Exploiting visual word co-occurrence for image retrieval

Miaojing Shi ‡, Xinghai Sun ‡, Dacheng Tao #, Chao Xu ‡

‡ Key Laboratory of Machine Perception(MoE), Peking University, China
# Centre for Quantum Computation and Intelligent Systems, University of Technology, Sydney, Australia

2012-09-25
Outline

- IMAGE REPRESENTATION
  - Fast Visual Word Generation via High-order Predictor

- IMAGE RANKING
  - Improved Cosine Similarity Measure: Co-occurrence Weighting
Outline

- IMAGE REPRESENTATION
  - Fast Visual Word Generation via High-order Predictor

- IMAGE RANKING
  - Improved Cosine Similarity Measure: Co-occurrence Weighting Similarity Measure
Outline

- IMAGE REPRESENTATION
  - Fast Visual Word Generation via High-order Predictor

- IMAGE RANKING
  - Improved Cosine Similarity Measure: Co-occurrence Weighting
    - Similarity Measure
Outline

- IMAGE REPRESENTATION
  - Fast Visual Word Generation via High-order Predictor

- IMAGE RANKING
  - Improved Cosine Similarity Measure: Co-occurrence Weighting
    Similarity Measure
Inspiration

- An illustration of one of the co-occurring patterns collected from natural images. Blue and green dots are a pair of co-occurring visual words, which represent the co-occurring pattern. Although they are visually less structured, images in different domains share common co-occurring pattern.
Building of Co-occurrence Table

An illustration of different co-occurring patterns

- Bundling features in affine invariant region: \( N(w_m, w_n) = \sum_i N^i(w_m, w_n) \)

Insights:
- The images containing the same co-occurring pairs might be totally different;
- The locations of co-occurring pairs for the related images could be different;
- If a visual word co-occurs with a large number of words, its uniqueness and distinctiveness decline.
Building of Co-occurrence Table

An illustration of different co-occurring patterns

Bundling features in affine invariant region $N(w_m, w_n) = \sum_i N^i(w_m, w_n)$

Insights
- the images containing the same co-occurring pairs might be totally different;
- the locations of co-occurring pairs for the related images could be different;
- if a visual word co-occurs with a large number of words, its uniqueness and distinctiveness decline.

Exploiting visual word co-occurrence for image retrieval
Building of Co-occurrence Table

An illustration of different co-occurring patterns

- Bundling features in affine invariant region $N(w_m, w_n) = \sum_i N^i(w_m, w_n)$

- Insights
  - the images containing the same co-occurring pairs might be totally different;
  - the locations of co-occurring pairs for the related images could be different;
  - if a visual word co-occurs with a large number of words, its uniqueness and distinctiveness decline.

Exploiting visual word co-occurrence for image retrieval

VISUAL WORD CO-OCCURENCE
Building of Co-occurrence Table

An illustration of different co-occurring patterns

- Bundling features in affine invariant region $N(w_m, w_n) = \sum_i N^i(w_m, w_n)$

- Insights
  - the images containing the same co-occurring pairs might be totally different;
  - the locations of co-occurring pairs for the related images could be different;
  - if a visual word co-occurs with a large number of words, its uniqueness and distinctiveness decline.
An illustration of different co-occurring patterns

Bundling features in affine invariant region

\[ N(w_m, w_n) = \sum_i N^i(w_m, w_n) \]

Insights
- The images containing the same co-occurring pairs might be totally different;
- The locations of co-occurring pairs for the related images could be different;
- If a visual word co-occurs with a large number of words, its uniqueness and distinctiveness decline.

Exploiting visual word co-occurrence for image retrieval
Fast Visual Word Generation via High-order Predictor

- Building of Predictor
- Center Group & Neighbor Group
- Generating of Visual Word
Fast Visual Word Generation via High-order Predictor

- Building of Predictor
- Center Group & Neighbor Group
- Generating of Visual Word
Fast Visual Word Generation via High-order Predictor

- Building of Predictor
- Center Group & Neighbor Group
- Generating of Visual Word
Building of Predictor

- Confidence
  - $C - ?$ or $C - h - i - n - ?a!$

- Prediction

\[
\hat{vw}_0 = f(vw_1, vw_2, ..., vw_s) \tag{1}
\]

- Maximum posterior probability

\[
vw_0^* = \arg \max p(\hat{vw}_0 | vw_1, vw_2, ..., vw_s) \tag{2}
\]
Building of Predictor

- Confidence
  - $C - ?$ or $C - h - i - n - ?a!$

- Prediction
  \[
  \hat{vw}_0 = f(vw_1, vw_2, ..., vw_s) \tag{1}
  \]

- maximum posterior probability
  \[
  vw_0^* = \arg \max p(\hat{vw}_0 | vw_1, vw_2, ..., vw_s) \tag{2}
  \]
Conditional Independence Assumption

- **Decomposition** $p(vw_0 | vw_1, vw_2, ..., vw_s)$

- **Conditional Independence**: complex & unachievable

\[
p(w_s, w_0, ..., w_{s-1}) \approx p(w_s) \prod_{w_i \in S} p(w_i | w_s)
\]

\[
p(w_0, ..., w_{s-1}) \approx \prod_{w_i \in S} p(w_i)
\]  

- **Predictor**

\[
w_s^* = \arg \max_{w_s \in W} p(w_s | w_0, ..., w_{s-1})
\]

\[
\approx \arg \max_{w_s \in W} \frac{p(w_s) \prod_{w_i \in S} p(w_i | w_s)}{\prod_{w_i \in S} p(w_i)}
\]

\[
= \arg \max_{w_s \in W} \frac{N(w_s) \prod_{w_i \in S} N(w_i, w_s)}{\prod_{w_i \in S} N(w_i)}
\]
Neighbor Group & Center Group

high order predictor for visual word generation

Query Training Set
SIFT Features
\{q_0, \ldots, q_{s-1}, q_s, \ldots, q_k\}

CENTER GROUP
NEIGHBOR GROUP

Spatial Co-occurrence Matrix

Exploiting visual word co-occurrence for image retrieval
Representative Approaches

- Approximate Nearest Neighbor (ANN)
  - Tree
  
  \[
  b^* = \arg \min_{b \in \{0, 1, 2, \ldots, B\}} \text{dist}(b) := \{b | \forall q : \|q - c_b\|\} \tag{5}
  \]

- KD-tree, Random KD Trees (RKD)
- K-means-tree, Hierarchical K-means Tree (HKM)
- Fast Library for Approximate Nearest Neighbor (FLANN: ANN, RKD, HKM)

Exploiting visual word co-occurrence for image retrieval
Representative Approaches

- Approximate Nearest Neighbor (ANN)
- Tree

\[ b = \arg \min_{b \in \{0,1,2\ldots B\}} dist(b) := \{b|\forall q : \|q - c_b\|\} \] (5)

- KD-tree, Random KD Trees (RKD)
- K-means-tree, Hierarchal K-means Tree (HKM)

- Fast Library for Approximate Nearest Neighbor (FLANN: ANN, RKD, HKM)

Exploiting visual word co-occurrence for image retrieval
Representative Approaches

- Approximate Nearest Neighbor (ANN)
- Tree

\[ b^* = \arg\min_{b \in \{0, 1, 2, \ldots, B\}} \text{dist}(b) := \{b | \forall q : \|q - c_b\|\} \]  

- KD-tree, Random KD Trees (RKD)
- K-means-tree, Hierarchical K-means Tree (HKM)

- Fast Library for Approximate Nearest Neighbor (FLANN: ANN, RKD, HKM)
**Visual Word Generation Results**

**Oxford I**

- Vocabulary size = 100K
  - High-order predictor (red line) is clearly superior to FLANN (blue line);
  - High-order predictor exhibits a more appealing performance than the first order predictor;
  - Performance on the 100K vocabulary is better than that of the 1M vocabulary.

**Oxford II**

- Vocabulary size = 1M

Exploiting visual word co-occurrence for image retrieval
Visual Word Generation Results

- **vocabulary size = 100K**
  - high-order predictor (red line) is clearly superior to FLANN (blue line);
  - high-order predictor exhibits a more appealing performance than the first order predictor;

- **vocabulary size = 1M**
  - performance on the 100K vocabulary is better than that of the 1M vocabulary.
Visual Word Generation Results

**vocabulary size = 100K**
- high-order predictor (red line) is clearly superior to FLANN (blue line);
- high-order predictor exhibits a more appealing performance than the first order predictor;
- performance on the 100K vocabulary is better than that of the 1M vocabulary.

**vocabulary size = 1M**

---

Exploiting visual word co-occurrence for image retrieval
if a visual word co-occurs with a large number of words, its uniqueness and distinctiveness decline.
Noisy Assumption

- Since our approach focuses on near duplicated image retrieval, the number of relevant images for a given query should be small in large scale dataset. All irrelevant images in the database can be taken as noises that produces negative information in image ranking.

  - Cosine Similarity

    \[ Sim(x, y) = \frac{x^T y}{||x||||y||} \]  

  - Co-occurrence Weighting Similarity

    \[ Sim(x, y) = \frac{x^T (I - \frac{1}{\beta} \Sigma)y}{||x||||y||} \]
Overview of Image Ranking

Co-occurrence weighting similarity: $y^T \Sigma x$
Image Ranking Results

Oxford dataset

Paris dataset

Exploiting visual word co-occurrence for image retrieval

IMAGE RANKING
Representative Approaches

- **HA**: Philbin et al. Object retrieval with large vocabularies and fast spatial matching. CVPR’07
- **SA**: Philbin et al. Lost in quantization: Improving particular object retrieval in large. CVPR’08
- **CM**: Yang et al. Contextual image retrieval model. CIVR’10
- **CVV**: Zhang et al. Building contextual visual vocabulary for large-scale image applications. ACM MM’10
- **SCQE**: Li et al. Query expansion by spatial co-occurrence for image retrieval. ACM MM’11
Comparisons I

Comparisons with different vocabularies

Comparisons with representative approaches

Exploiting visual word co-occurrence for image retrieval
Comparisons II

Performance of Co-Sim embedded on hard-and soft-assigned vocabularies with or without bounding box (BB) on the Oxford dataset.

<table>
<thead>
<tr>
<th></th>
<th>BB</th>
<th>Co-Sim</th>
<th>Ox (100K)</th>
<th>Ox (1M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard</td>
<td></td>
<td></td>
<td>0.514</td>
<td>0.613</td>
</tr>
<tr>
<td>Hard</td>
<td>+</td>
<td></td>
<td><strong>0.584</strong></td>
<td><strong>0.660</strong></td>
</tr>
<tr>
<td>Hard</td>
<td>+</td>
<td>+</td>
<td>0.514</td>
<td>0.613</td>
</tr>
<tr>
<td>Hard</td>
<td>+</td>
<td>+</td>
<td><strong>0.577</strong></td>
<td><strong>0.648</strong></td>
</tr>
<tr>
<td>Soft</td>
<td></td>
<td></td>
<td>0.529</td>
<td>0.640</td>
</tr>
<tr>
<td>Soft</td>
<td>+</td>
<td></td>
<td><strong>0.602</strong></td>
<td><strong>0.719</strong></td>
</tr>
<tr>
<td>Soft</td>
<td>+</td>
<td>+</td>
<td>0.554</td>
<td>0.673</td>
</tr>
<tr>
<td>Soft</td>
<td>+</td>
<td>+</td>
<td><strong>0.611</strong></td>
<td><strong>0.730</strong></td>
</tr>
</tbody>
</table>

With vs. without query bounding box. Visual bounding boxes are often manually labeled for query images to get rid of the nonsensical parts of images. We compare the results of our Co-Sim with and without query bounding box. It shows that the new scheme without bounding box apparently outperforms the results of hard- and soft-assignment even with bounding box. We suggest that this is because, the proposed similarity, while it does not need a manually labeled bounding box, functions intrinsically as a virtual bounding box, with the contribution of nonsensical words being smoothed down to trivia.
Comparisons III

Detailed experimental mAPs in comparisons with representative approaches.

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Ox(100K)</th>
<th>Ox(1M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philbin CVPR’07</td>
<td>0.514</td>
<td>0.613</td>
</tr>
<tr>
<td>SA CVPR’08</td>
<td>0.554</td>
<td>0.676</td>
</tr>
<tr>
<td>CM CIVR’10</td>
<td>0.545</td>
<td>0.658</td>
</tr>
<tr>
<td>CVV MM’10</td>
<td>0.565</td>
<td>0.661</td>
</tr>
<tr>
<td>SCQE MM’11</td>
<td>0.539</td>
<td>0.655</td>
</tr>
<tr>
<td>Co-Sim</td>
<td>0.584</td>
<td>0.660</td>
</tr>
<tr>
<td>Co-Sim(Soft)</td>
<td><strong>0.611</strong></td>
<td><strong>0.730</strong></td>
</tr>
</tbody>
</table>

Large scale retrieval results in comparison with state-of-the-art models.

<table>
<thead>
<tr>
<th>Vocabulary</th>
<th>Ox+I1(1M)</th>
<th>Ox+I2(1M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philbin CVPR’07</td>
<td>0.566</td>
<td>0.499</td>
</tr>
<tr>
<td>SA CVPR’08</td>
<td>0.603</td>
<td>0.534</td>
</tr>
<tr>
<td>CVV MM’10</td>
<td>0.610</td>
<td>0.549</td>
</tr>
<tr>
<td>SCQE MM’11</td>
<td>0.616</td>
<td>0.574</td>
</tr>
<tr>
<td>Co-Sim</td>
<td><strong>0.630</strong></td>
<td><strong>0.615</strong></td>
</tr>
</tbody>
</table>

Exploiting visual word co-occurrence for image retrieval
Question and Answer

- Thank you