Lehrstuhl für Mustererkennung

FAU Friedrich-Alexander Universität Erlangen-Nürnberg
Faculty of Engineering

Manifold Learning-based Data Sampling for Model Training

Shuqing Chen[1], Sabrina Dorn[2,3], Michael Lell[4], Marc Kachelrieß[2,3], Andreas Maier[1]

[1] *Pattern Recognition Lab, Department of Computer Science, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen, Germany*

[2] *German Cancer Research Center (DKFZ), Heidelberg, Germany*

[3] *Ruprecht-Karls-University Heidelberg, Heidelberg, Germany*

[4] *University Hospital Nuremberg, Paracelsus Medical University, Nuremberg, Germany*

Introduction

- Random training data sampling can cause inaccurate model in case of **small data sample size** and **nonuniform data sample distribution.**
- This manifold learning-based approach can solve the data sampling problem and improve the model training.
- The approach can be employed for different machine learning methods.
- The tests of two methods showed a **largest Dice improvement with 0.244.**

Materials and Methods

Goal:

- To avoid the bias due to the nonuniform data sampling
- To keep the distribution of the selected data sets similar

Steps:

- **Data Representation**: project the high-dimensional volumetric medical data into a low-dimensional visualization plane using manifold learning techniques
- **Data Clustering**: divided the data into different classes using clustering techniques and choose a reasonable clustering with help of the visualization
- **Data Selection**: build training data set, validation data set, and test data set by selecting data randomly from each class

Results and Discussion

Experiment 1: multi-organ segmentation with atlas registration [1]:

- 20 CT volumes in total, 17 for training, 3 for test
- Manifold learning: locally linear embedding (LLE) [2]
- Clustering: k-Mean (k=3)
- Max. improvement of Dice = 0.09

Experiment 2: multi-organ segmentation with cascaded U-Net [3]:

- 42 DECT volumes in total, 30 for training, 6 for validation, 6 for test
- Manifold learning: LLE
- Clustering: k-Mean (k=3)
- Max. improvement of Dice = 0.244

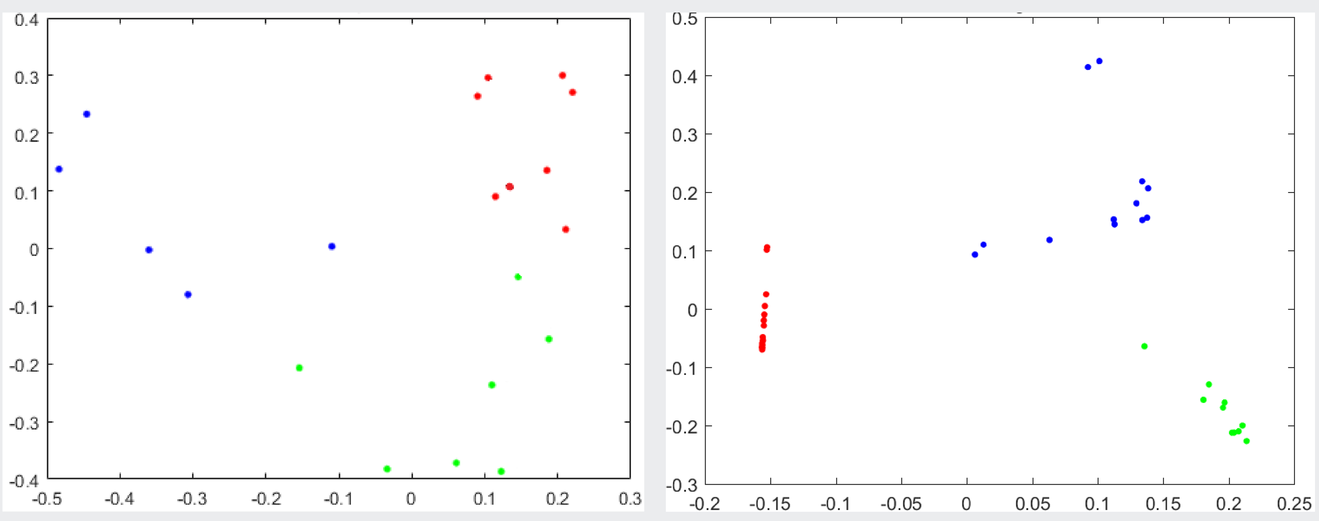

Figure 1: Data representation and clustering. Colors denote classes.

a) CT data distribution for atlas registration b) DECT data distribution for cascaded U-Net

	Right Lung	Left Lung	Right Kidney	Left Kidney	Liver	Spleen
Random Data Selection						
Dice	0.960±0.014	0.957±0.015	0.794±0.140	0.731±0.214	0.900±0.034	0.813±0.104
Proposed Method						
Dice	0.965±0.009	0.960±0.010	0.834±0.080	0.821±0.121	0.912±0.024	0.842±0.051
Improvement of the average						
Diff.	0.005	0.003	0.040	0.090	0.012	0.029

Table 1: Comparison of multi-organ segmentation using atlas registration on CT images with random data selection and the proposed method.

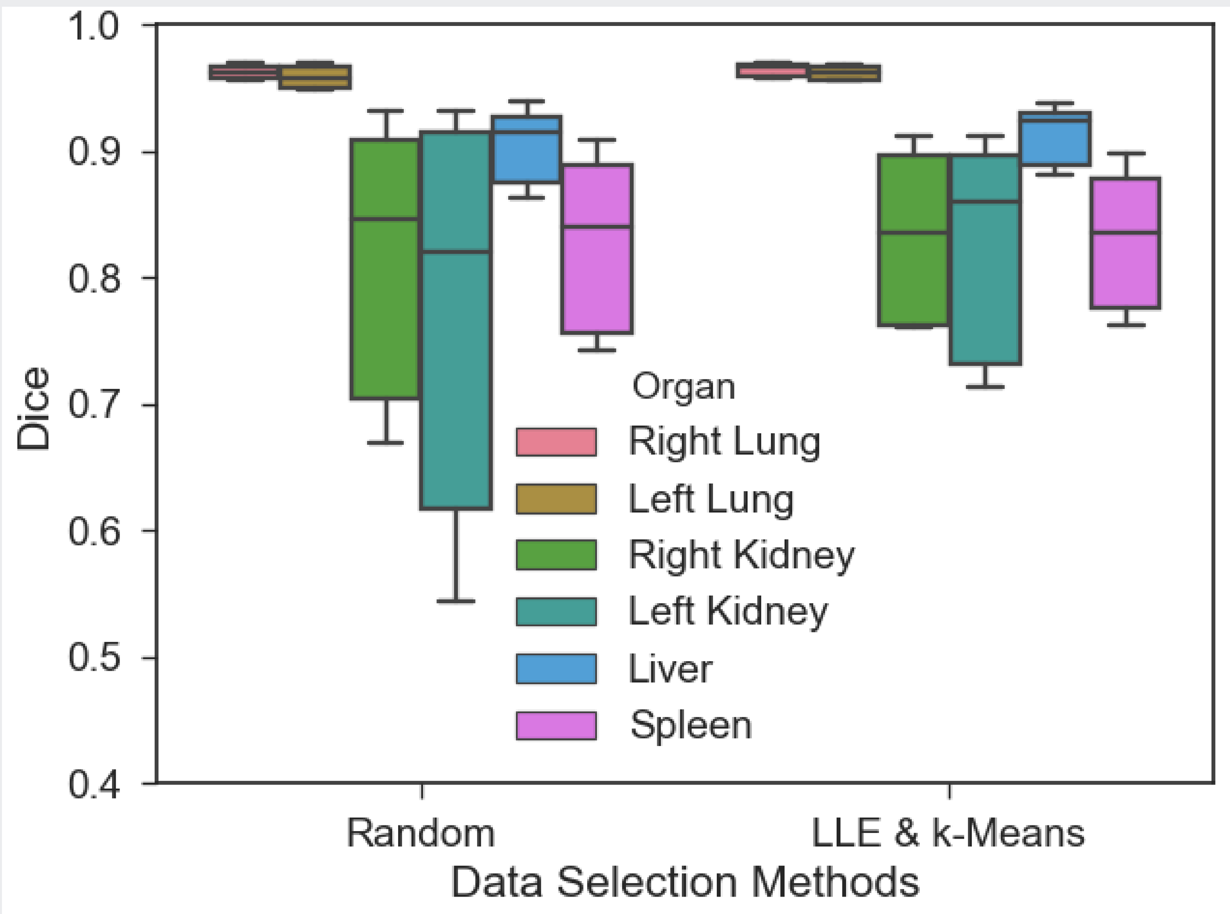

Figure 2: Comparison of multi-organ segmentation using cascaded U-Net on DECT images with random data selection and the proposed method.

	Right Kidney	Left Kidney	Liver	Spleen
Random Data Selection				
Dice	0.905±0.020	0.856±0.047	0.919±0.015	0.652±0.188
Proposed Method				
Dice	0.905±0.034	0.863±0.071	0.934±0.011	0.896±0.032
Improvement of the average				
Diff.	0.000	0.007	0.015	0.244

Table 2: Comparison of multi-organ segmentation using cascaded U-Net on DECT images with random data selection and the proposed method.

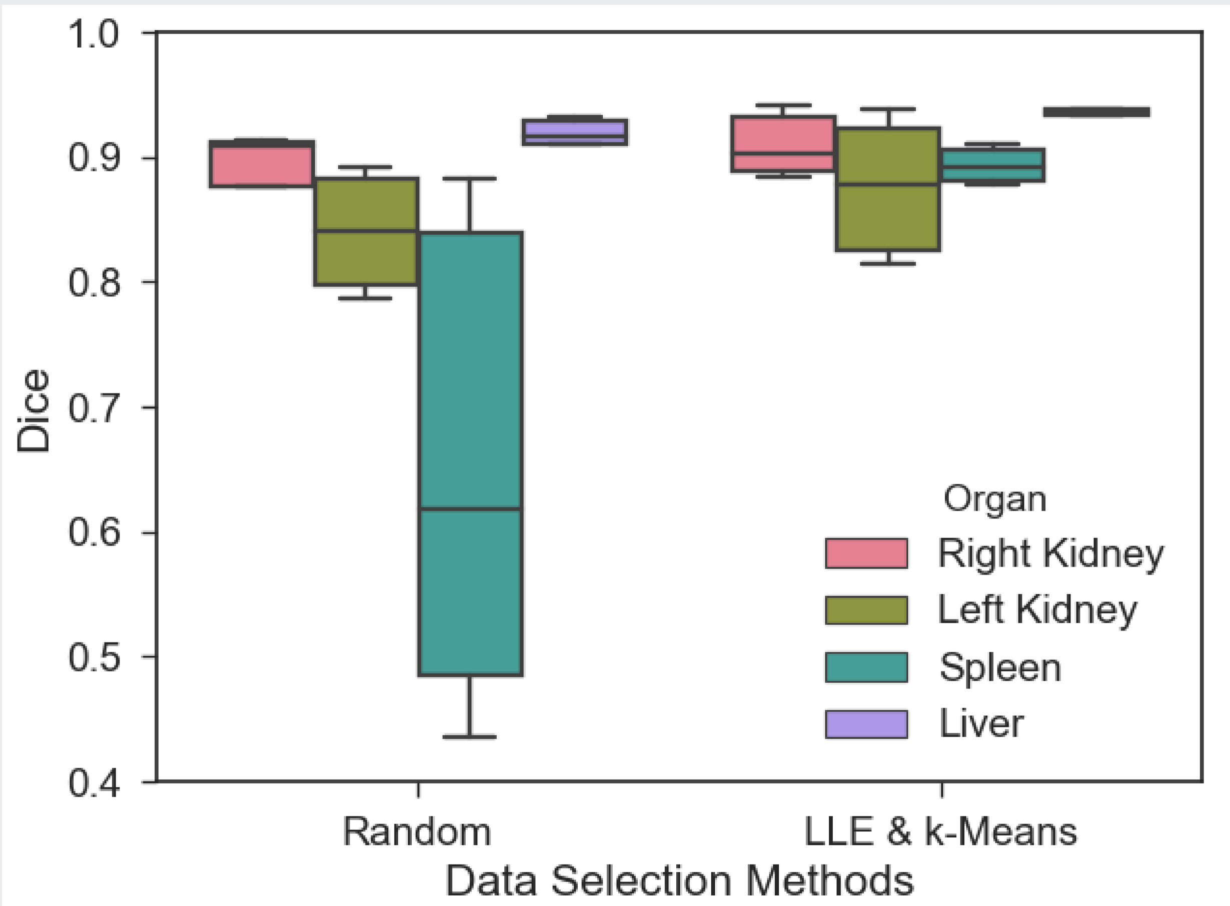

Figure 3: Comparison of multi-organ segmentation using cascaded U-Net on DECT images with random data selection and the proposed method.

Conclusions

- Data sampling is an important part in machine learning for data with small sample size and data with nonuniform sample distribution.
- Manifold learning-based sampling method can improve the data sampling and the model training.
- The bias caused by data selection can be reduced by using the proposed approach.

Contact

Shuqing.Chen@fau.de

http://www5.cs.fau.de/~chen

References

[1] Chen et al.: A feasibility study of automatic multi-organ segmentation using probabilistic atlas. Proc BVM, 2017

[2] Maaten et al.: Dimensionality reduction: a comparative review. 2008

[3] Chen et al.: Towards automatic abdominal multi-organ segmentation in dual energy CT using cascaded 3D fully convolutional network. arXiv. 2017

Acknowledgment: The authors gratefully acknowledge the support of the German Research Foundation (DFG). Note that the concepts and information presented in this paper are based on research, and they are not commercially available.