Weakly Supervised Learning for Attribute Localization in Outdoor Scenes
Shuo Wang\textsuperscript{1,2}, Jungseock Joo\textsuperscript{2}, Yizhou Wang\textsuperscript{1} and Song Chun Zhu\textsuperscript{2}
\textsuperscript{1}Nat’l Engineering Lab for Video Technology, Peking University; \textsuperscript{2}Center for Vision, Cognition, Learning and Arts, UCLA

Scene configuration: spatial layouts of a scene which are composed by the objects and regions of varying shapes
Scene attributes: described by text, contain the nouns and adjectives, corresponding to semantic meanings of the objects/regions and their characteristics
Hierarchical Space Tiling (HST):
- HST-geo: quantizes the huge scene configuration space
- HST-att: model the noun attribute as an appearance-Or node having a mixture of adjectives
Association matrix: measures the co-occurrence of a local region and an object, e.g., the grassland always appears at the bottom area of an image.
Weakly supervised method: given a collection of natural images associated with attributes in text, where the precise localization of each attributes left unknown, we simultaneously learn the scene configurations and attributes

Input: images + text descriptions
1. Learn HST-geo:
   (a) For each image, do multi-scale segmentation
   (b) Infer the optimal configuration for each image based-on the multi-scale segmentation
   (c) Update HST-geo, then repeat (b) and (c) until convergence
2. Pursue association matrix by non-maximum suppression
3. Jointly inference: for an image and its text description, infer the optimal configuration and attribute localization
4. Repeat 2 & 3 until convergence

Dataset
- 1226 images (256×256) from 12 categories
- 17 noun attributes and 30 noun+adjective attribute pairs
- Ground truth bounding box for evaluation
- \url{http://www.stat.ucla.edu/shuo.wang/SceneAtt.rar}

Experiments
- The association of noun attributes and scene parts
- Attribute localization results and comparison

Learning & Inference

Scene attribute recognition
\begin{tabular}{|c|c|c|c|}
\hline
 & cKernel+SVM & BoW+SPM & HST-geo & HST-att \\
\hline
MAP(\%) & 64.48 & 53.11 & 51.67 & 67.58 \\
\hline
\end{tabular}

Scene attribute localization
\begin{tabular}{|c|c|c|c|}
\hline
 & SW-FS & HST-geo & HST-att \\
\hline
MAP(\%) & 33.88 & 32.55 & 50.22 \\
\hline
\end{tabular}