Comparing Image Labeling Based on Art Historical Criteria to Automatic Clustering

Johannes Schuiki[1], Miriam Landkammer[2], Isabella Nicka[2] and Andreas Uhl[1]

[1]Department of Artificial Intelligence and Human Interfaces, University of Salzburg, Jakob-Haringer-Straße 2, 5020 Salzburg, Austria

[2]Institute for Medieval and Early Modern Material Culture, University of Salzburg, Körnermarkt 13, 3500 Krems an der Donau, Austria

ISPA 2025 Paper #19

IMAREAL
INSTITUT FÜR REALIENKUNDE DES MITTELALTERS UND DER FRÜHEN NEUZEIT

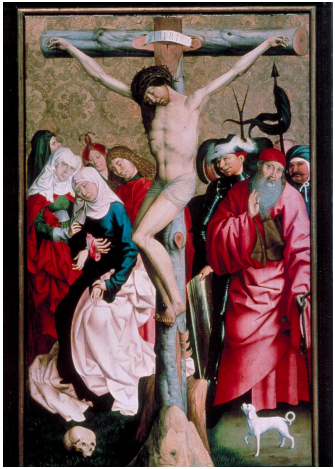
LAND SALZBURG

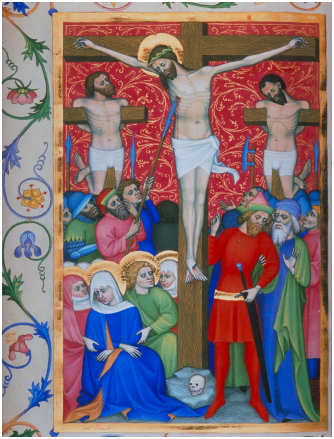
WAVELAB
Multimedia Signal Processing and Security Lab

Introduction

- digital humanities project; computer vision & art history
- overall topic: analysis of painted material
- trend in medieval Europe 14th and 15th century: draw/paint material realistically
- a prominent example where material wood appears on paintings: crucifixion of Jesus.

Depiction of Crucifixion of Jesus





Intro

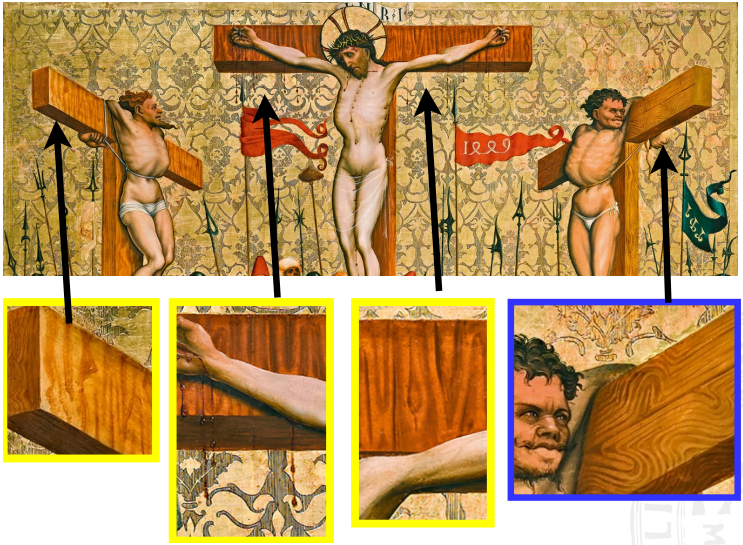
Art-historical perspective:

- how artists painted materials (in this case: wood)
- which wood textures were used more frequently than others and why
- cumbersome to do manually through large corpora

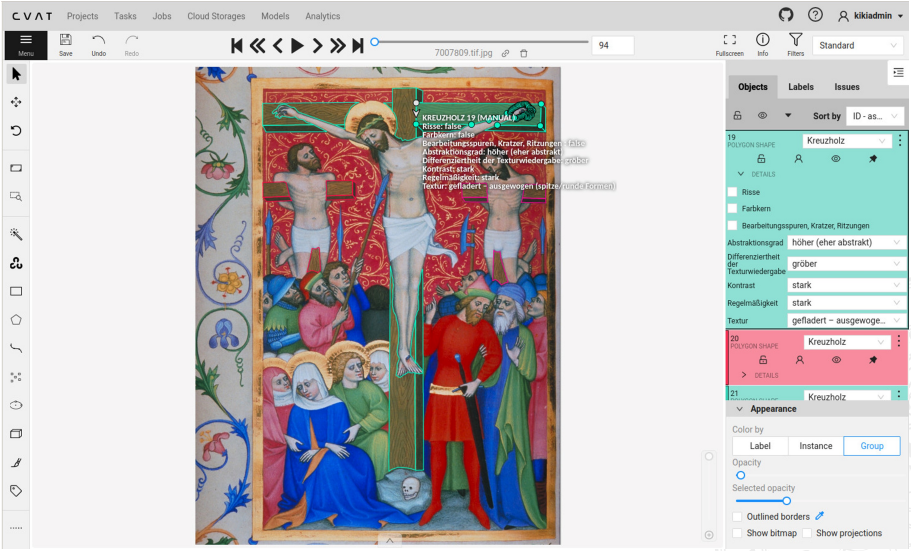
Computer science perspective:

- (eventually) deliver tools to do such analyses more efficiently

Intro: Example of wood texture interpretation

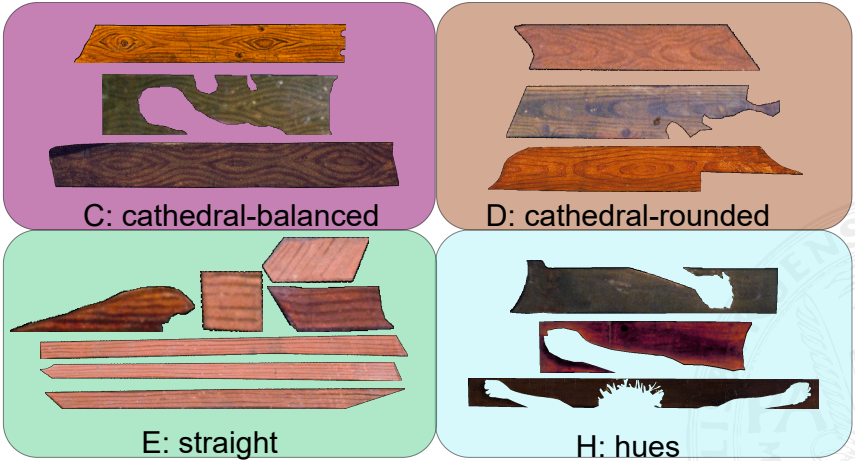


Intro: Manual Segmentation using CVAT

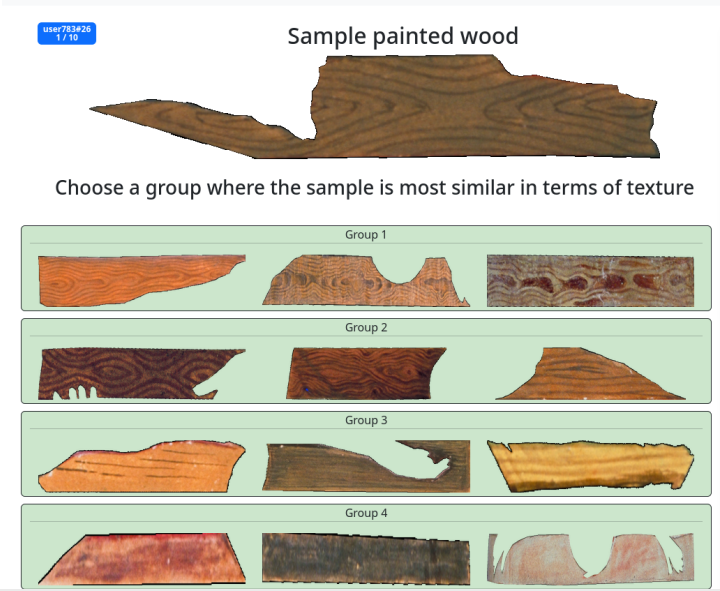


Intro: Some statistics and examples

In total: 287 Images were annotated, resulting in 4107 pieces of wooden cross. (most of them too small)



Motivation: Science-to-public event



Motivation: Science-to-public event

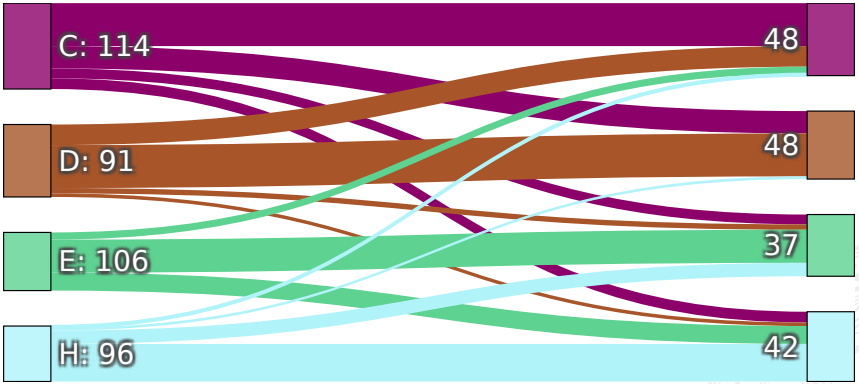


Figure: Class assignments from visitors at the 'Long Night of Research' (Lange Nacht der Forschung)

Motivation: ResNet-50 embeddings

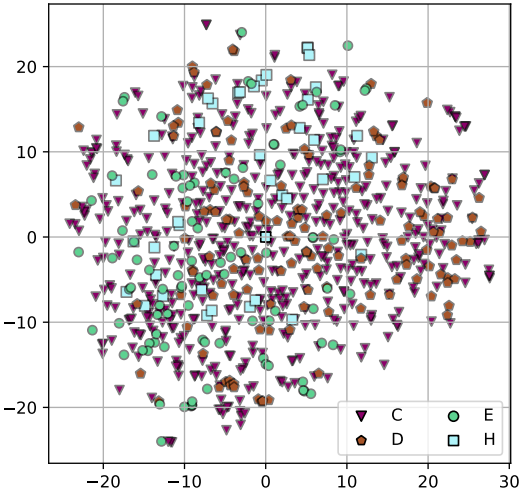


Figure: t-SNE plot of ResNet-50 embeddings of all wooden pieces

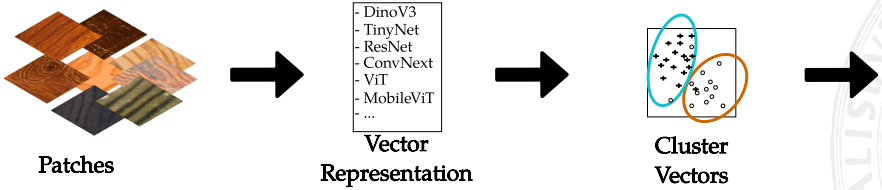
This study

Hypothesis: Labels according to art historians are unsuited.

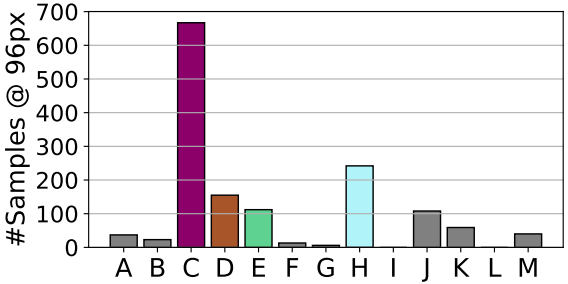
Question: Can we measure ”intrinsic” correspondence of labels (ground truth) and actual image content?

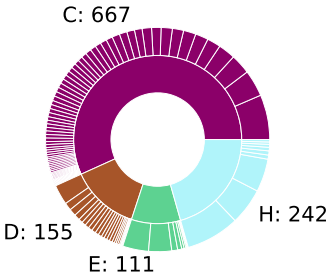
Task: Find ”natural representation” of image content / texture and check its separability according to ground truth labels

Approach to measure this / experiment design:



Patch Dataset Distribution





Legend

- A: no texture
- B: cathedral-pointed
- C: cathedral-balanced
- D: cathedral-rounded
- E: straight
- F: undulating
- G: dotted
- H: hues
- I: burl
- J: diffuse
- K: defect
- L: growth rings
- M: bark

Methods: Patch Embeddings

Classical Features:

- Dense SIFT & Fisher Vector Encoding (+ PCA)

Pre-trained (classic ImageNet) models:

- TinyNet
- ViT Tiny
- ResNet-18
- MobileViT
- ConvNext Tiny (Dino V3)
- ViT Small (Dino V3)

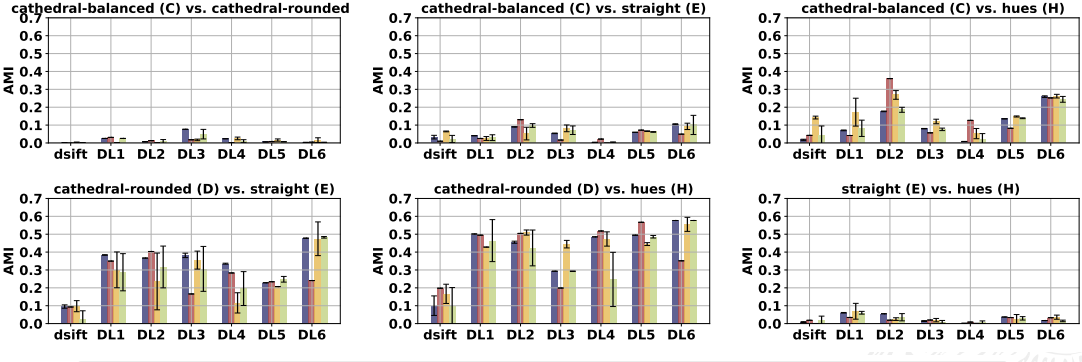
Methods: Clustering Methods

Clustering Methods used:

- Vanilla k-Means clustering[1]
- Hierarchical / agglomerative clustering[1]
- Deep Embedded Clustering (DEC)[2]
- Subspace k-Means[2]

[1] implementations used from scikit-learn
[2] implementations used from https://github.com/collinleiber/ClustPy

Results


cathedral-balanced (C) vs. cathedral-rounded
cathedral-balanced (C) vs. straight (E)
cathedral-balanced (C) vs. hues (H)
cathedral-rounded (D) vs. straight (E)
cathedral-rounded (D) vs. hues (H)
straight (E) vs. hues (H)
AMI
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
dsift
DL1
DL2
DL3
DL4
DL5
DL6
Legend
kmeans clustering
hierarchical clustering
deep embedded clustering
subspace kmeans
dsift = dense sift + fisher enc.
DL1 = tinynet_a
DL2 = vit_tiny_patch16_224
DL3 = resnet18
DL4 = mobilevit_s
DL5 = dinov3 convnext_tiny
DL6 = dinov3 vit_small

Conclusion

- Expert labels do not necessarily align with automatic, texture-based grouping (nor with laymen assignments at science to public event)
  - → rethink texture classes
- Subjective bias could also be a factor
  - → assign multiple labels to each piece and combine them

The End

Q & A

Interpretation of AMI value

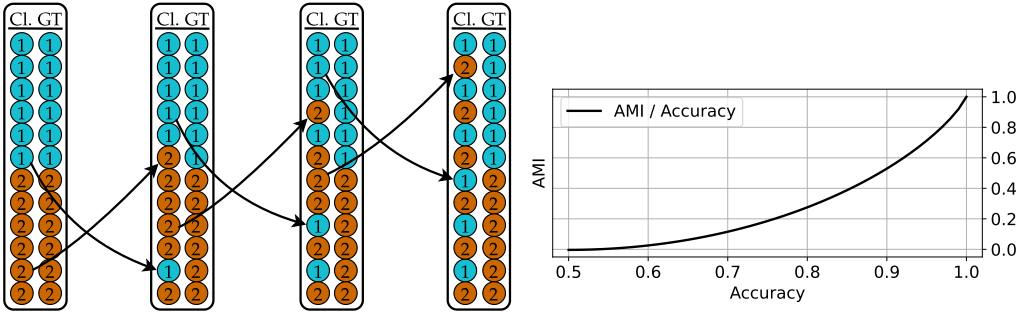


Figure: Left: Simulated alienation of ground truth ; Right: Resulting relation