Automatic Image Colorization and Rough Sketch Cleanup by Deep Learning

Hiroshi Ishikawa
Department of Computer Science & Engineering
Waseda University

Joint work with:

Satoshi Iizuka  Edgar Simo-Serra  Kazuma Sasaki
Colorization
Colorization of Black-and-white Pictures
Fully-automatic colorization
Colorization of Old Films

The Lost World (1925)
Related Work

- **Scribble-based** [Levin+ 2004; Yatziv+ 2004; An+ 2009; Xu+ 2013; Endo+ 2016]
  - Specify colors with scribbles
  - Require manual inputs
- **Reference image-based** [Chia+ 2011; Gupta+ 2012]
  - Transfer colors of reference images
  - Require very similar images
Related Work

- Automatic colorization with hand-crafted features [Cheng+ 2015]
  - Uses existing multiple image features
  - Computes chrominance via a shallow neural network
  - Depends on the performance of semantic segmentation
  - Only handles simple outdoor scenes
In our work

- End-to-end network jointly learns **global and local features** for automatic image colorization
  - Fusion layer merges the global and local features
  - Classification labels are exploited for learning

---

Network model

- Two branches: local features and global features
- Composition of four networks
Low-Level Features Network

- Convolutional layers (3x3 kernel)
- No pooling
- Lower resolution with strides > 1
Scaling by Strides

- **Flat-convolution**
  - Stride 1
  - Stride 2

- **Down-convolution**

- **Up-convolution**
  (Not used in colorization.
   Used in the sketch simplification)
Network model

- Network Global Features
- Mid-Level Features Network
- Colorization Network
- Fusion Layer
- Luminance
- Chrominance
- Upsampling

- Low-Level Features Network
- Global Features Network
- Scaling
- Shared weights
Compute a **global** 256-dimensional vector representation of the image.
Mid-Level Features Network

- Extracts mid-level features such as texture

Low-Level Features Network

Mid-Level Features Network

Global Features Network

Colorization Network

Luminance

Scaling

Shared weights

$\frac{H}{8} \times \frac{W}{8}$

512 256

Upsampling

Chrominance
Fusion Layer

- Low-Level Features Network
- Mid-Level Features Network
- Global Features Network
- Fusion Layer
- Colorization Network
- Chrominance
- Upsampling
- Luminance

Scaling

Shared weights
Fusion Layer

- Combines the global and mid-level features
• Compute chrominance from the fused features
• Restore the image to the input resolution
Training

- Mean Squared Error (MSE) and as loss function for chrominance
- Cross-entropy loss for classification network
- Optimization using ADADELTA [Zeiler 2012]
  - Adaptively sets a learning rate
Joint Training with Classification

- Training for classification jointly with the colorization
- Classification network connected to the global features
Dataset

- MIT Places Scene Dataset [Zhou+ 2014]
- 2.3M training images with 205 scene labels
  - 256 × 256 pixels

Abbey
Airport terminal
Aquarium
Baseball field
Dining room
Forest road
Gas station
Gift shop
Colorization Results
Computational Time

- Colorize within a few seconds

<table>
<thead>
<tr>
<th>Image Size</th>
<th>Pixels</th>
<th>CPU (s)</th>
<th>GPU (s)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>224 × 224†</td>
<td>50,176</td>
<td>0.399</td>
<td>0.080</td>
<td>5.0×</td>
</tr>
<tr>
<td>512 × 512</td>
<td>262,144</td>
<td>1.676</td>
<td>0.339</td>
<td>4.9×</td>
</tr>
<tr>
<td>1024 × 1024</td>
<td>1,048,576</td>
<td>5.629</td>
<td>1.084</td>
<td>5.2×</td>
</tr>
<tr>
<td>2048 × 2048</td>
<td>4,194,304</td>
<td>20.116</td>
<td>4.218</td>
<td>4.8×</td>
</tr>
</tbody>
</table>
Colorization of MIT Places Dataset
Comparisons

Input | [Cheng+ 2015] | Ours (w/o global features) | Ours (w/ global features)
Effectiveness of Global Features

Input  w/o global features  w/ global features
Colorization of Historical Photographs

Mount Moran, 1941
Scott's Run, 1937
Youngsters, 1912
Burns Basement, 1910
Colorization of Historical Photographs

California National Park, 1936
Homes, 1936
Spinners, 1910
Doffer Boys, 1909
Colorization of Historical Photographs

Community Center, 1936

North Dome, 1936

Norris Dam, 1933

Miner, 1937
Limitations

- Saturation tend to be low

- Does not recover **true** colors
Rough Sketch Simplification
Rough Sketch Simplification

Play Speed x300
Video: David Revoy, www.davidrevoy.com
Rough Sketch Simplification

Input: ![Image of rough sketch]

Output: ![Image of simplified sketch]

Rough Sketch Simplification

Rough Target

Rough Target

Rough Target
Related work

- Sketch simplification
  - Progressive Online Modification
    - Igarashi et al. 1997, Bae et al. 2008, Grimm and Joshi 2012, Fišer et al. 2015
  
- Stroke Reduction

- Stroke Grouping

- Vector input

- Vectorization
  - Model Fitting (Bezier, ...)
  - Gradient-based approaches
  - Require fairly clean input sketches

Noris et al. 2015
Network model

- 23 convolutional layers
- Output has the same resolution as the input
Learning

- Trained from scratch (about 3 weeks)
- Using 424x424px patches
- Weighted Mean Square Error loss
- Batch Normalization [Ioffe and Szegedy 2015] is critical
- Optimized with ADADELTA [Zeiler 2012]
Sketch dataset

- 68 pairs of rough and target sketches
- 5 illustrators

Sketch dataset

Extracted patches
Inverse Dataset Creation

- Data quality is critical
- Creating target sketches from rough sketches has misalignments
- Creating rough sketches from target sketches properly aligns

Standard

Inverse Creation
Data augmentation

- 68 pairs is insufficient
- Scaling, random cropping, flipping, and rotation
- Additional augmentation: tone, slur, and noise
Sketch Simplification Results
Sketch Simplification Experimental Environment

- Intel(R) Core i7-5960X CPU @ 3.00GHz
- NVIDIA GeForce TITAN X GPU
- Torch framework

- About 3 weeks for training
- Simplification time

<table>
<thead>
<tr>
<th>Image Size</th>
<th>Pixels</th>
<th>CPU (s)</th>
<th>GPU (s)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>320 × 320</td>
<td>102,400</td>
<td>2.014</td>
<td>0.047</td>
<td>42.9×</td>
</tr>
<tr>
<td>640 × 640</td>
<td>409,600</td>
<td>7.533</td>
<td>0.159</td>
<td>47.4×</td>
</tr>
<tr>
<td>1024 × 1024</td>
<td>1,048,576</td>
<td>19.463</td>
<td>0.397</td>
<td>49.0×</td>
</tr>
</tbody>
</table>
Comparison

Input  Adobe Live Trace  Potrace  Ours
Comparison

Input  Adobe Live Trace  Potrace  Ours
Comparison

Liu et al. 2015

Ours

(a) Fairy
(b) Mouse
(c) Duck
(d) Car

Vector Input
Raster Input
Results
Results
Results
Conclusion

- Automatic Image colorization
  - Fuse global and local information
  - Joint training of colorization and classification

- Automatic Rough Sketch Simplification
  - Convolutional networks are suited to image processing
  - Proper data is crucial for training