Stereo GrabCut: Object Extraction for Stereo Images

State Key Laboratory for Novel Software Technology, Nanjing University

juran@smail.nju.edu.com
Stereo image: an old and young media

- A stereo image is made up of two images taken from two slightly different views to simulate the human stereo vision.

- Based on binocular disparity, which is first described by Sir Charles Wheatston in 1838.
Easier to acquire, display and transfer

HTC Evo 3D — the first 3D smartphone
Stereo camera
Stereo digital video camera
3D projector
3D TV
Search results on Flickr
Applications

• 3D movie production

Remap the disparity range for comfortable 3D viewing experience.
Applications

- Image editing
  - Background replacement
  - Stereo photo
  - Color popping
Object extraction for stereo images

- Segmentation should be consistent for both views.
- Stereo images have implicit depth information.

Comparison of GrabCut and Stereo GrabCut
User interface

Step 1. The user drags a compact rectangle around an object to get an initial segmentation.
Step 2. The user scribble with a foreground and background brush to revise the initial result.
Approach

- Establish a correspondence term using stereo matching.
- Pre-estimation of foreground/background from depth map using saliency analysis.
Stereo matching

- Stereo matching
  - Accurate
  - Fast

ELAS[2] is a GCP (ground control points) based algorithm and works in nearly real time.
Consistent graph cut

- Global energy function

\[ E(A) = \sum_{p \in P_l \cup P_r} R_p(A_p) + \lambda_B \sum_{\{p,q\} \in N_B} B_{\{p,q\}} |A_p - A_q| + \lambda_C \sum_{\{p_l,q_r\} \in N_C} C_{\{p_l,q_r\}} |A_{p_l} - A_{q_r}| \]

- Region term
- Boundary term
- Consistency term

- Graph cut model[3][4]

(a) Classical graph cut model

(b) A simple extension to classical graph cut model. The graph is constructed by simply linking the graphs of the left image and the right image at the terminal nodes.

(c) Consistent graph cut model. The graph extends (b) by adding correspondence edges.
**Depth saliency and pre-estimation**

- Basic assumption: salient regions are more likely to differ from background in depth.
- Saliency value definition:
  \[
  S(p_i) = \sum_{p_k \in S_B} |d(p_i) - d(p_k)|, \quad p_i \in S_O
  \]
- Histogram based speed up:
  \[
  S_D(d(p_i)) = \sum_{0 \leq d_k \leq D_{\text{max}}} f_{d_k} |d_k - d(p_i)|, \quad p_i \in S_O
  \]
- Pre-estimation of foreground/background:
  \[
  L_p = \begin{cases} 
  "Object" & \text{if } S_D(p_i) \geq S_f \\
  "Background" & \text{if } S_D(p_i) \leq S_b \\
  "Unsure" & \text{otherwise}
  \end{cases}
  \]
# Depth saliency and pre-estimation

<table>
<thead>
<tr>
<th>Color image</th>
<th>Disparity map</th>
<th>HC[12] (color)</th>
<th>HC[12] (depth)</th>
<th>Our Approach</th>
<th>Pre-estimation by our approach</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Color image" /></td>
<td><img src="image2.png" alt="Disparity map" /></td>
<td><img src="image3.png" alt="HC[12] (color)" /></td>
<td><img src="image4.png" alt="HC[12] (depth)" /></td>
<td><img src="image5.png" alt="Our Approach" /></td>
<td><img src="image6.png" alt="Pre-estimation by our approach" /></td>
</tr>
</tbody>
</table>
FG/BG color modeling

• Gaussian mixture model:

\[ P(c_i | \mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(c_i | \mu_k, \Sigma_k) \]

• Use the pre-estimation of FG/BG as samples and initialize the color model using K-means

• Global energy function

\[ E(A) = \sum_{p \in P_l \cup P_r} R_p(A_p) + \lambda_B \sum_{\{p, q\} \in N_B} B_{\{p, q\}} |A_p - A_q| + \lambda_C \sum_{\{p_{lqr}\} \in N_C} C_{\{p_{lqr}\}} |A_{pl} - A_{qr}| \]

• Region term

\[ R_p(A_p) = \begin{cases} 
- \log P(c_p | \mu_F, \Sigma_F), & \text{if } A_p \in \text{Foreground} \\
- \log P(c_p | \mu_B, \Sigma_B), & \text{if } A_p \in \text{Background} 
\end{cases} \]
Boundary and correspondence term

- Global energy function

\[ E(A) = \sum_{p \in P_l \cup P_r} R_p(A_p) + \lambda_B \sum_{\{p,q\} \in N_B} B_{\{p,q\}} |A_p - A_q| + \lambda_C \sum_{\{p_l,q_r\} \in N_C} C_{\{p_l,q_r\}} |A_{p_l} - A_{q_r}| \]

- Boundary term

\[ B_{\{p,q\}} = \exp \left( - \|c_p - c_q\|_2 \right) \]

- Correspondence term

\[ C_{\{p_l,q_r\}} = \exp \left( - \|c_{p_l} - c_{q_r}\|_2 \right) \]
**Further editing**

- Users scribble on the initial result with a foreground and background brush.
- The color models, consistent graph model and optimal flow are re-computed.
Evaluation

- Dataset
  - www.flickr.com
  - Stereo photos taken in real world
Evaluation
Consistency and accuracy

- **Consistency evaluation**

\[
\frac{|C_l| + |C_r|}{|N_l| + |N_r|}
\]

<table>
<thead>
<tr>
<th>Approach</th>
<th>Input by machine</th>
<th>Input by user</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rect</td>
<td>Stroke</td>
</tr>
<tr>
<td>GrabCut[5]</td>
<td>95.93</td>
<td>-</td>
</tr>
<tr>
<td>Stereo GrabCut</td>
<td>99.44</td>
<td>-</td>
</tr>
</tbody>
</table>

- **Accuracy evaluation**

\[
\frac{N_{gr} \cap N_{rs}}{N_{gr} \cup N_{rs}}
\]

<table>
<thead>
<tr>
<th>Approach</th>
<th>Input by machine</th>
<th>Input by user</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rect</td>
<td>Stroke</td>
</tr>
<tr>
<td>GrabCut[5]</td>
<td>81.52</td>
<td>-</td>
</tr>
<tr>
<td>StereoCut[4]</td>
<td>-</td>
<td>91.49</td>
</tr>
<tr>
<td>Stereo GrabCut</td>
<td>87.36</td>
<td>-</td>
</tr>
</tbody>
</table>
Running time

- Test on a 2.4GHz Intel T8300 CPU with 2GB RAM
- Acceptable for user interaction
Future work

• Apply Stereo GrabCut to further applications
  ➢ Stereo image editing

➢ Supervised segmentation and labeling
References

Thank you!

Q&A