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

Frank Dittrich, Vivek Sharma, Heinz Woern and Sule Yalilgan

Karlsruher Institut für Technologie

Høgskolen i Gjøvik

Institut für Prozessrechentechnik, Automation und Robotik (IPR)

Introduction

Domain: Scene Analysis in Safe Human-Robot Collaboration & Safe-Human-Robot-Interaction.

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}

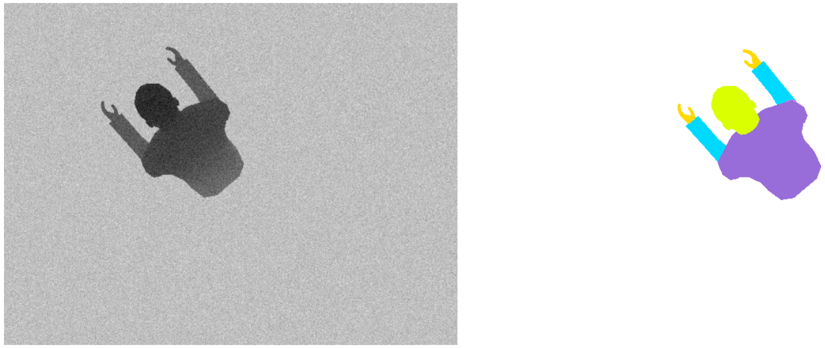

Figure 1: Synthetic generated depth data and it's corresponding ground truth image.

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

Karlsruher Institut für Technologie

- 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.

$$S_{scaled} = \lambda \, x \, S_{original}$$

$$\{\lambda_{min}\text{x}168=160,\ \lambda_{max}\text{x}168=190\}$$

- For testing data ground truth, we use Automatic Annotation approach.

Setup

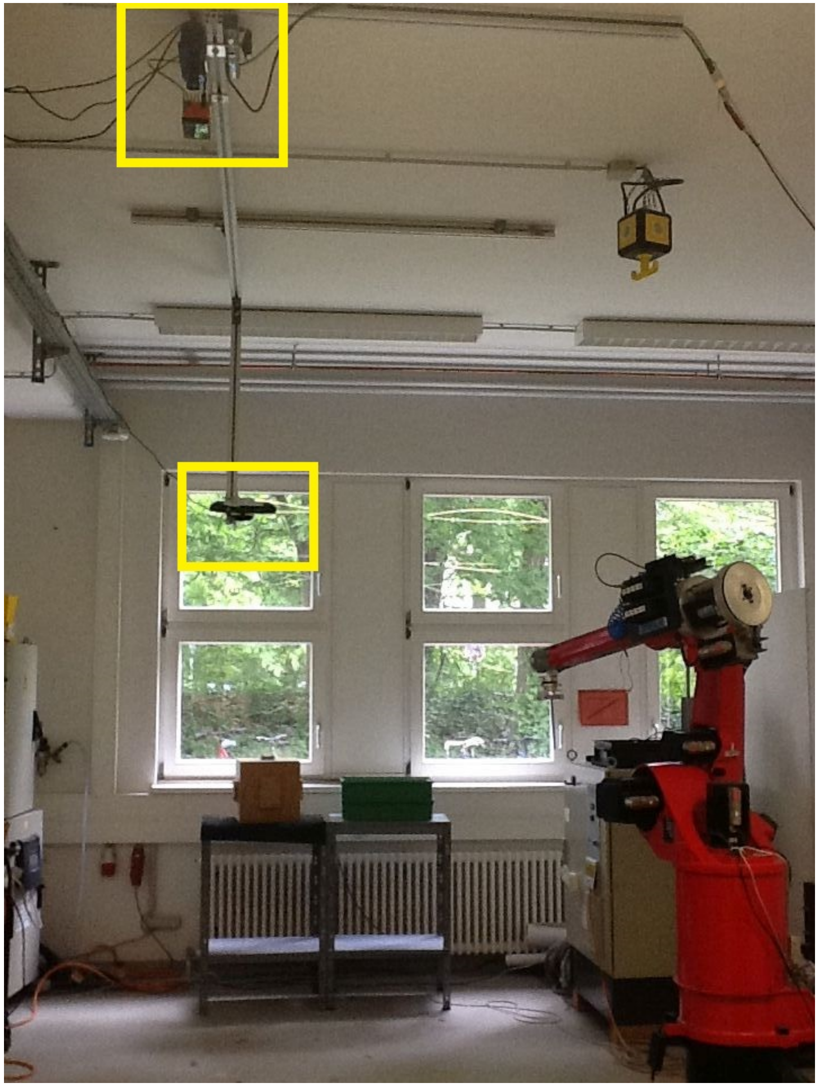

Figure 2: KINECT skeleon tracking setup.

Institut für Prozessrechentechnik, Automation und Robotik (IPR)
Prof. Dr.-Ing. H.Wörn

Training Data: Human

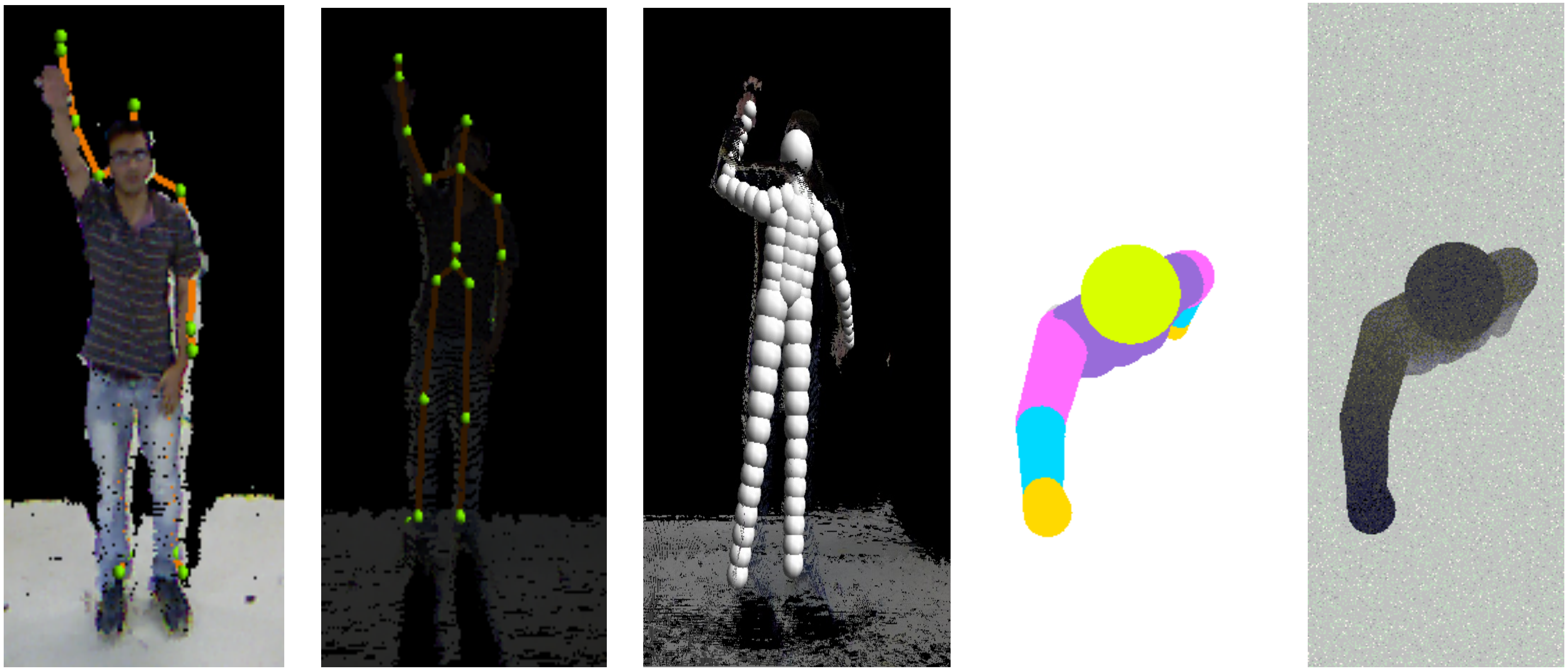

Figure 3: *Left:* KINECT skeleon tracking. *Center:* Coarse approximation of the human body, modeled by small set of 173 spheres arraged along the skeleton estimate. *Right:* Finer sphere approximation of the human body, modeled by a larger ser of spheres in the V-REP environment.

Training Data: Human

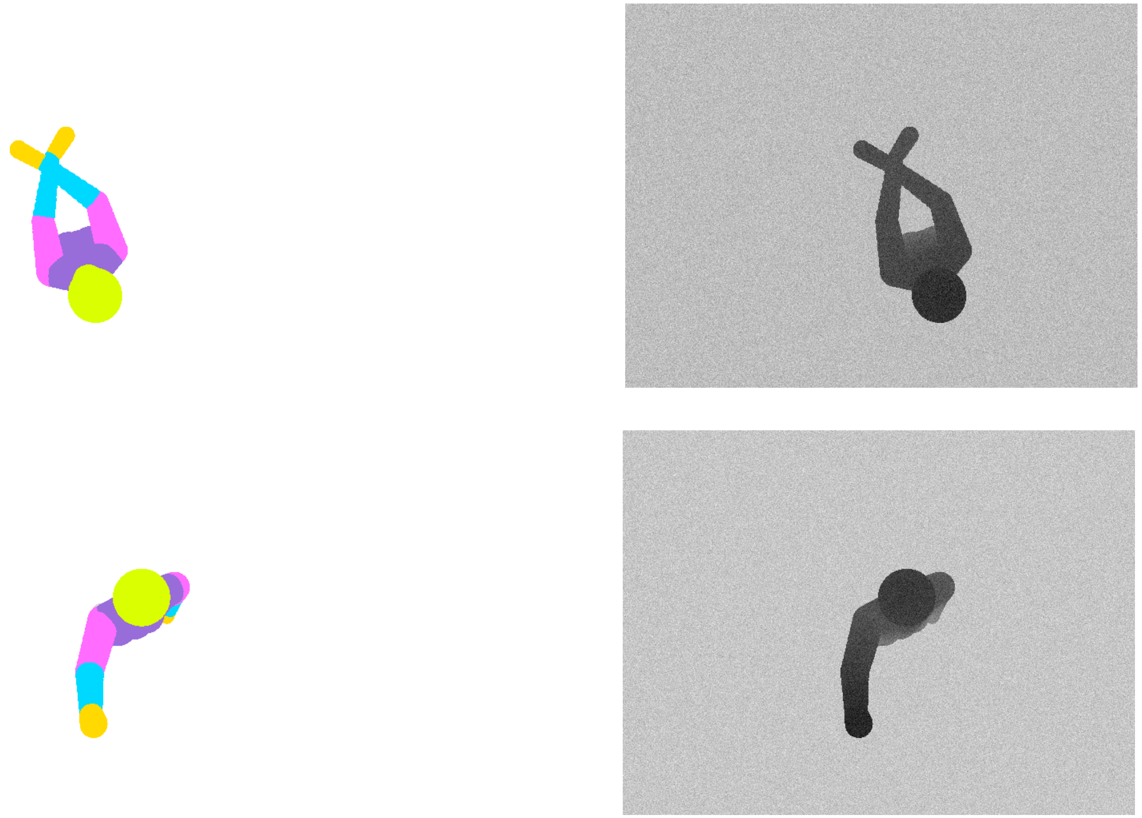

Figure 4: Synthetic depth data generated with a snythetic KINECT sensor of human, groundtruth(*left*) and synthetic depth frame with additive white Gaussian Noise(*right*).

Institut für Prozessrechentechnik, Automation und Robotik (IPR)
Prof. Dr.-Ing. H.Wörn

Testing Data

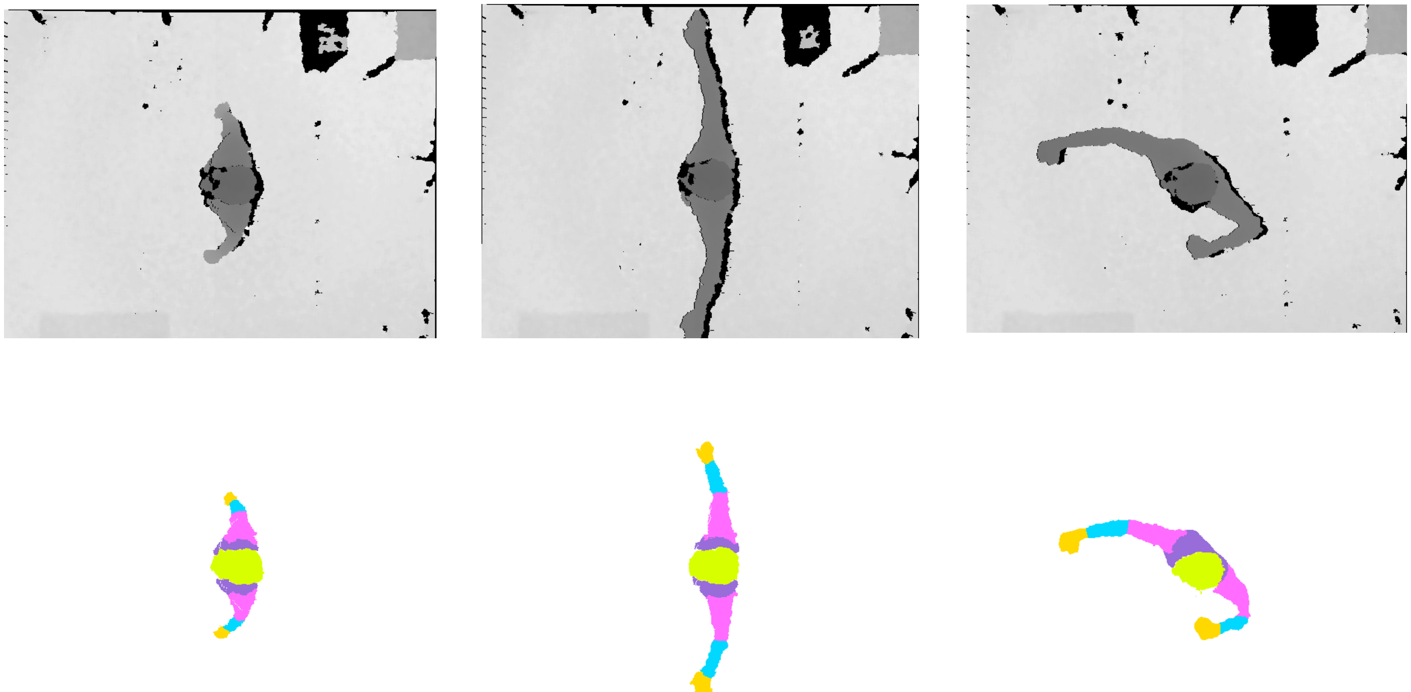

Figure 5: Real world depth data of only human. (*Top)* Real world depth frames and (*Bottom*) corresponding ground truth data.

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:

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

$$f_{i,j} = \mathrm{d_o}\left(s_x + (i - w_p/2), s_y + (j - h_p/2)\right) \;,$$

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

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.

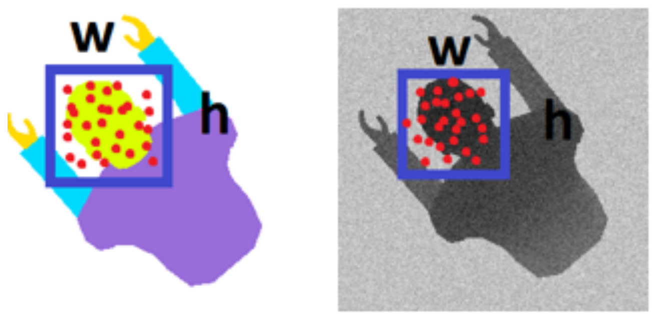

Figure 6: Feature extraction of object class using a rectangular patch, parallel to the image coordinate system and centered at the same position.

Institut für Prozessrechentechnik, Automation und Robotik (IPR)
Prof. Dr.-Ing. H.Wörn

Optimized Feature Selection

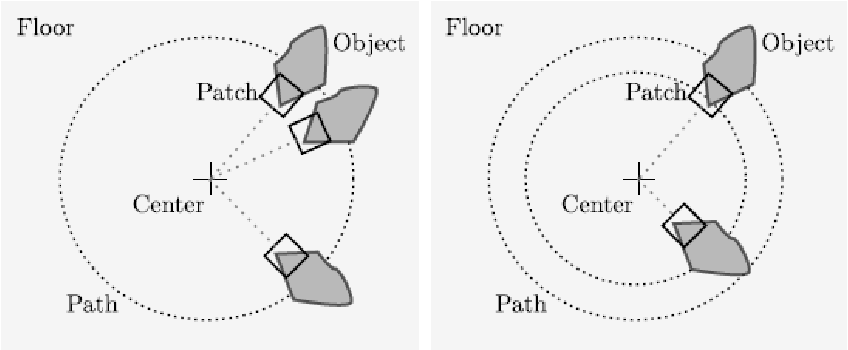

Figure 7: Feature patch adaptation

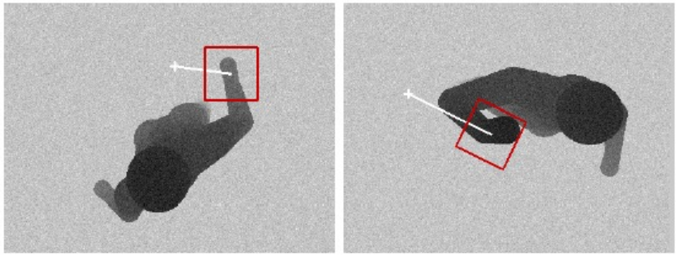

Figure 8: Feature extraction of the hand pixel sample using a rectangular region.

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

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

$$\mathrm{t}(i,j) = (\mathbf{b}_0, \mathbf{b}_1) \cdot \begin{pmatrix} i - w_p/2 \\ j - h_p/2 \end{pmatrix} + \begin{pmatrix} s_x \\ s_y \end{pmatrix} \quad ,$$

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.

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 implemetation, predictions can be understood as empirical distribution and high classification performance

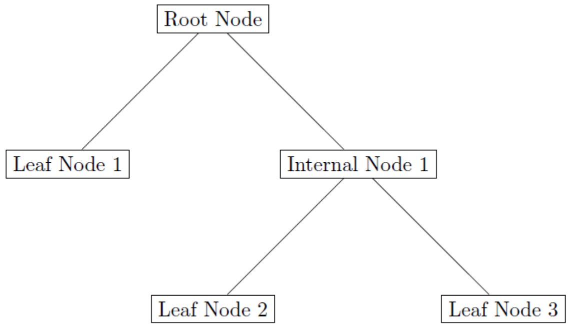

Figure 9: Structure of decision tree with root node, Internal nodes and leaf nodes, along with decision criteria to split.

Evaluation

Karlsruher Institut für Technologie

- 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 feaure 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.

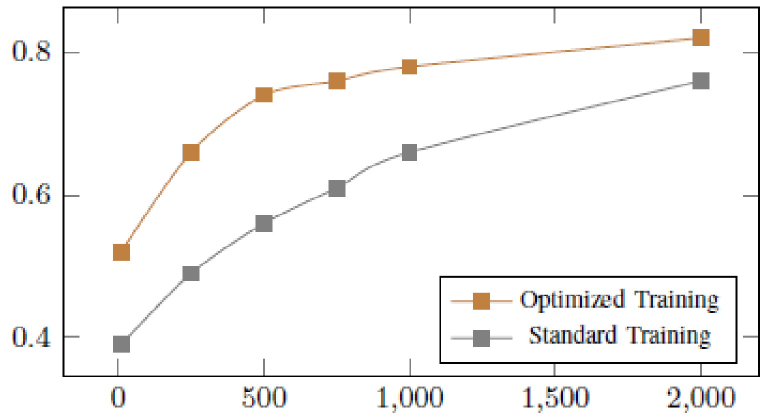

Figure 10: Comparison of the standard and optimized training strategy using average recall measure as a function of synthetic depth frames.

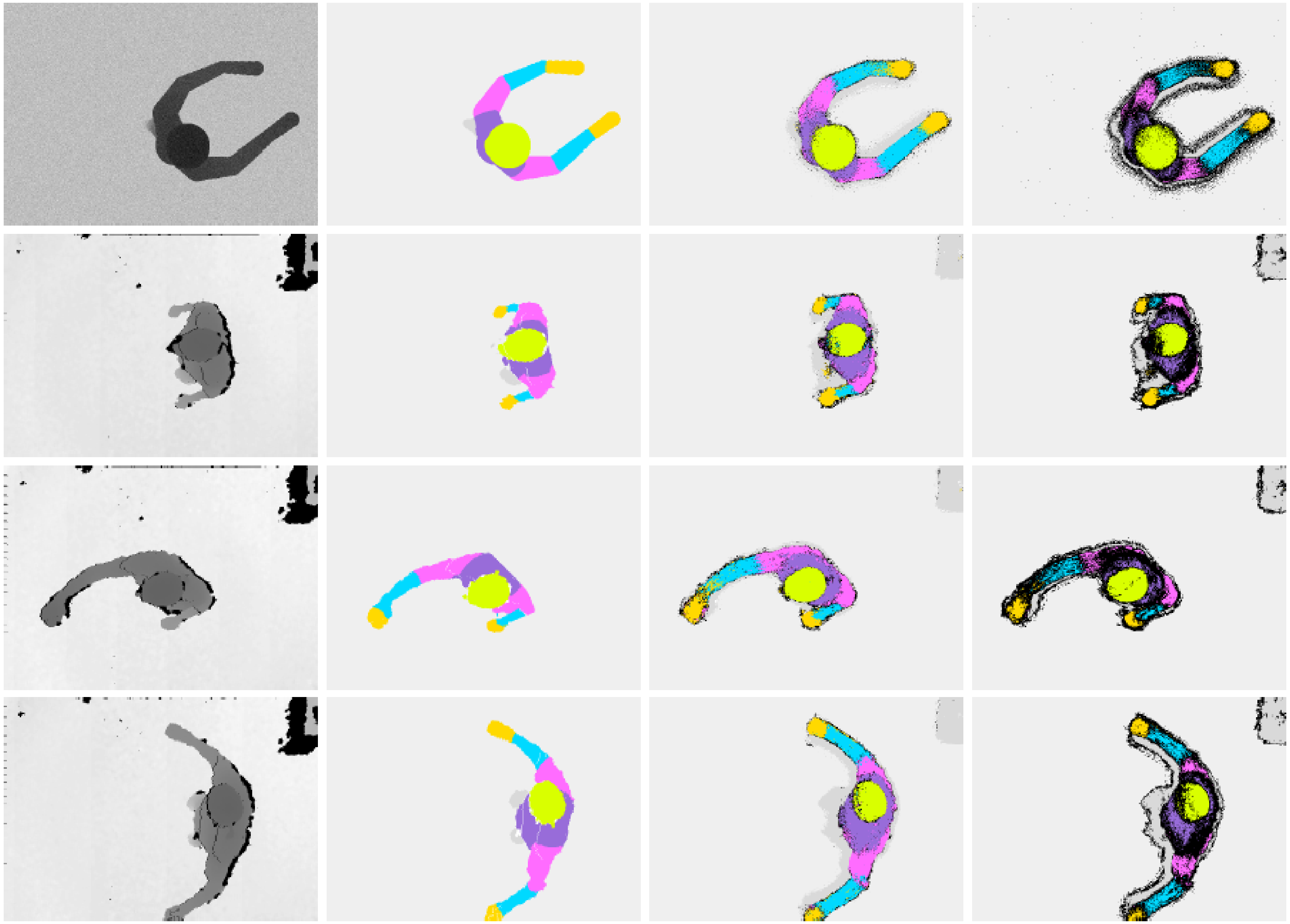

Figure 11: Prediction results based on synthetic and real-world data with prediction probability thresholding of 0.5 and 0.75 respectively

KIT Karlsruher Institut für Technologie

Confusion Matrix

Using Synthetic Data

		Bg	He	UB	UA	LA	Ha	L
Bg	(Background)	**0.95**	0.00	0.00	0.00	0.00	0.00	0.05
He	(Head)	0.00	**0.93**	0.05	0.01	0.01	0.00	0.00
UB	(Upper Body)	0.00	0.03	**0.87**	0.08	0.00	0.00	0.02
UA	(Upper Arm)	0.00	0.00	0.16	**0.80**	0.04	0.00	0.00
LA	(Lower Arm)	0.00	0.00	0.02	0.14	**0.78**	0.06	0.00
Ha	(Hand)	0.00	0.00	0.00	0.02	0.23	**0.75**	0.00
L	(Legs)	0.00	0.00	0.04	0.00	0.01	0.00	**0.95**

Using Real-World Data

	Bg	He	UB	UA	LA	Ha	L
Bg	**0.95**	0.00	0.00	0.00	0.00	0.00	0.05
He	0.00	**0.84**	0.08	0.02	0.05	0.01	0.00
UB	0.00	0.00	**0.83**	0.15	0.02	0.00	0.00
UA	0.00	0.00	0.19	**0.67**	0.13	0.01	0.00
LA	0.00	0.00	0.00	0.05	**0.77**	0.18	0.00
Ha	0.00	0.00	0.00	0.04	0.15	**0.81**	0.00
L	0.03	0.00	0.04	0.02	0.01	0.03	**0.87**

Confusion Matrix based Quality Measures

	Avg	Bg	He	UB	UA	LA	Ha	L
Recall_Synth	**0.86**	0.95	0.93	0.86	0.79	0.77	0.75	0.94
Precision_Synth	**0.71**	1.00	0.97	0.79	0.77	0.72	0.63	0.11
Recall_Real	**0.82**	0.94	0.84	0.83	0.67	0.76	0.80	0.87
Precision_Real	**0.61**	1.00	0.99	0.70	0.65	0.48	0.46	0.03

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

- [1]. Decision Forests for Computer Vision and Medical Image Analysis. A. Criminisi and J. Shotton, Springer 2013, Advances in Computer Vision and Patter Recognition(ACVPR).
- [2]. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout and Context. Jamie Shotton, John Winn, Carsten Rother, Antonio Criminisi. 2007
- [3]. http://coppeliarobotics.com/
- [4] Jorg Stuckler, Nenad Biresev, and Sven Behnke. Semantic mapping using object-class segmentation of RGB-D images. In IROS, pages 3005–3010. IEEE, 2012.
- [5] Dumont et al. Fast Multi-class Image Annotation with Random Subwindows and Multiple Output Randomized Trees. In Alpesh Ranchordas and Helder Arajo, editors, VISAPP (2), pages 196–203. INSTICC Press, 2009.
- [6] Kontschieder et al. Structured class-labels in random forests for semantic image labelling. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 2190–2197, November 2011.
- [7] Shotton et al. Real-time Human Pose Recognition in Parts from Single Depth Images. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR ’11, pages 1297–1304. IEEE Computer Society, 2011.

Thanks ☺