Neighbourhood Consensus Networks

Authors:
Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic, Akihiko Torii, Tomas Pajdla Josef Sivic

Presenter:
Eric Page
Problem: Dense Visual Correspondences

AKA Pixel Matching
Problem 2: Repetitive or Textureless Patterns
Background

• Classical approach:
  • Combine feature detection algorithm (SIFT) with nearest neighbor search
  • Compare similarity between points using Euclidian distance, Manhattan distance, cosine similarities etc.

• Modern improvements:
  • Supervised deep learning techniques to replace SIFT
  • Modest improvements over classical methods
  • Deep learning not nearly as influential as in other applications
Background: Neighbourhood Consensus

$I_a$ 

$a$: support region in $I_a$

$s_i = 2$

$s_j = 0$

true match $x_i$

supporting matches for $x_i$

$I_b$

$b$: support region in $I_b$

false match $x_j$
Method

1. Dense feature extraction and matching
2. Neighbourhood consensus network
3. Soft mutual nearest neighbor filtering
4. Extraction of correspondences
5. Weakly supervised training loss
Dense feature extraction and matching

Cosine Similarity

\[ c_{ijkl} = \frac{\langle f_{ij}^A, f_{kl}^B \rangle}{\| f_{ij}^A \|^2 \| f_{kl}^B \|^2} \]

\[ \langle a, b \rangle = ab^T \]

Inner Product

i.e. Dot Product!
Correlation Map

\[
c_{ijkl} = \frac{\langle f_{ij}^A, f_{kl}^B \rangle}{\|f_{ij}^A\|_2 \|f_{kl}^B\|_2}
\]

\[
c_{0011} = 1 \quad \text{Perfect Match!}
\]
\[
c_{1101} = 1 \quad \text{Perfect Match!}
\]
\[
c_{1000} \approx .5
\]
\[
c_{0001} \approx 0
\]
Neighbourhood Consensus Network

$\mathcal{N}_A \times \mathcal{N}_B$

4D input matches

$\mathcal{N}_1$ 4D filters (conv. layer 1)

$\mathcal{N}_2$ (4D $\times$ $\mathcal{N}_1$) filters (conv. layer 2)

1 (4D $\times$ $\mathcal{N}_2$) filter (conv. layer 3)

4D filtered matches
4D Convolution Explanation
Convolutions

\[ h(i) = f \ast w = \sum_{d_x=1} \sum_{d_y=1} f(i + d_x).w(d_x) \]

1D Convolution

\[ h(i,j) = f \ast w = \sum_{d_x=1} \sum_{d_y=1} f(i + d_x, j + d_y).w(d_x, d_y) \]

2D Convolution

\[ h(i,j,k) = f \ast w = \sum_{d_x=1} \sum_{d_y=1} \sum_{d_z=1} f(i + d_x, j + d_y, k + d_z).w(d_x, d_y, d_z) \]

3D Convolution

\[ h(i,j,k,l) = f \ast w = \sum_{d_x=1} \sum_{d_y=1} \sum_{d_z=1} \sum_{d_w=1} f(i + d_x, j + d_y, k + d_z, l + d_w).w(d_x, d_y, d_z, d_w) \]

4D Convolution
Neighbourhood Consensus Network Order Invariance

\[ \tilde{c} = N(c) + (N(c^T))^T, \]
Soft Mutual Nearest Neighbour Filtering

• Filters out matches that are not mutual nearest neighbours
Soft Mutual Nearest Neighbour Filtering

• Keeps matches that are mutual nearest neighbors
Soft Mutual Nearest Neighbour Filtering

\[
(f_{ab}^A, f_{cd}^B) \text{ mutual N.N.} \iff \begin{cases} (a, b) = \arg \min_{ij} \| f_{ij}^A - f_{cd}^B \| \\
(c, d) = \arg \min_{kl} \| f_{ab}^A - f_{kl}^B \|. \end{cases}
\]

Not Differentiable!

\[
\hat{c} = M(c), \quad \text{where} \quad \hat{c}_{ijkl} = r_{ijkl}^A r_{ijkl}^B c_{ijkl},
\]

\[
r_{ijkl}^A = \frac{c_{ijkl}}{\max_{ab} c_{abkl}}, \quad \text{and} \quad r_{ijkl}^B = \frac{c_{ijkl}}{\max_{cd} c_{ijcd}}.
\]
Extracting Correspondences from the Correlation Map

\[ s_{ijkl}^A = \frac{\exp(c_{ijkl})}{\sum_{ab} \exp(c_{abkl})} \quad \text{and} \quad s_{ijkl}^B = \frac{\exp(c_{ijkl})}{\sum_{cd} \exp(c_{ijcd})}. \]
Weakly-Supervised Training Loss

- Training pairs \((I^A, I^B)\)
- Labelled with \(y=1\) for positive pairs
- Labelled with \(y=-1\) for negative pairs

\[
\mathcal{L}(I^A, I^B) = -y \left( \bar{s}^A + \bar{s}^B \right),
\]

\[
s^A_{ijkl} = \frac{\exp(c_{ijkl})}{\sum_{ab} \exp(c_{abkl})} \quad \text{and} \quad s^B_{ijkl} = \frac{\exp(c_{ijkl})}{\sum_{cd} \exp(c_{ijcd})}.
\]
Experiments

Category Level Matching

Instance Level Matching
Category-Level Matching Dataset

- Proposal Flow (PF) - Pascal
- 1,351 image pairs
- 20 object categories
Instance-Level Matching Datasets

• **Training**
  - Indoor Venues Dataset
  - 3,861 positive image pairs
  - 89 different venues in 6 cities

• **Testing**
  - Indoor Visual Localization (InLoc) Dataset
  - 10K perspective cutouts extracted from 227 panoramas
  - Additional 356 query images captured at different times using smart-phones
Experiment Implementation

• Category-Level Matching
  • Feature Resolution: 25x25
  • 3 layers of 5x5x5x5 filters

• Instance-Level Matching
  • Feature Resolution: 200x150
  • Downsampled after computing 4-D correlation map
  • 2 layers of 3x3x3x3 filters

• Both Experiments
  • ResNet-101 Initialized on ImageNet
  • Trained for 5 epochs with lr = 5x10^-4
  • Fine-Tuned for 5 epochs with lr = 1x10^-5
## Quantitative Results

### Category Level
**PF-Pascal**

<table>
<thead>
<tr>
<th>Method</th>
<th>PCK</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCNet-AG+ [12]</td>
<td>72.2</td>
</tr>
<tr>
<td>CNNGeo [28]</td>
<td>71.9</td>
</tr>
<tr>
<td>WeakAlign [29]</td>
<td>75.8</td>
</tr>
<tr>
<td><strong>NC-Net</strong></td>
<td><strong>78.9</strong></td>
</tr>
</tbody>
</table>

Table 1: **Results for semantic keypoint transfer.** We show the rate (%) of correctly transferred keypoints within thresh. $\alpha = 0.1$.

### Instance Level
**InLoc**

<table>
<thead>
<tr>
<th>Dist. (m)</th>
<th>SparsePE [41]</th>
<th>DensePE [41]</th>
<th>DensePE + MNN</th>
<th>DensePE + NC-Net [41]</th>
<th>InLoc + MNN</th>
<th>InLoc + NC-Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.25</td>
<td>21.3</td>
<td>35.3</td>
<td>31.9</td>
<td>37.1</td>
<td>38.9</td>
<td>37.1</td>
</tr>
<tr>
<td>0.50</td>
<td>30.7</td>
<td>47.4</td>
<td>50.5</td>
<td>53.5</td>
<td>56.5</td>
<td>60.2</td>
</tr>
<tr>
<td>1.00</td>
<td>42.6</td>
<td>57.1</td>
<td>62.0</td>
<td>62.9</td>
<td>69.9</td>
<td>72.0</td>
</tr>
<tr>
<td>2.00</td>
<td>48.3</td>
<td>61.1</td>
<td>64.7</td>
<td>66.3</td>
<td>74.2</td>
<td>76.3</td>
</tr>
</tbody>
</table>

Table 2: **Comparison of indoor localization methods.** We show the rate (%) of correctly localized queries within a given distance (m) and 10° angular error.
Results Category-Level

Ours

Ground-truth
Results Category-Level
Results Category-Level
Results Category-Level
Results Category-Level
Results Category-Level
Results: Instance Level

Ours (InLoc+NC-Net) vs Baseline (InLoc)

0.36m, 1.93° vs 8.11m, 48.38°
Results: Instance Level

Ours (InLoc+NC-Net)

Baseline (DensePE)

0.31m, 0.91°

7.53m, 45.04°
Results: Instance Level

**Ours (InLoc+NC-Net)**

**Baseline (DensePE)**

0.75m, 2.31°

4.44m, 2.25°
Results: Instance Level

Ours (InLoc+NC-Net) vs. Baseline (InLoc)

- Ours: 0.02m, 1.07°
- Baseline: 2.73m, 3.23°
Results: Instance Level

<table>
<thead>
<tr>
<th>Ours (InLoc+NC-Net)</th>
<th>Baseline (InLoc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image of Ours]</td>
<td>![Image of Baseline]</td>
</tr>
<tr>
<td>0.41m, 0.83°</td>
<td>6.65m, 23.86°</td>
</tr>
</tbody>
</table>
Questions?
References

• https://jwbian.net/Papers/GMS_CVPR17.pdf

• https://www.google.com/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwiuge6z2sTgAhWjzIMKHSAwDFkQjRxE6BAgBEAU&url=https%3A%2F%2Ffoursquare.com%2Fv%2Fibm-building%2F4b0c6790f964a520973c23e3%2Fphotos&psig=AOvVaw1lpWmlc9ETfG_x6ukrBs2B&ust=1550559373567182

Additional Results: Category-Level

<table>
<thead>
<tr>
<th>Ours</th>
<th>Ground-truth</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Bus Image]</td>
<td>![Bus Image]</td>
</tr>
<tr>
<td>![Train Image]</td>
<td>![Train Image]</td>
</tr>
</tbody>
</table>
Additional Results Category-Level
Results Category-Level