Learning Robust Color Name Models from Web Images

Boris Schauerte, Rainer Stiefelhagen
Computer Vision for Human Computer Interaction, Karlsruhe Institute of Technology, Germany

Motivation: Why learn color terms? What is an “unnatural” mistake?

Why: Reliable color naming and recognition is necessary for natural human-computer/robot interaction, because it is one of the most common attributes used to communicate and reference objects.

Unnatural: Color terms have fuzzy boundaries and thus, e.g., describing the same object as “yellow” or “orange” may be equally appropriate. However, describing the object as “blue” would be a “unnatural” mistake that humans hardly ever make.

Contributions
- State-of-the-art color naming with an emphasis on reducing “unnatural” naming mistakes
- ~14% improvement towards assigning natural, human-like color names
- Combines salient object detection, KLD outlier reduction, and supervised latent Dirichlet allocation

Outlier Reduction
Remove outliers in the training data (see figure at the top, 3rd row)

1) Estimate a simple, initial color model and calculate the Kullback-Leibler divergence (KLD) between each image and the initial model

\[ d_{KLD}(P(z|w), P(w|d)) = \sum_{w} P(w|d) \ln \left( \frac{P(w|d)}{P(z|w)} \right) \]

with \( P(z|w) = \frac{1}{N} \sum_{d} P(w|d). \)

2) Rank the images and select a subset with the lowest KLD ratio

\[ R_{KLD}^d = \frac{d_{KLD}^e}{\min_{z \neq z} d_{KLD}^e} \]

Evaluation

Training Data Set: Google-512
512 images from Google’s image search for each of the 11 basic English color terms

Test Data Set: eBay+
440 segmented eBay images of objects (4 object classes, 10 evaluation images for each of the 11 basic color terms); plus, for each image, the color term labels of 5 human subjects

<table>
<thead>
<tr>
<th></th>
<th>Cars</th>
<th>Pott.</th>
<th>Shoes</th>
<th>Dress</th>
<th>Total</th>
<th>Dist*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>73.63</td>
<td>80.90</td>
<td>91.82</td>
<td>90.00</td>
<td>84.09</td>
<td>0.50</td>
</tr>
<tr>
<td>X² rank PLSA</td>
<td>71.82</td>
<td>83.64</td>
<td>92.73</td>
<td>86.36</td>
<td>83.64</td>
<td>0.73</td>
</tr>
<tr>
<td>Human</td>
<td>92.73</td>
<td>87.82</td>
<td>90.18</td>
<td>91.99</td>
<td>90.64</td>
<td></td>
</tr>
</tbody>
</table>

* Dist: distance between the confusion matrices of the human observers and the trained model

Contact
{schauerte, rainer.stiefelhagen}@kit.edu
Computer Vision for Human Computer Interaction
http://cvhci.anthropomatik.kit.edu