Distinctive 3D local deep descriptors

Fabio Poiesi and Davide Boscaini
Motivation

Compact descriptor

Efficient to compute

Generalise across sensor modalities

Learnable end-to-end
# 3D descriptors for PCD (overview)

<table>
<thead>
<tr>
<th></th>
<th>Hand crafted</th>
<th>Data driven</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="image1" alt="FPFH" /></td>
<td><img src="image2" alt="PPF" /></td>
</tr>
<tr>
<td></td>
<td><img src="image6" alt="SHOT" /></td>
<td><img src="image7" alt="TOLDI" /></td>
</tr>
</tbody>
</table>

LRF: Local Reference Frame
References

How do we learn DIPs?

[Qi2017b] Qi et al., “PointNet: Deep learning on point sets for 3D classification and segmentation,” CVPR 2017
TOLDI LRF \cite{Yang2017,Gojcic2019}

\[
\hat{\Sigma}_S = \frac{1}{|S|} \sum_{p_i \in S} (p_i - p)(p_i - p)^T
\]

\[
S = \{p_i : \|p_i - p\|_2 \leq r_{LRF}\}
\]

\[
\hat{n}_p, \quad \text{if } \sum_{p_i \in S} \langle \hat{n}_p, \overrightarrow{p_i p} \rangle \geq 0
\]

\[
-\hat{n}_p, \quad \text{otherwise}
\]

\[\hat{y}_p = \hat{x}_p \times \hat{z}_p\]

\[\hat{x}_p = \frac{1}{\| \sum_{p_i \in S} \| \alpha_i \beta_i v_i \|_2 \} \sum_{p_i \in S} \alpha_i \beta_i v_i\]

\[v_i = \overrightarrow{pp_i} - \langle \overrightarrow{pp_i}, \hat{z}_p \rangle \hat{z}_p\]

\[\alpha_i = (r_{LRF} - \|p - p_i\|^2)\]

\[\beta_i = \langle \overrightarrow{pp_i}, \hat{z}_p \rangle^2\]

\cite{Yang2017} Yang et al., “TOLDI: An effective and robust approach for 3D local shape description,” Pattern Recognition 2017

\cite{Gojcic2019} Gojcic et al., “The Perfect Match: 3D Point Cloud Matching with Smoothed Densities,” CVPR 2019
PointNet [Qi2017]

- **input**: LRF-rotated patch points \( n = \text{number of points} \)
- **output**: descriptor \( d = 32 \)
- trained via Siamese approach

[Qi2017b] Qi et al., “PointNet: Deep learning on point sets for 3D classification and segmentation,” CVPR 2017
Hardest contrastive loss

Intuition: within a batch, make the descriptors of positive pairs close in the embedding space while making the other descriptors distant

\[ \ell_h = \frac{1}{b} \sum_{(f, f') \in C_+} \left( \frac{1}{|C_+|} [d(f, f') - m_+]^2 \right) + \frac{1}{2|C_+|} \left[ m_+ - \min_{f \in C} d(f, \tilde{f}) \right]^2 + \frac{1}{2|C_+|} \left[ m_+ - \min_{f' \in C} d(f', \tilde{f'}) \right]^2 \]

- \( f \) descriptor
- \((f, f')\) pair of anchors
- \((f, f')\) hardest negatives
- \( m, m_+ \) margins

Descriptors (illustration)
- positive
- negative
Chamfer loss

\[ \ell_c(\mathcal{X}) = \frac{1}{2n} \left( \sum_{x \in \mathcal{X}} \min_{x' \in \mathcal{X}'} \|Ax - A'x'\|_2^2 + \sum_{x' \in \mathcal{X}'} \min_{x \in \mathcal{X}} \|Ax - A'x'\|_2^2 \right) \]

\[ A \in \mathbb{R}^{3 \times 3} \text{ unconstrained} \]

tried \[ \ell_{\text{reg}} = \|I - AA^T\|_F^2 \rightarrow A \rightarrow I \rightarrow \text{no contribution} \]
Experiments

• Training
  • 3DMatch dataset
  • about 16K point-cloud pairs
  • each pair is 256 descriptors
  • 40 epochs

• Testing
  • 3DMatch, ETH

• Evaluation
  • Feature Matching Recall [Deng2018]

[Deng2018] Deng et al., "PPFNet: Global Context Aware Local Features for Robust 3D Point Matching," CVPR 2018

[Zeng2017] Zeng et al., "3DMatch: Learning the matching of local 3D geometry in range scans," CVPR 2017

### Feature-matching recall on the 3DMatch dataset [4].

<table>
<thead>
<tr>
<th>Method</th>
<th>3DMatch</th>
<th>3DMatchRotated</th>
<th>Feat. dim.</th>
<th>Time [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ε</td>
<td>std</td>
<td>Ε</td>
<td>std</td>
</tr>
<tr>
<td>SHOT [17]</td>
<td>.238</td>
<td>.109</td>
<td>.234</td>
<td>.095</td>
</tr>
<tr>
<td>FPFH [16]</td>
<td>.359</td>
<td>.134</td>
<td>.364</td>
<td>.136</td>
</tr>
<tr>
<td>USC [36]</td>
<td>.400</td>
<td>.125</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CGF [37]</td>
<td>.582</td>
<td>.142</td>
<td>.585</td>
<td>.140</td>
</tr>
<tr>
<td>Folding [5]</td>
<td>.613</td>
<td>.087</td>
<td>.023</td>
<td>.010</td>
</tr>
<tr>
<td>PPFNet [6]</td>
<td>.623</td>
<td>.108</td>
<td>.003</td>
<td>.005</td>
</tr>
<tr>
<td>DirectReg [8]</td>
<td>.746</td>
<td>.094</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CapsuleNet [9]</td>
<td>.807</td>
<td>.062</td>
<td>.807</td>
<td>.062</td>
</tr>
<tr>
<td>PerfectMatch [10]</td>
<td>.947</td>
<td>.027</td>
<td>.949</td>
<td>.024</td>
</tr>
<tr>
<td>D3Feat [12]</td>
<td><strong>.958</strong></td>
<td><strong>.029</strong></td>
<td><strong>.955</strong></td>
<td><strong>.035</strong></td>
</tr>
<tr>
<td>DIP</td>
<td>.948</td>
<td>.046</td>
<td>.946</td>
<td>.046</td>
</tr>
</tbody>
</table>
## ETH Dataset

### Feature-matching recall on the ETH dataset [22].

<table>
<thead>
<tr>
<th>Method</th>
<th>Gazebo Summer</th>
<th>Gazebo Winter</th>
<th>Wood Autumn</th>
<th>Wood Summer</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPFH [16]</td>
<td>.386</td>
<td>.142</td>
<td>.148</td>
<td>.208</td>
<td>.221</td>
</tr>
<tr>
<td>SHOT [17]</td>
<td>.739</td>
<td>.457</td>
<td>.609</td>
<td>.640</td>
<td>.611</td>
</tr>
<tr>
<td>3DMatch [4]</td>
<td>.228</td>
<td>.083</td>
<td>.139</td>
<td>.224</td>
<td>.169</td>
</tr>
<tr>
<td>CGF [37]</td>
<td>.375</td>
<td>.138</td>
<td>.104</td>
<td>.192</td>
<td>.202</td>
</tr>
<tr>
<td>PerfectMatch [10]</td>
<td>.913</td>
<td>.841</td>
<td>.678</td>
<td>.728</td>
<td>.790</td>
</tr>
<tr>
<td>D3Feat [12]</td>
<td>.859</td>
<td>.630</td>
<td>.496</td>
<td>.480</td>
<td>.563</td>
</tr>
<tr>
<td>DIP</td>
<td>.908</td>
<td>.886</td>
<td>.965</td>
<td>.952</td>
<td>.928</td>
</tr>
</tbody>
</table>
Conclusions

• Compact descriptor
• Efficient to compute
• Generalise across sensor modalities
• Learnable end-to-end

• Some interesting new research can be explored
  • e.g. 6D pose estimation

CODE AVAILABLE
https://github.com/fabiopoiesi/dip