NTT at TRECVID 2015: Instance Search


*Communication Science Laboratories, NTT Corporation
**Media Intelligence Laboratories, NTT Corporation
Introduction

- Local feature-based image retrieval is still the most widely used solution for instance search from videos
  - Spatial verification has been widely proved to be successful in this solution

- RANSAC [Philbin+CVPR07][Zhu+TRECVID14]
  - One of the most widely used spatial verification methods
  - Advantage: effective in the rejection of mismatches
  - Disadvantage: quadratic time in the number of SIFT correspondences; have to be founded on a compromise reranking framework
  - Disadvantage: not consider the sensitivity of spatial verification in terms of large 3D viewpoint changes
Effort

- **Complexity**
  - Solution: Ensemble of Weak Geometric Relations (EWGR) [Wu&Kashino+BMVC15]
  - Impose multiple pairwise geometric constraints on pairs of correspondences
  - Advantage: leverage a spatial neighborhood constraint to reduce the complexity from quadratic time to linear time in the number of correspondences

- **Large 3D Viewpoint Change**
  - Problem: local features (even a Hessian affine region detector) are invariant to anisotropic transformations only to a limited extent
  - Solution: Angle Free (AF) [Shimamura+MVA15]
  - Convert each image into a set of affine transformed images to augment the information used for retrieval
System Overview

- **Topic**
  - **Test Videos**
  - **BOVW (Full Search)**
  - **EWGR (Full Search)**
  - **AF (Reranking)**
  - **RL1**
  - **RL2**
  - **RL3**
  - **RL4**
  - **RL5**

Copyright © 2014 NTT Corp. All Rights Reserved.
Preprocessing

- **Keyframe Sampling**
  - Minimum Frame Rate: 6 frames per second
  - #Keyframe: 9,752,650

- **Feature Detection & Description**
  - Hessian affine region detector [Mikolajczyk&Schmid+IJCV04] with rotation switched off
  - #Root SIFT [Arandjelovic&Zisserman+CVPR12]: 15B

- **Vocabulary Construction**
  - Random Sampling: 100M root SIFTs
  - Approximate $k$-means [Philbin+CVPR07] based on a randomized KD-tree

- **Word Assignment**
  - Topic Image: soft assignment [Philbin+CVPR08] with $k = 3$
  - Test Keyframe: hard assignment
Bag of Visual Words (BOVW)

- **Image Encoding**
  - Topic Image: an ROI and a non-ROI TFIDF histogram with 1M dimensions
  - Test Keyframe: a 1M-dimensional TFIDF histogram

- **Similarity Computation**
  - Inverted Index
  - Image-Level Cosine Similarity
  - Weighted average in which the ROI and non-ROI weights were 0.9 and 0.1
  - Shot-Level Average Pooling
Issue on MAP Evaluation Tools

- **Ground Truth (INS.SEARCH.QRELS.TV15)**
  - Label = 1: Relevant
  - Label = 0: Nonrelevant
  - Remainder: Unjudged

- **Version of MAP Evaluation Tool**
  - A: All the unjudged shots are removed from the retrieved set (Our Tool)
  - B: Treat all the unjudged shots the same as “Nonrelevant” (TREC_EVAL_VIDEO)

- **MAPs shown in this report**
  - INS14: A
  - INS15: B
### Performance of Instance Search Based on BOVW

<table>
<thead>
<tr>
<th>Configuration</th>
<th>$\sigma^2$</th>
<th>MAP (INS14)</th>
<th>MAP (INS15)</th>
<th>Time (INS14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>26.0</td>
<td>–</td>
<td>3.73</td>
</tr>
<tr>
<td>2</td>
<td>.1</td>
<td>26.3</td>
<td>–</td>
<td>3.87</td>
</tr>
<tr>
<td>3</td>
<td>.01</td>
<td>27.4</td>
<td>28.4</td>
<td>3.88</td>
</tr>
<tr>
<td>4</td>
<td>.001</td>
<td>27.3</td>
<td>–</td>
<td>3.45</td>
</tr>
</tbody>
</table>

- $\sigma^2$ is the scalar of the exponential function of soft assignment [Philbin+CVPR08]
- “Time” excludes I/O time and the time taken for feature extraction and ranking, and is in units of second per topic
Ensemble of Weak Geometric Relations (EWGR)

[Ensemble of Weak Geometric Relations (Inliers)]

[Neighor, Scaling, Rotation, Relative Position]

[Wu&Kashino+BMVC15]
Complexity

- **Spatial Neighborhood Constraint**
  - Disregard pairs of correspondences if they have a large gap in the image space
  - A correspondence pair \((c_a, c_b)\) is disregarded if \(c_a \notin \mathbb{N}_k(c_b)\) or \(c_b \notin \mathbb{N}_k(c_a)\)
    - \(\mathbb{N}_k(c)\) is the \(k\)-NNs of \(c\) in the image space

- **Great Advantage in Efficiency**
  - Reduce the complexity of all the subsequent verifications from \(\Theta(|C|^2)\) to \(\Theta(\min(|C|, k) |C|) \leq \Theta(k |C|)\)
    - \(|C|\) is the number of correspondences
  - Linear time in \(|C|\) for a fixed \(k\)

- **\(k\)-NN search in the image space**
  - Solution: Randomized KD-Tree [Muja&Lowe+VISAPP09]
  - Complexity: \(\Theta(k |C| \log |C|)\) for a standard KD-tree in theory, and \(\Theta(k |C|)\) for a randomized KD-tree in practice
Performance of Instance Search Based on EWGR

<table>
<thead>
<tr>
<th>Configuration</th>
<th>$k$</th>
<th>$\epsilon_\theta$</th>
<th>$\epsilon_v$</th>
<th>MAP (INS14)</th>
<th>MAP (INS15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>27.4</td>
<td>28.4</td>
</tr>
<tr>
<td>EWGR</td>
<td>80</td>
<td>$\pi/8$</td>
<td>1</td>
<td>29.58</td>
<td>29.94</td>
</tr>
</tbody>
</table>

- Processing time on a per topic basis
  - INS14: 31.5 minutes (1 CPU) and conjecturally 24 seconds (20 CPUs)
  - INS15: 27.0 minutes (1 CPU) and conjecturally 20 seconds (20 CPUs)
  - It should be noted that EWGR searched the full database containing 9.8M images
Top-8 EWGR Mismatches (#9147)
Top-8 EWGR Mismatches (#9129)
Top-8 EWGR Mismatches (#9151)
Angle Free (AF)

Transformation Simulation

[Shimamura+MVA15]

[Morel&Yu+SIAMJIS09]
4D Hough Voting

3D Rotation & Scaling

Voting map

3D Rotation & Scaling

Matches

Query
Reference

Voting

Voting map

Viewpoint changes

Query
Object
Reference

Relative pose \((\Delta \theta, \Delta \varphi, \Delta \psi)\)
## Performance of Instance Search Based on AF

<table>
<thead>
<tr>
<th>Run ID</th>
<th>BOVW</th>
<th>EWGR</th>
<th>AF</th>
<th>#Reranking</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ROI</td>
<td>Full</td>
</tr>
<tr>
<td>–</td>
<td>◎</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>NTT_1</td>
<td>◎</td>
<td>◎</td>
<td>◎</td>
<td>10,000</td>
<td>3,000</td>
</tr>
<tr>
<td>NTT_2</td>
<td>◎</td>
<td>◎</td>
<td>◎</td>
<td>6,000</td>
<td>6,000</td>
</tr>
<tr>
<td>NTT_3</td>
<td>◎</td>
<td>–</td>
<td>◎</td>
<td>10,000</td>
<td>3,000</td>
</tr>
<tr>
<td>NTT_4</td>
<td>◎</td>
<td>◎</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### Fusion Options

<table>
<thead>
<tr>
<th>Run ID</th>
<th>BOVW</th>
<th>EWGR</th>
<th>AF</th>
<th>#Reranking</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ROI</td>
<td>Full</td>
</tr>
<tr>
<td>–</td>
<td>◎</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>NTT_1</td>
<td>◎</td>
<td>◎</td>
<td>◎</td>
<td>10,000</td>
<td>3,000</td>
</tr>
<tr>
<td>NTT_2</td>
<td>◎</td>
<td>◎</td>
<td>◎</td>
<td>6,000</td>
<td>6,000</td>
</tr>
<tr>
<td>NTT_3</td>
<td>◎</td>
<td>–</td>
<td>◎</td>
<td>10,000</td>
<td>3,000</td>
</tr>
<tr>
<td>NTT_4</td>
<td>◎</td>
<td>◎</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

### ROI Weight

<table>
<thead>
<tr>
<th>Fusion</th>
<th>ROI Weight</th>
<th>Number of Top Results for Reranking</th>
<th>#Reranking</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full</td>
<td>1,000     2,000     3,000     6,000     10,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>– –</td>
<td>1 0</td>
<td>28.65 28.93 29.08 29.46 –</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EWGR</td>
<td>1 0</td>
<td>– – – –    30.50 30.76 30.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>– 1</td>
<td>1 1</td>
<td>– – – –    30.56 30.77 –</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EWGR</td>
<td>1 1</td>
<td>– – – –    31.64 31.78 31.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>– 0</td>
<td>0 1</td>
<td>28.73 29.86 30.10 30.14 –</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EWGR</td>
<td>0 1</td>
<td>– – – –    31.46 31.20 30.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
EWGR Misses Rescued by AF (#9148)
EWGR Misses Rescued by AF (#9130)
EWGR Misses Rescued by AF (#9150)
Conclusion

▪ Conclusion
  ▪ Spatial verification is successful in the instance search of near-rigid objects, but has no role in the instance search of deformable objects
  ▪ The use of a spatial neighborhood constraint reduces the complexity from quadratic time to linear time in the number of correspondences
  ▪ Depending on the configuration of local feature detectors, spatial verification is sensitive to globally different but locally similar patterns
  ▪ AF handles large 3D viewpoint changes, small instances and occlusions better than can be expected, but requires much longer processing time because of the greatly enlarged number of images

▪ Future Subject
  ▪ Preprocessing Revisit: the correct MAP of our system based on BOVW was only 19.7% (INS14) even if we used a frame rate of 6 frames per second