Image representations using local image descriptors

Sanparith.Marukatat@nectec.or.th
Image classification

- Food & Restaurant domain
- Unconstrained images
- Manual tags
- Food / Non-food
Current trend

- Deep Learning: CNN
  - As Feature Extractor & Classifier
  - As Feature Extractor
  - Pro: problem-dependent Feature Extractor
- Cons
  - Neural Networks training
  - Requires special hardware
- Another solution: traditional image descriptors + classifier
Image descriptors

- Global descriptors
  - Histogram of colors, Histogram of Local Binary Patterns (LBPs)
  - Gabor filters, GIST
- Local descriptors
  - SIFT, SURF
  - MSER
- Global descriptor using local descriptors
  - Bag of Features (BoF)
  - Spatial Pyramid Matching (SPM)
  - Locality-constrained Linear Coding (LLC)
SIFT & BoF

- Extract SIFT vectors from all images
- 128 dimensions
- Clustering SIFT vectors (k-means)
- 1 cluster = 1 visual word
- Input image
  - Extract SIFT vectors (Dense SIFT)
  - Put each SIFT vector into the closest cluster
  - Histogram of visual words
Some results

- 5-fold cross validation
- GIST + SVM with RBF kernel: 85.57%
- SIFT-BoF (2000)+ SVM with kernels

1. Histogram Intersection kernel: \( k(x, y) = \sum_i \min\{x_i, y_i\} \) 89.69
2. Chi-square kernel: \( k(x, y) = \sum_i \frac{x_i y_i}{(x_i+y_i)} \) 89.68
3. Hellinger kernel: \( k(x, y) = \sqrt{\sum_i x_i y_i} \) 84.27
4. Jensen-Shannon kernel: \( k(x, y) = \frac{1}{2} \sum_i \left( x_i \log_2 \frac{x_i+y_i}{x_i} + y_i \log_2 \frac{x_i+y_i}{y_i} \right) \) 89.03
SPM

- Multi-scale analysis
  - Compute descriptors at different resolutions
  - Compute descriptors at basic resolution but summing at different resolutions

- SPM
  - Dense SIFT
  - Level $l$ splits image into $2^l \times 2^l = 4^l$ areas
  - In each area, compute histogram of $M$ visual words
  - Concatenate into a big histogram with weighting
Sparse SPM

- SPM are often used with SVM using RBF kernel
- SPM: $M=2000$, $L=3 \rightarrow 42,000$ bins but most are 0
- Sparse SPM
  - **Sparse coding** instead of basic cluster assignment
  - **Max-pooling** instead of sum
- Not histogram, used with linear SVM

\[
\begin{align*}
\min_{u_1, \ldots, u_m} & \sum_{i=1}^{M} \|x_i - V u_i\|^2 + \lambda |u_i| \\
\text{subject to } & \|v_k\| \leq 1.
\end{align*}\\
\begin{align*}
\min_{u_1, \ldots, u_m} & \sum_{i=1}^{M} \|x_i - V u_i\|^2 \\
\text{subject to } & \text{Card}(u_i) = 1, |u_i| = 1, u_i \geq 0, \forall i
\end{align*}
\]
LLC

• Locality is more important than sparsity

• Locality = neighbors in the codebook

• LLC problem

\[
\min_C \sum_{i=1}^{M} \|x_i - Bc_i\|^2 + \lambda \|d_i \odot c_i\|^2 \\
\text{subject to } 1^T c_i = 1
\]

\[
d_{im} = \exp \left( \frac{1}{Z \sigma} \|x_i - b_m\|^2 \right) \\
Z = \max_k \|x_i - b_k\|^2
\]
Results

• M=2000, L=3, K=5 : 42,000 dimensions
• Non-zero ~2,900 dimensions
• 0.39 sec/image
• Linear SVM
  • LLC with SIFT: 91.48%
  • LLC with **U-RI-LBP**: 90.20%
Thank you
Q & A
References

• **SPM:** Beyond Bag of Features: Spatial Pyramid Matching for Natural Scene Categories, CVPR 2006

• **ScSPM:** Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification, CVPR 2009

• **LLC:** Locality-constrained Linear Coding for Image Classification, CVPR 2010