Pixelwise Object Class Segmentation based on Synthetic Data using an Optimized Training Strategy.

Frank Dittrich, Vivek Sharma, Heinz Woern and Sule Yalilgan
Introduction


Project: AMICA (Ifab, Reis Robotics and MRK-Systems).
Problem Statement

- In the industrial workspace environment:
  - There is no spatial and temporal separation between human worker and industrial-grade components and robots.

- We focus on the
  - Intuitive and natural human-robot interaction.
  - Safety considerations and measures in a shared work environment.
  - The realization of cooperative process.
  - The workflow optimization.
Goal

- The goal is to have correct classification.
- Random decision forest in our research is being used for object class segmentation in real time.
- Application is intended in research scenarios related to safe human-robot cooperation and interaction in the industrial domain.
State of the Art

- Shotton et. al. [7] proposed human body part segmentation as a basis of human pose segmentation, RGB-D pixel centered patch, with motion capture data to detailed and articulated 3D human body models in a virtual environment.

- Stückler et. al. [4] used depth and RGB. Decisions: simple difference tests on the normalized sums of the random features sub-spaces.

- Dumont et. al. [5] used depth and RGB. Decisions: thresholds tests of random dimensions of the feature space.

- Kontscheider et. al. [6] used depth and label context of RGB, comparable to CRF based approach of 4 neighborhood pairwise potentials.
Collection of Data

- **Synthetic Data Generated:**
  - **Depth** frame with additive white Gaussian noise.
  - **RGB** Image (ground truth).
  - **Data Instances**: human (head, body, upper-arm, lower-arm, hands, legs).

- **Unlimited amount of data can be generated.**
  - 640X480\{1(Depth, Float),3(RGB),Integer\}

Figure 1: Synthetic generated depth data and it's corresponding ground truth image.
Robot Simulator

- V-REP
  - Virtual Robot Experimentation Platform [3]
    - Integrated Development Environment (IDE)
    - Distributed Control Architecture
    - Remote API Client
    - Supports: C/C++, Python, Lua, Java, Matlab, Octave or Urbi
    - Free for academic and research purpose
Human Multicolor Data

- Real world choreographies via KINECT skeleton tracking data from a calibrated multi-sensor setup.
- Synthetic representation of 3D human model based on a set of spheres in virtual environment (V-REP)
- Scaling factor for height ranging between 160-190 cm’s.
  \[ S_{scaled} = \lambda \times S_{original} \]
  \{\lambda_{min} \times 168 = 160, \lambda_{max} \times 168 = 190\}
- For testing data ground truth, we use Automatic Annotation approach.
Setup

Figure 2: KINECT skeleton tracking setup.
Training Data: Human

Figure 3: *Left:* KINECT skeleton tracking. *Center:* Coarse approximation of the human body, modeled by small set of 173 spheres arranged along the skeleton estimate. *Right:* Finer sphere approximation of the human body, modeled by a larger set of spheres in the V-REP environment.
Figure 4: Synthetic depth data generated with a synthetic KINECT sensor of human, groundtruth(*left*) and synthetic depth frame with additive white Gaussian Noise(*right*).
Figure 5: Real world depth data of only human. (Top) Real world depth frames and (Bottom) corresponding ground truth data.
Standard Feature Selection

The features are depth information only, centered at the pixel sample patch of constant size. The ordered depth values are then used as the feature description $f$ of the object class sample $s$:

$$f(s) = (f[1:w_p,1], f[1:w_p,2], \ldots, f[1:w_p,h_p]) \in \mathbb{R}^{w_p \cdot h_p},$$

$$f_{i,j} = d_o(s_x + (i - w_p/2), s_y + (j - h_p/2)),$$

$$(i, j) \in \{1, \ldots, w_p\} \times \{1, \ldots, h_p\},$$

Where $(s_x, s_y)$ is the position of sample in the depth frame, $d_o(i, j)$ depicts the operator which returns the depth value of the position $(i, j)$ in the depth frame.

Figure 6: Feature extraction of object class using a rectangular patch, parallel to the image coordinate system and centered at the same position.
Figure 7: Feature patch adaptation

\[ \tilde{f}(s) = \left( \tilde{f}_{[1:w_p],1}, \tilde{f}_{[1:w_p],2}, \ldots, \tilde{f}_{[1:w_p],h_p} \right) \in \mathbb{R}^{w_p \times h_p}, \]

\[ \tilde{f}_{i,j} = d_o(t(i, j)), (i, j) \in \{1, \ldots, w_p\} \times \{1, \ldots, h_p\}, \]

\[ t(i, j) = (b_0, b_1) \cdot \left( \begin{array}{c} \frac{i - w_p/2}{s_x} \\ \frac{j - h_p/2}{s_y} \end{array} \right), \]

where the function \( t \) transforms the patch position \((i, j)\) into a global frame position, using the basis vectors \( b_0 \) and \( b_1 \) of the rotated region coordinate system. \( b_0 \) is the displacement of the pixel sample and \( b_1 \) is the orthogonality constraint.

Figure 8: Feature extraction of the hand pixel sample using a rectangular region.
Classification Approach

- Classification Approach: Random Decision Forest (RDF) [1]
  - Why RDF only?
    - Provides higher accuracy on previous unseen data
    - An ensemble of n binary decision trees is called as Forest.
    - Bagging and randomized node optimization
    - Multi-class classification, fast training, high generalization, easy implementation, predictions can be understood as empirical distribution and high classification performance

Figure 9: Structure of decision tree with root node, Internal nodes and leaf nodes, along with decision criteria to split.
Evaluation

- For the evaluation of the overall segmentation approach, the most optimal parameter setup was used with:
  - Forest size $T = 5$
  - Fixed patch size $(w,h) = (64,64)$
  - Maximum tree depth $D = 15$
  - For the randomization (Ro) in the training process 100 thresholds and 100 feature functions
  - Training is based on synthetic depth frames with additive white Gaussian noise using a std of 15 cm
  - In total 5000 depth frames were generated, 2000 depth frames (F) were chosen in random for training (Data), 300 pixel positions per object class (PC) were chosen uniform in random.

- PC with Intel i7 CPU with 4 core processor, 250GB SSD and 4 GB RAM, pixel prediction for a frame width 640 X 480 pixels.
Figure 10: Comparison of the standard and optimized training strategy using average recall measure as a function of synthetic depth frames.
Figure 11: Prediction results based on synthetic and real-world data with prediction probability thresholding of 0.5 and 0.75 respectively
Confusion Matrix

Using Synthetic Data

<table>
<thead>
<tr>
<th></th>
<th>Bg</th>
<th>He</th>
<th>UB</th>
<th>UA</th>
<th>LA</th>
<th>Ha</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bg</td>
<td>0.95</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>He</td>
<td>0.00</td>
<td>0.93</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>UB</td>
<td>0.00</td>
<td>0.03</td>
<td>0.87</td>
<td>0.08</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>UA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.16</td>
<td>0.80</td>
<td>0.04</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>LA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.14</td>
<td>0.78</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>Ha</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.23</td>
<td>0.75</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>L</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.95</td>
</tr>
</tbody>
</table>

Using Real-World Data

<table>
<thead>
<tr>
<th></th>
<th>Bg</th>
<th>He</th>
<th>UB</th>
<th>UA</th>
<th>LA</th>
<th>Ha</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bg</td>
<td>0.95</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
</tr>
<tr>
<td>He</td>
<td>0.00</td>
<td>0.84</td>
<td>0.08</td>
<td>0.02</td>
<td>0.05</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>UB</td>
<td>0.00</td>
<td>0.00</td>
<td>0.83</td>
<td>0.15</td>
<td>0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>UA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.19</td>
<td>0.67</td>
<td>0.13</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>LA</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.77</td>
<td>0.18</td>
<td>0.00</td>
</tr>
<tr>
<td>Ha</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.15</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>L</td>
<td>0.03</td>
<td>0.00</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Confusion Matrix based Quality Measures

<table>
<thead>
<tr>
<th></th>
<th>Avg</th>
<th>Bg</th>
<th>He</th>
<th>UB</th>
<th>UA</th>
<th>LA</th>
<th>Ha</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall_Synth</td>
<td>0.86</td>
<td>0.95</td>
<td>0.93</td>
<td>0.86</td>
<td>0.79</td>
<td>0.77</td>
<td>0.75</td>
<td>0.94</td>
</tr>
<tr>
<td>Precision_Synth</td>
<td>0.71</td>
<td>1.00</td>
<td>0.97</td>
<td>0.79</td>
<td>0.77</td>
<td>0.72</td>
<td>0.63</td>
<td>0.11</td>
</tr>
<tr>
<td>Recall_Real</td>
<td>0.82</td>
<td>0.94</td>
<td>0.84</td>
<td>0.83</td>
<td>0.67</td>
<td>0.76</td>
<td>0.80</td>
<td>0.87</td>
</tr>
<tr>
<td>Precision_Real</td>
<td>0.61</td>
<td>1.00</td>
<td>0.99</td>
<td>0.70</td>
<td>0.65</td>
<td>0.48</td>
<td>0.46</td>
<td>0.03</td>
</tr>
</tbody>
</table>
Conclusion

- A generic classification approach for pixelwise labeling of object classes, applied to the problem of human body part segmentation in RGB-D data from a ceiling sensor.
- As an innovation, we presented an optimized training strategy which allows for a reduced number of training frames, while preserving the classification performance.
- Goal of using depth only data, works efficiently. High precision and recall values proves that in both cases of synthetic and real world data, it is supported.
- The use of the KINECT skeleton tracking based synthetic data generation.
- RDF with linear feature response shows better results than Axis aligned.
- New data set has been established, and is available on lease for scientific research and academia. It is a top-view dataset.
- High performance of the overall system and the suitability of synthetic training data for the segmentation of the real-world data.

Limitations:
- Pixel count vs training frames, trade-off.
- Tree depth: undefitting vs overfitting.
Future work:
- Parametric.
- Bayesian optimization technique.
- More human localized body parts.
- Human height with more variability.
References


Thanks 😊