

Relation between  
Agreement Measures on Human Labeling and  
Machine Learning Performance:  
Results from an Art History Domain

Becky Passonneau, Columbia University

Tom Lippincott, Columbia University

Tae Yano, Carnegie Mellon University

Judith Klavans, University of Maryland

# FSC Image/Text Set: AHSC


- Images: *ARTstor Art History Survey Collection*; 4000 works of art and architecture
- Texts: two from a concordance of a dozen art history surveys used in creating the AHSC
- Meets our criteria: Curated, minimal cataloging, image/text association
- Characteristics of the texts:
  - Neolithic art to 20<sup>th</sup> century
  - About 30 chapters each; 20-40 plates per chapter (surrogate images freely available on the web )
  - Document encoding: TEI Lite
  - One to four paragraphs per image

# Image Indexer's Workbench

CLiMB Cataloger Workbench: The Flight into Egypt

Export View Help

Image Image Information



**Text**

Mary and Joseph's flight to protect Jesus from Herod's slaughter of Hebrew babies is recounted by Matthew in the Bible. The subject often decorated predellas (small scenes at the base of altarpieces), but this painting is too large to be a predella panel. It most likely was not the central section of an altarpiece either, since those were usually meditative, devotional images rather than narrative ones like this. Perhaps it was made for a religious confraternity. Such scuole were among the most important patrons of Venetian painters. Their commissions accounted for Carpaccio's best-known works -- large, bustling scenes that are full of detail and provide valuable information about life in renaissance Venice. Here, the distant village and covered boat gliding past offer a hint of Carpaccio's delight in storytelling. While Bellini began to use layered oil glazes to soften the edges of his forms, Carpaccio continued to favor a harder (and increasingly old-fashioned) use of line, which in this case enhanced his narrative purpose. Hard contours accentuate the gait of the ass and the long stride of Joseph, and they help frame the Virgin and Child in a way that almost enthrones them on their humble mount. In contrast, the luminous undersides of the clouds reveal the influence of Bellini's treatment of light.

**Subject Terms**

**Term Under Consideration**

panel

**Terms Assigned**

Term	Normalized Term	Subject ID
predella	predellas(visual works components, components by specific context, ... Top of the AAT hi...	AAT:300003745
narrative	narrative(artistical devices, artistic concepts, ... Top of the AAT hierarchies)	AAT:300055903
village	villages(settlements by form: scale, settlements by form, ... Top of the AAT hierarchies)	AAT:300008372

**AAT Browser(6)** TGN Browser(3) ULAN Browser(25)

Search Term: panel  Show Partial Match

- panel [size: card photographs]
- panel paintings [paintings by form]
- panels [ornament areas]
- panels [wood by form or function]
- panels [surface element components]
- panels [costume components]

**Selected Record Description**

Refers generally to paintings on a wooden support, including smaller portable paintings and medium-sized paintings, such as altarpieces, for which several planks of wood were joined to form a larger panel. The term is typically used to refer specifically to paintings on a wooden support in Western art, generally dating from ancient Greece and Rome through the Renaissance, after which time canvas became the standard support for paintings in this size range. Panel paintings are still common today in Greek and Russian Orthodox icons.

**Selected Record Hierarchy**

Top of the AAT hierarchies

- Objects Facet
  - Visual and Verbal Communication
    - Visual Works
      - visual works
        - visual works by medium or technique
          - paintings
            - paintings by form
              - panel paintings

Show Full Display

Legend: Selected Term Terms With Children

# Example



Ram and Tree. Offering stand from Ur. c. 2600 B.C.

A far more realistic style is found in Sumerian sculpture . . . put together from varied substances such as wood, gold leaf, and lapis lazuli. Some assemblages . . . roughly contemporary with the Tell Asmar figures, have been found in the tombs at Ur . . . including the fascinating object shown in an offering stand in the shape of a ram rearing up against a flowering tree.

<p>

<semcat type="implementation">. . . substances such as wood, gold leaf, and lapis . </semcat>

<semcat type="historical\_context"> . . . contemporary with the Tell Asmar figures . . . </semcat>

<semcat type="image\_content"> . . . offering stand in the shape of a ram rearing up against a flowering tree.</semcat> . . .</p>

# Motivation



Ram and Tree. Offering stand from Ur. c. 2600 B.C.

- Allow indexer's to choose what type of metadata to look for
  - Add descriptors about the work
  - Add descriptors about provenance
- Allow end user's to constrain the semantics of a search term
  - OF: Tell Asmar figures
  - Same Period: Tell Asmar figures

# Functional Semantic Categories

Category Label	Rough Description
<b>Image Content</b>	<b>Describes the appearance or other objective features of the depicted object</b>
Interpretation	The author provides his or her interpretation of the work
<b>Implementation</b>	<b>Explains artistic methods/materials used in the work, including style, techniques</b>
Comparison	Comparison to another art work in order to make/develop an art historical claim
Biographic	Information about the artist, patron or other people involved in creating the work
<b>Historical Context</b>	<b>Description of historical, social, cultural context</b>
Significance	Explanation of art historical significance

# Table of Results from Pilot Annotations

Exp	Dataset	#Labels	#Anns	Alpha (MASI)
1	I: 13 images, 52 paragraphs	any	2	0.76
2	II: 9 images, 24 paragraphs	any	2	0.93
3	II: (ditto)	two	5	0.46
4a	III: 10 images, 24 paragraphs	one	7	0.24
4b	III: 10 images, 159 sentences	one	7	0.30

- Comparable range to previous work

# Summary of IA Results

- Semi-controlled study
  - IA decreases when restricted to one label per item
  - IA decreases with more annotators
- Pairwise IA for experiments varied widely
  - For 4a, 0.46 to -0.10 (7 annotators)
  - For 4b, same range
- IA varied greatly with the image/text unit
  - High of 0.40 for 7 annotators in 4a (units 1, 9)
  - Low of 0.02 for 7 annotators in 4a (unit 5)



# Conclusions from Pilot Annotation Experiments

To optimize annotation quality for our large scale effort (50-75 images and 600-900 sentences):

- Allow multiple labels
- [Develop annotation interface](#) (with online training)
- Use many annotators, post-select the highest quality annotations
- Partition the data in many ways

# Specific Questions

- Does ML performance correlate with IA among X annotators on class labels?
  - Compute IA for each class
  - Rank the X classes
- Does ML performance correlate with IA across Y annotators on a given class?
  - Compute Y-1 pairwise IA values for each annotator
  - Rank the Y annotators
  - *Swap in* each next annotator's labels

# Data

- Three binary classifications, 1A per class
  - Historical Context: 0.39
  - Image Content: 0.21
  - Implementation: 0.19
- Training data: 100 paragraphs labeled by D
- Test data: Single label per annotator
  - 24 paragraphs labeled by six remaining annotators in Exp 4
  - 6 paragraphs labeled by two annotators in Exp 2

# Annotators' Average Pairwise IA, for all FSC labels

Annotator	Avg. Pairwise IA (sd)	IA Year 1, Year 2
A	0.32 (0.12)	
A'	0.31 (0.10)	0.34
A''	0.28 (0.13)	
B	0.21 (0.15)	0.88
C	0.17 (0.11)	
D	0.14 (0.14)	
E	0.10 (0.16)	

# Machine Learning

- Naïve bayes, binary classifiers
  - Performs better than multinomial NB on small datasets
  - Performs well when independence assumption is violated
- Three feature sets
  - Bag-of-words (BOW)
  - Part-of-speech (POS): 4-level backoff tagger
  - Both

# Annotator *Swap* Experiments

- For each classifier *and* for each feature set
  - Disjoint training/testing data
    - Train on same 100 paragraphs, annotated by D
    - Test by swapping in annotations of 24 paragraphs by A, A', A'', B, C, E (plus the 6 paragraph training set)
  - 10-fold cross validation on 130 paragraphs
    - For the 24 paragraph set, swap in each next annotator
- Correlate:
  - Average ML performance on 3 classes with per-class IA
  - Individual learning runs with individual annotators

# Average ML per Condition Correlates with per-Class IA

- 6 runs X 3 feature sets X 2 evaluation paradigms
- Average learning performance correlates with IA among 6 annotators on bow and both, not on pos

	Train 100/Test 30			10-Fold Crossval 130		
	bow	pos	both	bow	pos	both
Historical Cont.	0.71	0.68	0.71	0.75	0.69	0.77
Image Content	0.57	0.72	0.57	0.63	0.69	0.63
Implementation	0.59	0.44	0.59	0.60	0.59	0.60
<b>Correlation</b>	<b>0.98</b>	0.46	<b>0.98</b>	<b>1.00</b>	0.58	<b>1.00</b>

# Individual ML Runs do not Correlate with Annotator Rank

Train100/Test30					
Historical Context		Image Content		Implementation	
bow	0.05	bow	-0.25	bow	-0.43
pos	0.18	pos	-0.75	pos	-0.01
both	0.59	both	0.42	both	-0.43
Crossval 130					
bow	0.11	bow	-0.06	bow	-0.77
pos	-0.87	pos	0.07	pos	0.46
both	0.71	both	0.14	both	-0.87



# Details:

## Individual Annotators/ML Runs

- Annotator A
  - Highest ranked annotator
  - Often the low(est) ML performance
- Annotator B
  - Mid-ranked
  - Often near top ML for Image Content and Implementation
- Annotator E
  - Lowest ranked annotator
  - Occasionally has highest ranked runs

# Details: Feature Sets

- BOW: high dimensionality, low generality
- POS: low dimensionality, high generality
- Whether BOW/POS/Both does well depends on
  - Which classifier
  - Which annotator's data
- POS > BOW for Image Content on average
- BOW > POS for Historical Context on average

# Conclusions

- We need to repeat experiment on larger dataset
- Semantic annotation requirements
  - No *a priori* best IA threshold
  - More qualitative analysis of label distributions
- ML correlated with per-class IA
- ML did not correlate with individuals' IA

# Discussion

- When using human labeled data for learning:
  - Data from a single annotator with high IA does not guarantee good learning data
  - Data from an annotator with poor IA does not guarantee the data is not good learning data
  - Different annotations may lead to different feature sets
- Learners should learn what a range of annotators do, not what one annotator does

# Current and Future Work

- Large-scale annotation effort: 5 annotators
  - Done: 50 images/600 sentences from two texts, same time period (Ancient Egypt)
  - To do: 50 images/600 sentences from two new time periods (Early Medieval Europe; other)
- Redo annotator swap experiment on larger datasets
- Multilabel learning
- Learning from multiple annotators
- Feature selection