Multi-Modal Image Retrieval for Complex Queries using Small Codes
Behjat Siddiquie*, Brandyn White*, Abhishek Sharma and Larry S. Davis

Problem and Motivation

Image Retrieval
- **Complex Queries**: Comprise of Objects, Attributes and Relationships.
- **Multi-Modal Queries**: Query can be an image, text or a sketch.
- **Scalability**: Approach should scale up to millions of images.

Approach

Semantic Representation
- Object/Attribute binary masks, with sketches as the primary form of representation.
- Encodes objects, attributes and relationships.
- Captures **Semantic Similarity** as opposed to **Visual Similarity**.

Image2semantic
- Semantic Texton Forests (Shotton et al.)
- Two stage random forest classifier: Utilizes local color/texture, spatial context and Image Level Priors.

Text2semantic
- Represent object configurations as an MRF.
- Sketches approximated by bounding boxes.
- Sample MRF to generate most likely sketches.

Multi-View Hashing
- Semantic Representations from each modality differ.
- Project different modalities (views) of the data to common low-dimensional subspace.
- Partial Least Squares (PLS) learns projection from each modality to the common subspace.
- Given two views \( X, Y \); PLS decomposes them as:
  \[
  X = TP^T + E; \quad Y = UQ^T + F; \quad U = TD + H
  \]
- Iteratively maximizes:
  \[
  \text{cov}(t_i, u_j)^2 = \max_{|w_{xi}|=1, |w_{yi}|=1} \text{cov}(Xw_{xi}, Yw_{yi})^2
  \]
- Map vectors in the subspace to binary codes to reduce memory burden and enable fast matching (Hamming dist.).
- Randomly rotate projection matrix to uniformly distribute information across all dimensions.
  \[
  W_x = W_x R
  \]
- Iterative Quantization (ITQ) to compute optimal rotation matrix that minimizes the quantization error (Gong et al.):
  \[
  Q(H, R) = ||H - XW_x R||^2
  \]

Results

Quantitative Results
- Single-Modality Hashing
- Multi-Modal Hashing

Qualitative Results
- Text Queries
- Sketch Queries
- Image Queries
- Large Scale Retrieval: Flickr (1M images)