How does Energy Minimization Improve Recognizing Human Poses for Safe Human-Robot Collaboration?

Vivek Sharma[◇†⋆], Frank Dittrich[†], Şule Yildirim-Yayilgan[⋆], Heinz Wörn[†]

[†]Karlsruhe Institute of Technology [⋆]Gjøvik University College [◇]KU Leuven, ESAT-PSI, iMinds

Email: {vivek.sharma, sule.yayilgan}@hig.no {frank.dittrich, woern}@kit.edu

Problem Statement

- In the industrial scenario humans and robots often share the same workspace posing a lot of threats to human safety issues.
- 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.
- We use a random decision forest (RDF) and a conditional random field (CRF) for pixelwise object class labeling of human body-parts using depth measurements obtained from KINECT RGB-D ceiling sensor.
- We use energy minimization (EM) method in order to improve recognition of human body parts.

Related Work

- Shotton et al. in [1], propose a segmentation approach purely based on pixelwise classification using boosted classifier.
- Shotton et al. in [2], demonstrate the application of segmentation of human body-parts for human pose segmentation in real-time using decision forests.
- Sharma et al. in [4], propose an optimized training strategy for pixelwise segmentation.

Data Collection

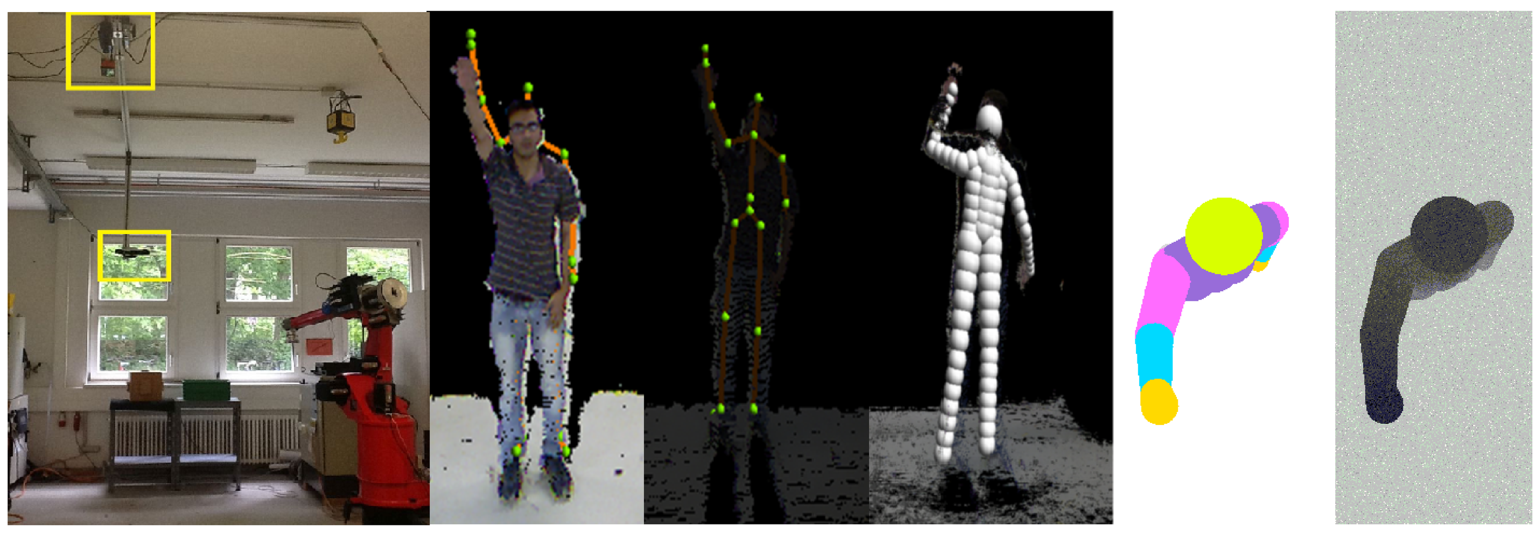

Figure 1: Synthetic human data generation. (*From Left to Right*) Multi-sensor KINECT skeleton tracking setup at our robotic workplace. Real-world human skeleton tracking using KINECT, skeletal joints of interest of real-world human, 3D human skeleton modeled on a set of 173 spheres, ground truth labeling of depth data and corresponding depth data (when KINECT sensor is above the human model at a height of 3.5 meters).

- Human body-parts: *head, body, upper-arm, lower-arm, hand and legs*.
- Poses and shape: *sitting, standing, walking, working, dancing, swinging, boxing, tilting, bending, bowing, and stretching* with combinations of angled arms, single and both arms and other combinations.
- Human height range: 160-190 cm.

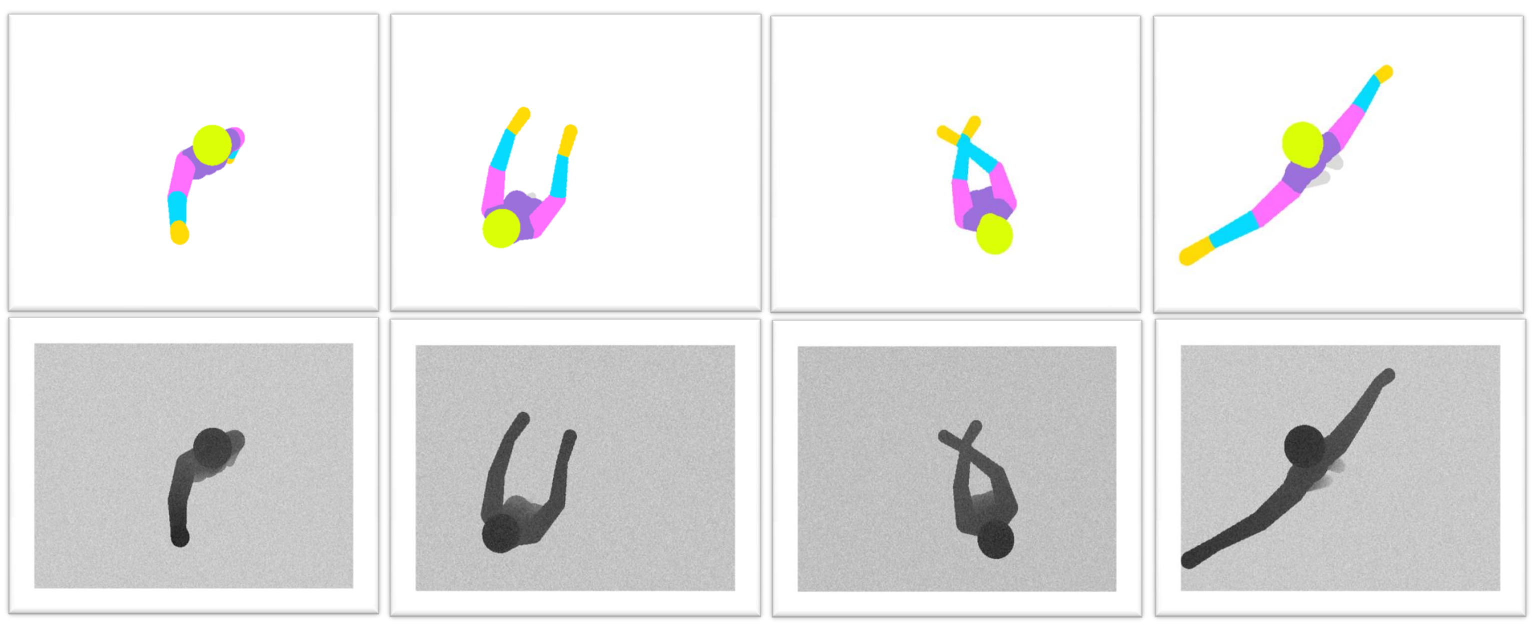

Figure 2: Synthetic human data for training. *Top*: Ground truth labels of depth data. *Bottom*: Corresponding synthetic depth data.

Proposed System

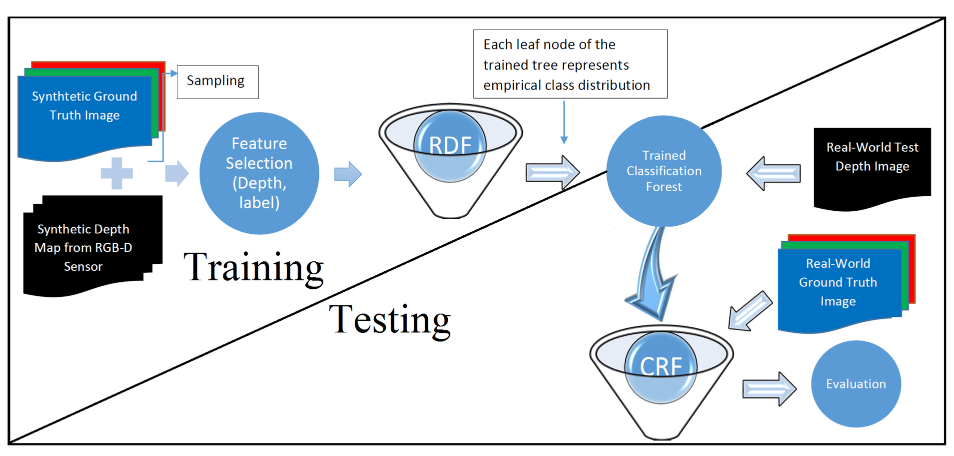

Figure 3: Schematic layout of the segmentation system.

Proposed Approach

- The EM or CRF energy is defined as:

$$E(\mathbf{x}) = \sum_{i \in \upsilon} \varphi_i(x_i) + \sum_{i \in \upsilon, j \in \eta} \varphi_{i,j}(x_i, x_j)$$

- Unary term $(\varphi_i(x_i))$ is the likelihood of an object label assigned to pixel i, obtained from the RDF classifier.
- Pairwise smooth term $(\varphi_{i,j}(x_i, x_j))$ is in the form of Potts model [3] which can be efficiently minimized by α-expansion.
- α-Expansion [3] built on graph cuts are meant for solving multi-labeling problems.

Results and Conclusion

	Avg	Head	Body	UArm	LArm	Hand	Legs
RDF_{mAR}	**0.780**	0.920	0.764	0.730	0.703	0.722	0.845
RDF_{mAP}	**0.569**	0.930	0.656	0.681	0.430	0.491	0.230
EM_{mAR}	**0.843**	0.946	0.835	0.849	0.651	0.791	0.987
EM_{mAP}	**0.725**	0.975	0.696	0.741	0.777	0.802	0.361

Table 1: mAR and mAP measures obtained for each of RDF and EM methods, using a confusion matrix and test real-world data

- We generate qualitative (see Figure.4) and quantitative (see Table.1) results in our tests with RDF and EM methods.
- EM improves the performance measures by approximately 12% in mean average-recall (mAR) and 15% in mean average precision (mAP) over the RDF performance measures.
- Quantitative results appear more meaningful for practicability review of Safe Human-Robot Collaboration.
- In [2], number of training frames (F) = 300K/tree with pixel-count-per object class (PC) = 2000 takes a lot of training time, has a high computational cost and has large memory consumption.
- In our case, F=1600/tree with PC=300 is sufficient for producing almost comparable results, with reduced computational expense and training time.
- Our work can distinguish subtle changes such as crossed-arms which was not possible in [2].

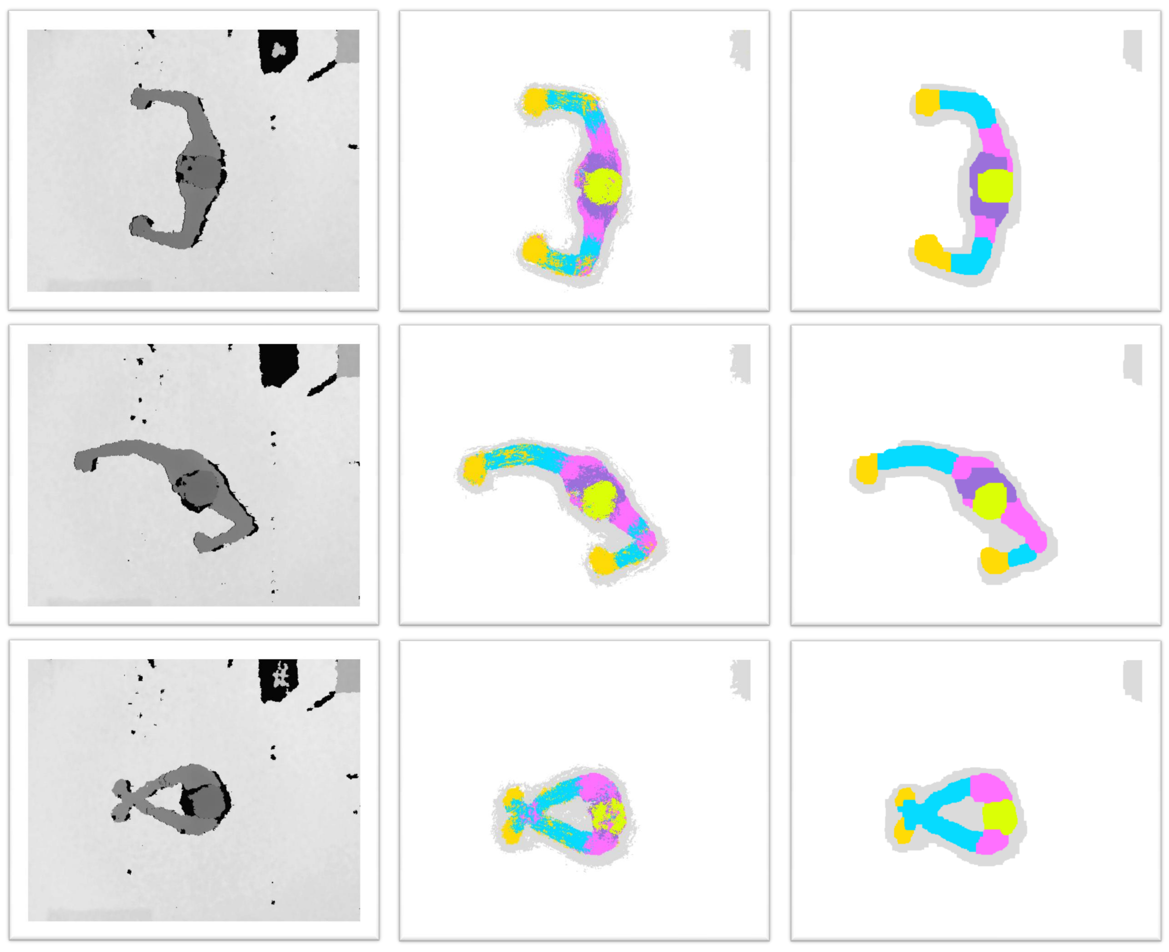

Figure 4: Prediction results based on real-world human test depth data. The first column shows the test real-world depth frames, the second and third column show the predictions obtained from RDFs and EM method.

Ackowledgements

This work is supported by the BMBF funded project AMIKA and the EU project ROVINA.

References

[1] Shotton, J., Winn, J., Rother, C., and Criminisi, A. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. J. Comput. Vision, 2009.

[2] Shotton, J., Girshick, R. B., Fitzgibbon, A. W., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A. and Blake, A.. Efficient human pose estimation from single depth images. IEEE Trans. Pattern Anal. Mach. Intell., 2013.

[3] Boykov, Y., Veksler, O., and Zabih, R.. Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell., 2001.

[4] Dittrich, F., Sharma, V., Woern, H. and Yayilgan, S. Pixelwise Object Class Segmentation based on Synthetic Data using an Optimized Training Strategy. IEEE Intl. Conf. on Networks & Soft Computing, 2014.