A Spatial-Color Layout Feature for Representing Galaxy Images

Yin Cui* Yongzhou Xiang* Kun Rong* Rogerio S. Feris* Liangliang Cao*
*Columbia University *IBM T. J. Watson Research Center

1. Background and Motivation

Hubble Ultra-Deep Field (HDF)
- 113,000,000 of the entire sky.
- 2700-day of total exposure time.
- More than 10,000 galaxies in the image.
- Over 100 billion galaxies in the entire universe.

Sloan Digital Sky Survey (SDSS) Dataset
- Naiming at exploring the tremendous number of galaxies in the universe.
- Has collected almost 1 million galaxies to date.
- More than 50 million galaxy images in the near future.
- Need an automatic way to deal with galaxy images.

Astronomical Galaxy Images
- Very noisy (dark-current noises) and contain a large area of dark background.
- Commonly used Computer Vision features (e.g., HOG, LBP, Gist, Color-histogram) do not work well.

The Hubble Sequence and Galaxy Evolution
- The most widely used galaxy representation scheme in Astronomy developed by Edwin Hubble in 1926.
- Galaxy evolves from the right side to the left side.
- Young and active stars are bluer, old stars are redder.
- As the stars become old and inactive, the galaxy they formed evolves from irregular to spiral to elliptical.
- An important cue: Local Color Distribution.

2. Spatial-Color Layout Feature (SCLF)

SCL Galaxy Detector
- Scale and rotation invariant for the galaxy.
- Robust to noise and changing of background.

Step I: Convert to grayscale and do edge-preserving bilateral filtering.
Step II: Binary segmentation using Otsu’s method.
Step III: Fill holes based on morphological operations.
Step IV: Find the center and orientation of the biggest connected component.
Step V: Align the galaxy by its center and orientation.

SCL Galaxy Descriptor
- Dark background → color distributions infer shape.
- Concatenated local color-histograms on spatial layout.
- Capture local and global galaxy shape information.

Comparison with other features

<table>
<thead>
<tr>
<th>Color-Hist</th>
<th>Local Shape</th>
<th>Color</th>
<th>Robust to noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
</tbody>
</table>

3. Galaxy Retrieval System

System Overview
- Galaxy Image Collection
- Feature Extraction
- Hashing
- Database
- Query
- One query: 35 ms on a single PC

Hashing
- Use Kernelized Locality Sensitive Hashing (KLSH) (B. Kulis and K. Grauman, 2009)
- Locality-sensitive for kernel functions.
- Sub-linear time in approximation.
- Test on three kernel functions:
  - Chi-Square Kernel, Histogram Intersection Kernel, Jensen-Shannon Divergence Kernel

4. Experiments

Evaluation by Galaxy Zoo Annotations
- Galaxy Zoo is a crowdsourcing project to collect annotations for galaxy images from SDSS dataset.
- Each image has 37 annotations from 30-50 people.

Performance Comparison
- 283,971 galaxy images with annotations from SDSS.
- Top K retrieved images are judged by evaluation SVM.
- Evaluation metric: mean Average Precision at depth K.

5. Future Work

Representing Galaxy by Attributes
- Build attribute detectors for galaxies based on the proposed feature and GalaxyZoo labels.

6. Summary

- A new feature designed for representing galaxy.
- An efficient large-scale galaxy retrieval system.
- An evaluation method based on crowdsourcing data.

7. Acknowledgement

- Thanks to Dr. Rogerio Feris and Dr. Liangliang Cao (last two authors). This work came from our course project in a class taught by them: