

- **Sequential tree reweighted message passing** (TRW-S) has been implemented in hybrid FPGA-CPU platform for high speed stereo matching.

- TRW-S shows **faster convergence** to lower minimum energy than belief propagation (BP).

- **Pipelined streaming architecture** is exploited for high throughput message passing.

<table>
<thead>
<tr>
<th></th>
<th>Run time</th>
<th>Acc_TRWS</th>
<th>Adjusted run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tsukuba</td>
<td>6.40 msec</td>
<td>82.1%</td>
<td>6.40 msec</td>
</tr>
<tr>
<td>Judging</td>
<td>6.40 msec</td>
<td>84.3%</td>
<td>7.43 msec</td>
</tr>
</tbody>
</table>
Thank You
## Performance Result

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tsukuba</strong></td>
<td>NVIDIA GeForce 7900 GTX</td>
<td>NVIDIA GeForce 8800 GTS</td>
<td>NVIDIA GeForce GTX 260</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>GPU</strong></td>
<td>(4 scales)</td>
<td>(B, T₁, T₀)</td>
<td>(3 scales)</td>
<td>T₀ = 5</td>
</tr>
<tr>
<td><strong># Iteration</strong></td>
<td>(5,5,10,2)</td>
<td>(12, 20, 5)</td>
<td>(9,6,2)</td>
<td></td>
</tr>
<tr>
<td><strong>Time (msec)</strong></td>
<td>80.8</td>
<td>97.3</td>
<td>61.4</td>
<td>26.10</td>
</tr>
<tr>
<td><strong>Min. Energy</strong></td>
<td>N/A</td>
<td>396,953</td>
<td>N/A</td>
<td>393,434</td>
</tr>
</tbody>
</table>