Intrinsic Correspondence of Classification Ground Truth and Image Content on the Example of Endoscopic Images

Johannes Schuiki • Andreas Uhl

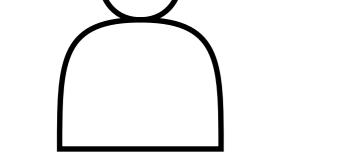
Paris Lodron University of Salzburg Department of Artificial Intelligence and Human Interfaces

Abstract

On what basis class labels ("ground truth") get assigned to images heavily depends on the application scenario, sometimes even without visual inspection of the data. Therefore, it can be of interest to evaluate whether distinguishing intrinsic structures exist within the image data. In this study, it is investigated if images from five small-scale endoscopic datasets where class labels were assigned based on domain-specific criteria can be algorithmically clustered into the desired classes. The image classification task is treated as a clustering comparison problem by comparing ground truth labels with clustering results derived from a variety of image representations.

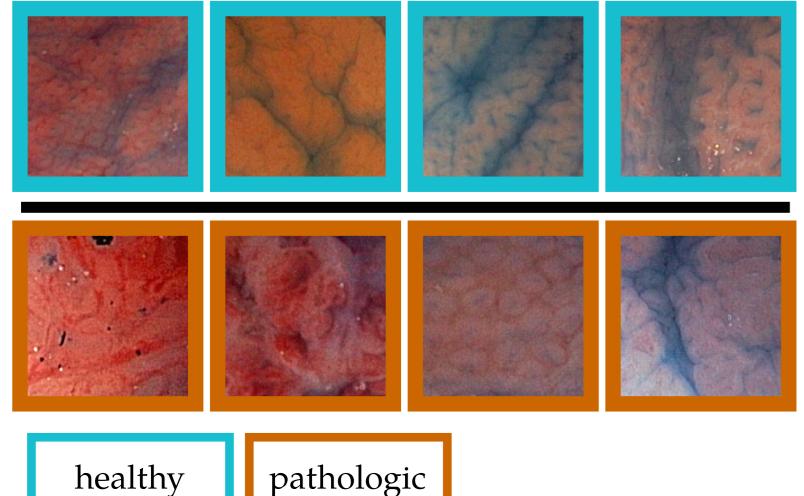
How to assign ground truth?





based on visual inspection by domain experts

Colonic Polyps High Magnification Endoscope



based on biopsy and visual inspection by domain experts

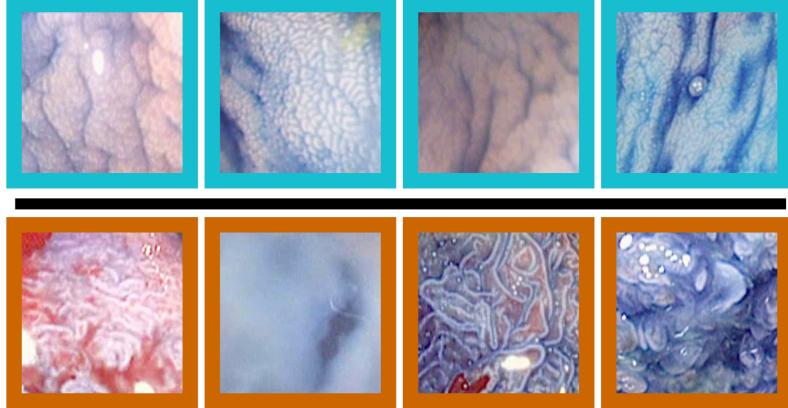
> Celiac Disease Narrow Band Imaging





based on biopsy / histological analysis

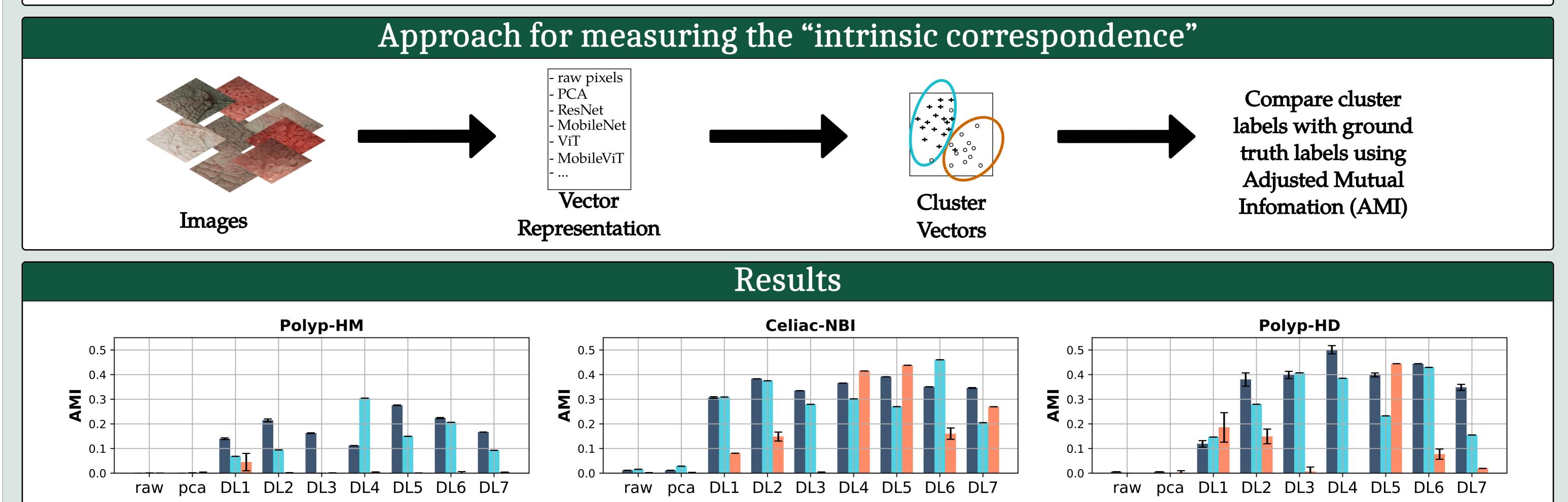
Colonic Polyps High Definition Endoscope



Research Question: Do labels correspond to visual content?

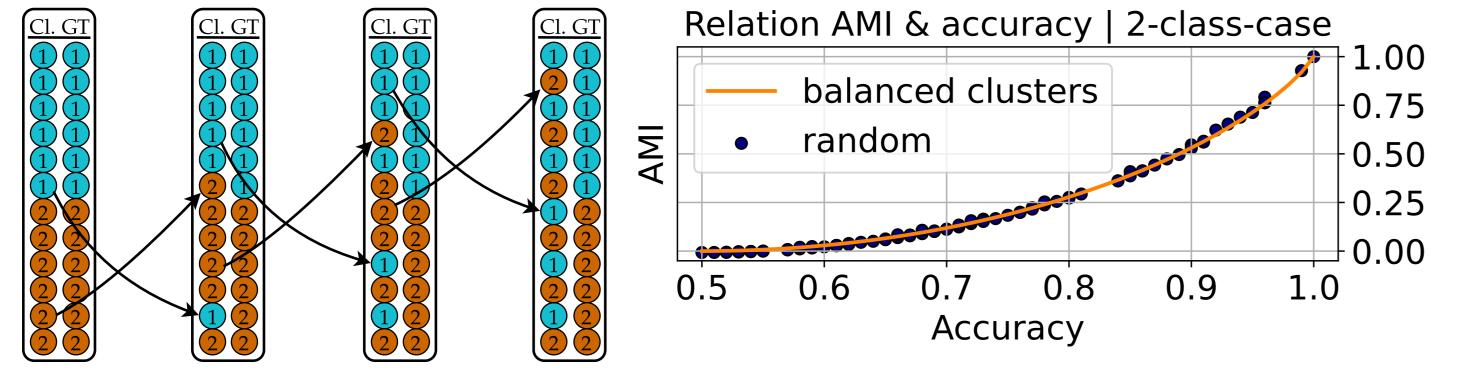
Why not just train a classifier?

Answer: Ideally, we want to find a "natural representation" of images and avoid learning something unintended by force \rightarrow intrinsic correspondence



kmeans clustering	spectral clustering	pca = first 108 PC	DL2 = tinynet	DL4 = resnet50	DL6 = efficientnet
hierachical clustering	raw = pixel vectorization	DL1 = mobilenetv3	DL3 = vit_tiny_patch16	DL5 = resnet18	DL7 = mobilevit

Simulation: Relationship AMI and accuracy



Discussion / Conclusion

▶ not all image labels reliably reflect the underlying visual characteristics
▶ expert-assigned labels do not always perfectly correspond to visual content (harmonizes with existing literature on inter-rater-variability)
▶ moderate to high correspondence (AMI ≥ 0.4) between ground truth and clustering results for biopsy based labeling
▶ approach for measuring "intrinsic correspondence" limited by absence of universally optimal "natural representation"

This work has been partially supported by the Salzburg State Government WISS2025 Research and Transfer Lab on "Artificial Intelligence in BioMedical Image Analysis (AIBIA)".