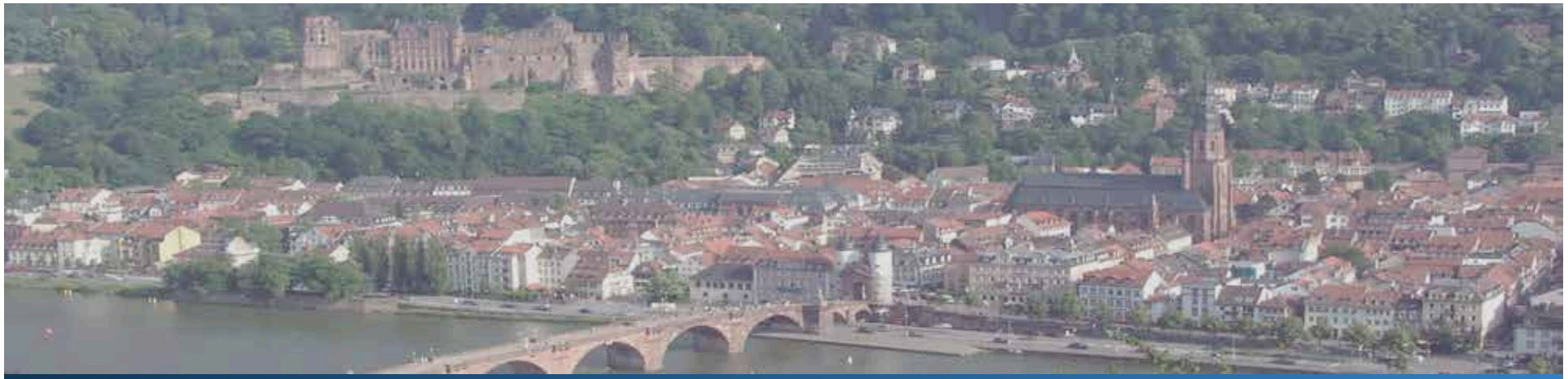
Image2Content: Visuelle Objekterkennung zum Datamining in den Geisteswissenschaften und die Bedeutung von Crowdsourcing

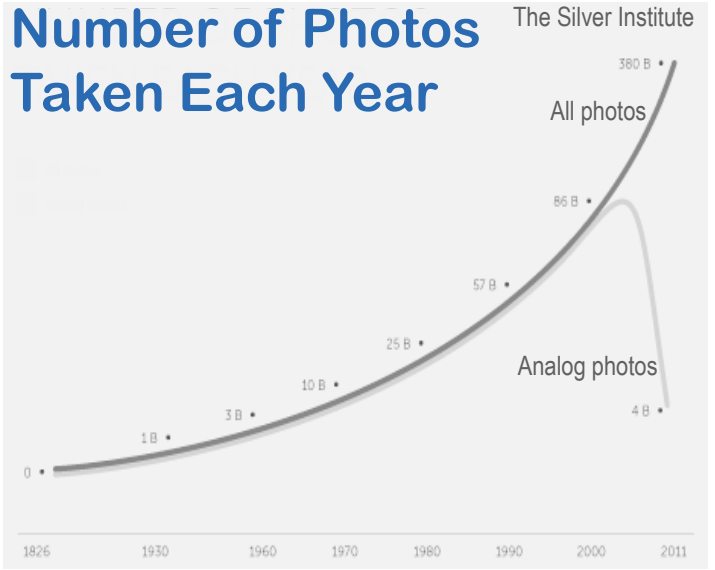
Björn Ommer

Computer Vision Group

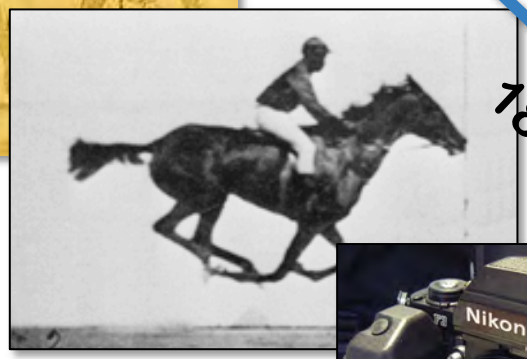
University of Heidelberg



Our World is Visual



1825



1872

1990

1996

1st digital consumer SLR

"When I took office, only high energy physicists had ever heard of what is called the World Wide Web... Now even my cat has it's own page." *- Bill Clinton*

>100bn images on facebook, **++6bn** images/month

72 hours of video uploaded to YouTube every minute

gettyimages
ARTSTOR
Picasa
HeidICON Die Heidelberger Bilddatenbank
facebook
flickr
YouTube

2012

Finding Relevant Content

- **Large-scale digitization in the Arts & Humanities:**

prometheus
Das verteilte digitale Bildarchiv für Forschung & Lehre **~1.7M images**

Deutsches Dokumentationszentrum für Kunstgeschichte
Bildarchiv Foto Marburg **~1M images**

artstor IMAGES **~108M images of artworks**

- **„We are drowning in information and starved for knowledge“, *John Naisbitt: Megatrends***

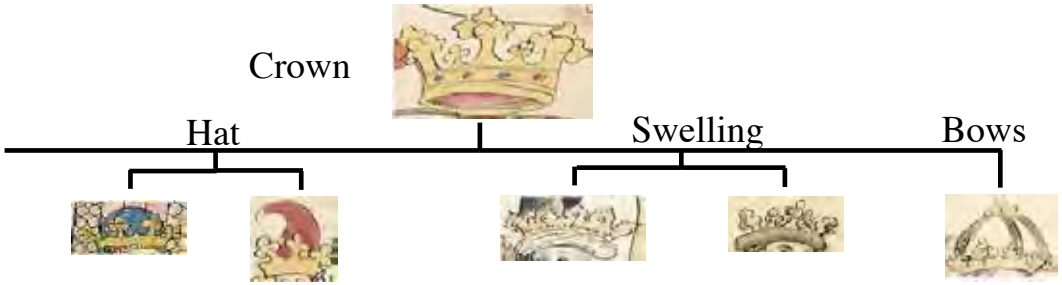
Analysis of Large Pre-Modern Image Datasets

Large scale digitization in the Humanities: Pibliotheca Palatina:

- **~270 000 pages**
- **~7 000 miniatures**

Textual annotations provided, but… NO labeling w.r.t.

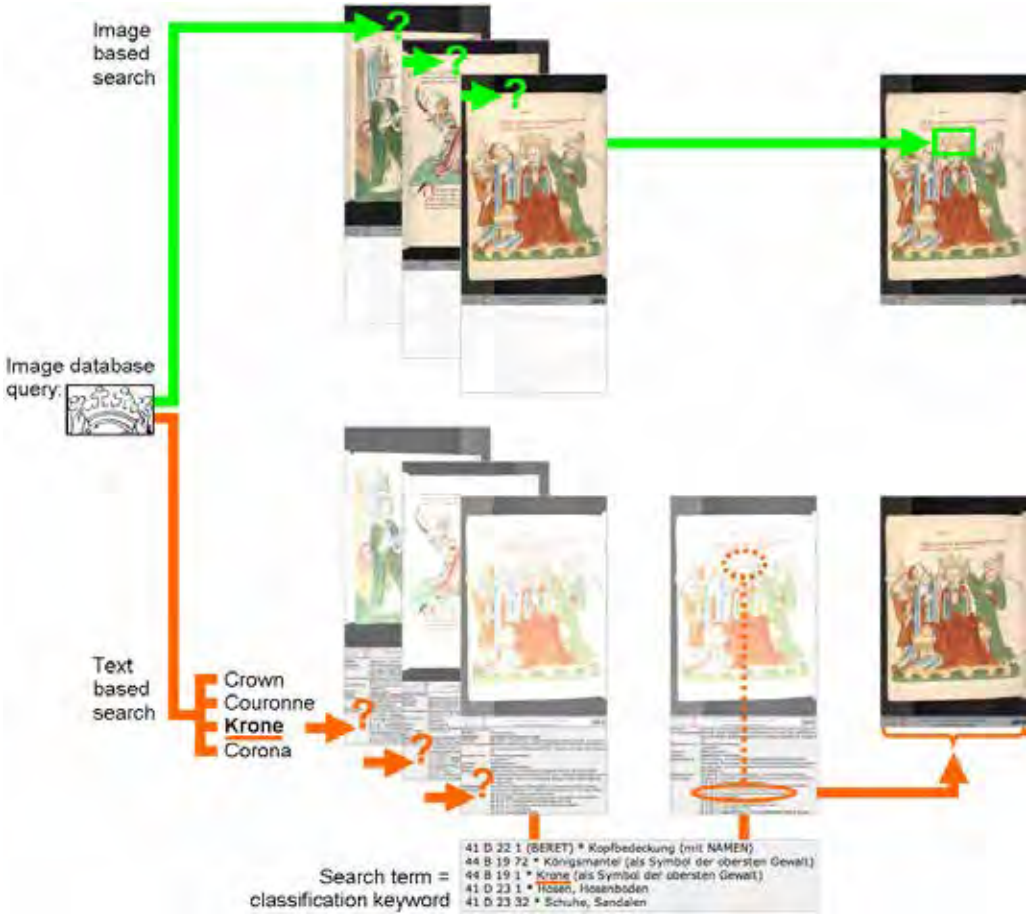
- **object locations**
- **relations betw. objects within images & betw. Images**
- **hierarchical nature of categories**

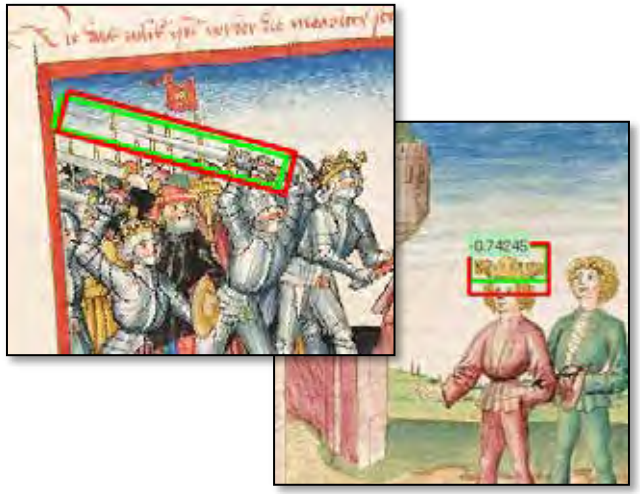


Symbols of power: >2.5K images related to crowns in Cod. Pal. Germ. ⇨ concentration game with > 2.5K cards!

Image Retrieval vs. Text-Based Search

- Object retrieval in images: search through images NOT through textual annotations

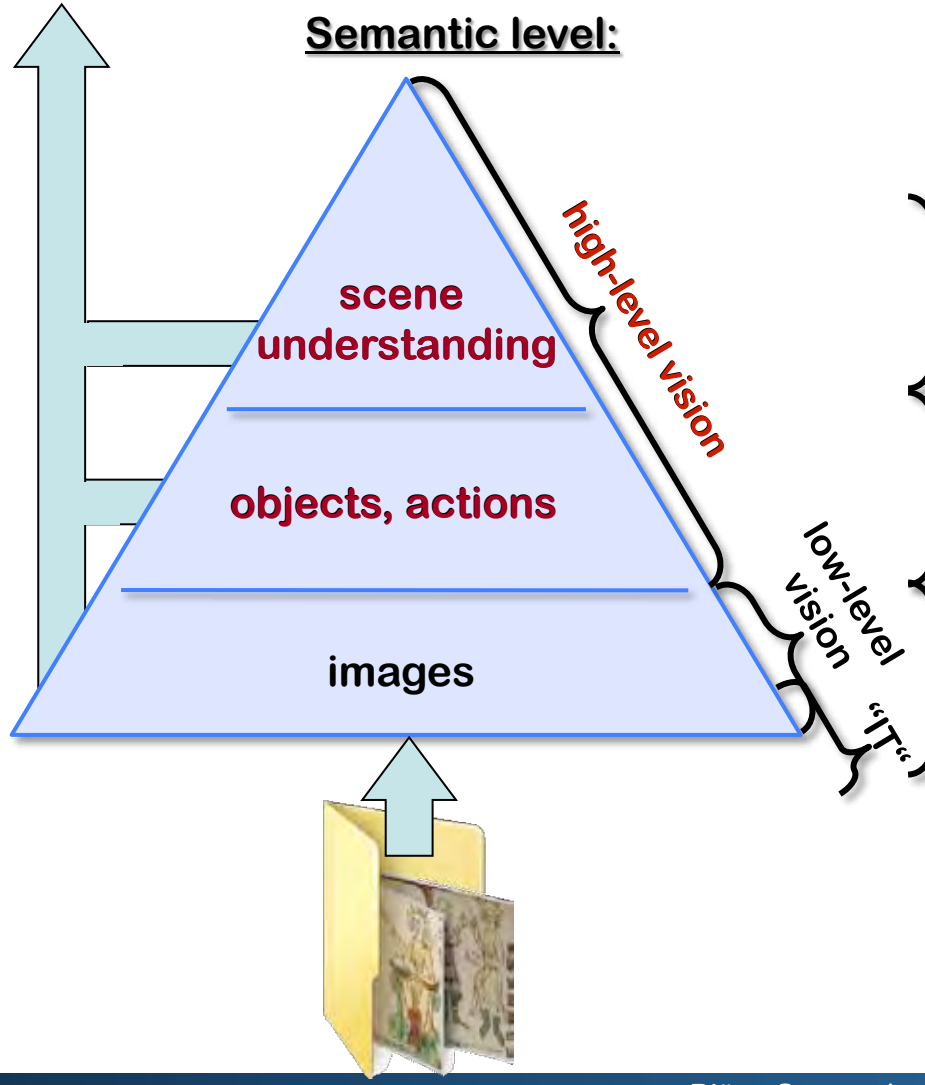




[Yarlagadda et al., ACCV'10 e-Heritage]

Computer Vision and the Humanities

Humanities

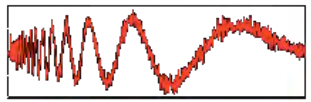


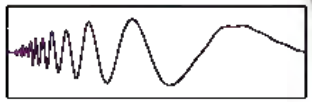
Efforts so far:

interactions, context

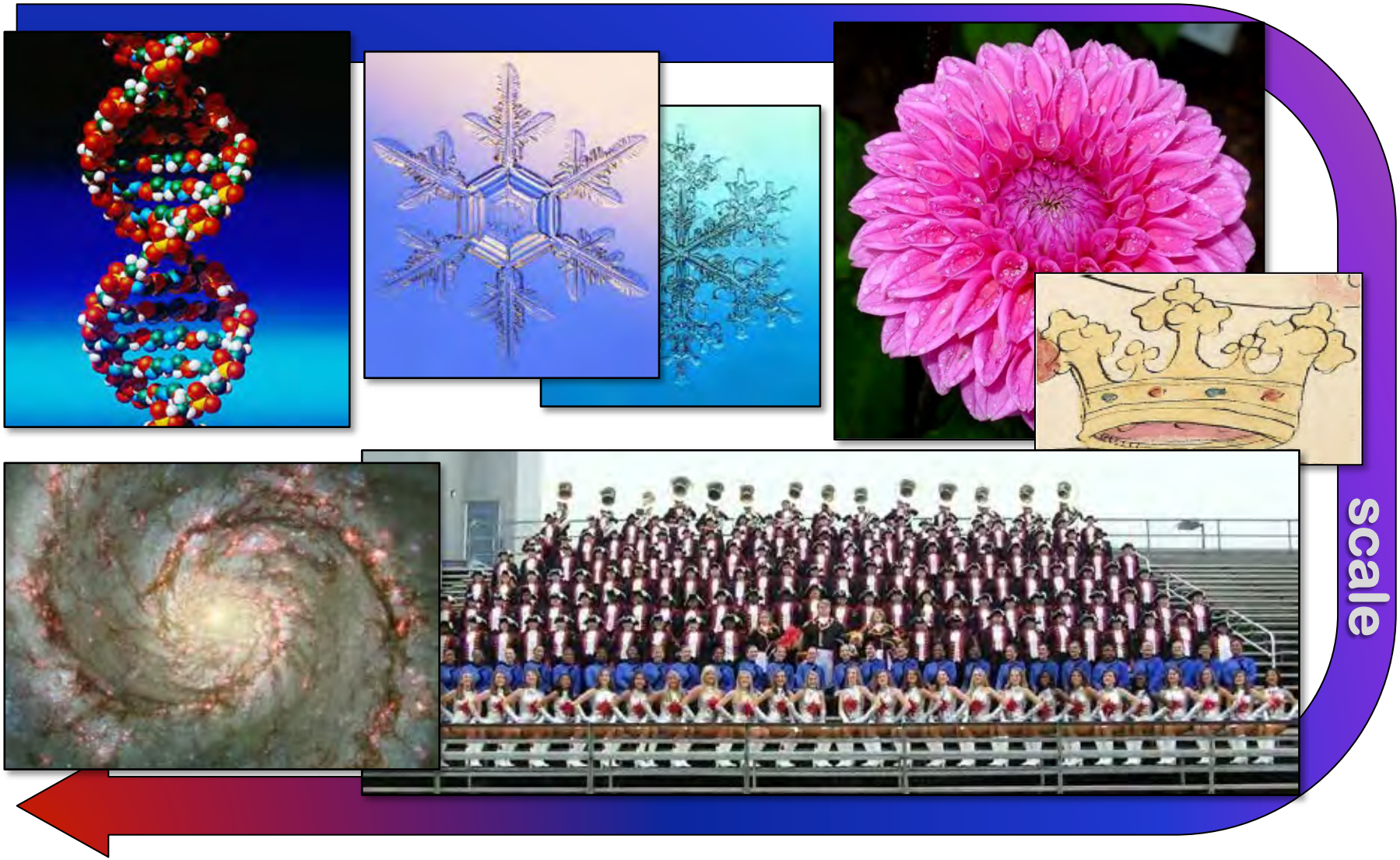
object & action recognition,
object statistics & analysis

digitization, signal processing, image databases, manual annotation, …





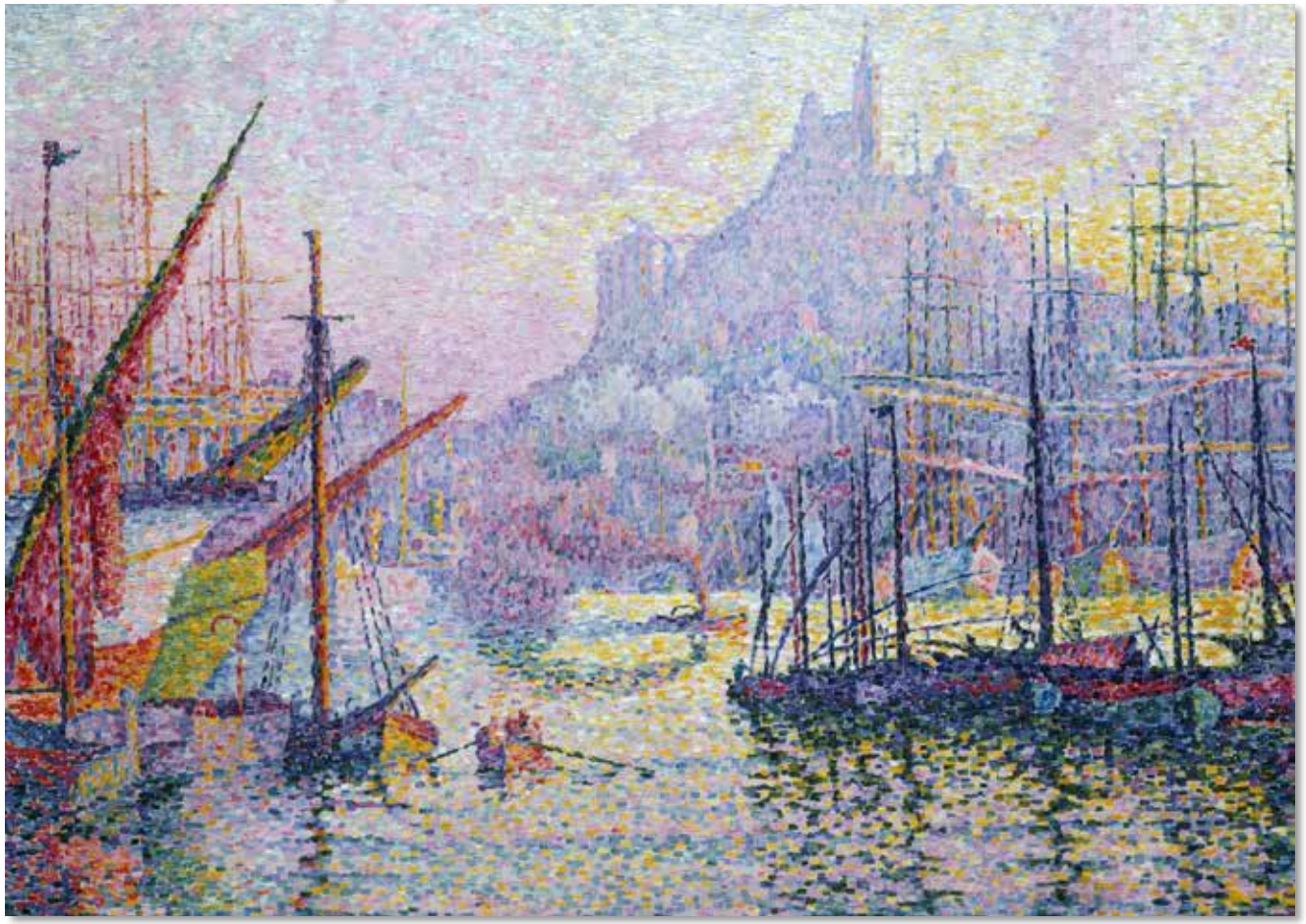
Our (Visual) World is Highly Structured



Patterns, Patterns, Everywhere ⇨ Machine Learning

Recognition >> Observing Pixels: The Semantic Gap

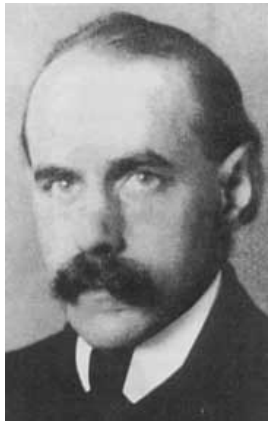
?



"I stand at the window and see a house, trees, sky. Theoretically I might say there were 327 brightnesses and nuances of colour. Do I have "327"? No. I have sky, house, and trees." **--Max Wertheimer**

Representing Structure

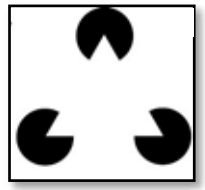
- Structure, esp. shape, is an emergent property
  ⇨ cannot be observed locally
- How can we represent what cannot be measured (directly)?



Max Wertheimer

⇨Perceptual Grouping

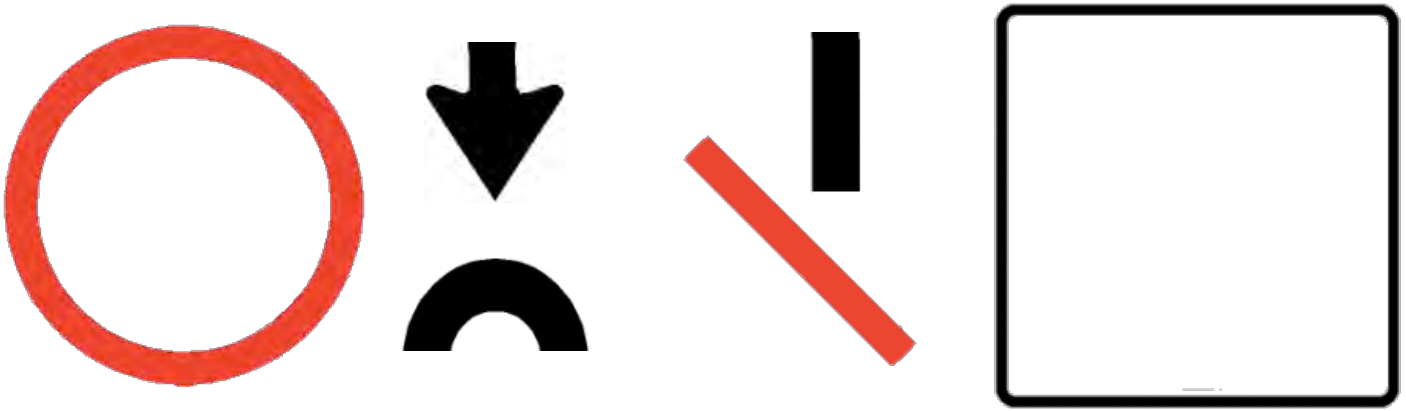
- Bottom-up grouping using relationships between perceptual entities
- Top-down grouping using prior know-ledge



⇨Compositional hierarchy bridges seman-tic gap btw. parts & whole object

Our Approach: Compositionality

- Simple, widely reusable parts & relations between them ⇨ Compositions

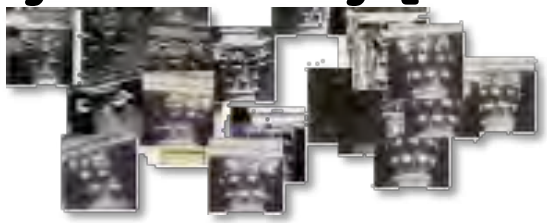


[Ommer et al., PAMI'10]

Our Ongoing Projects in the Humanities

1. Object detection [Cod. Pal. germ.]

2. Analysis of object category variability [CPG]

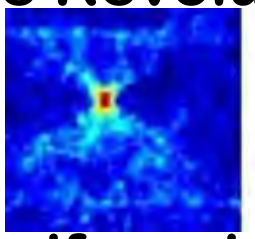


3. Architectural analysis

4. Registration of reproductions [Cod. Manesse]

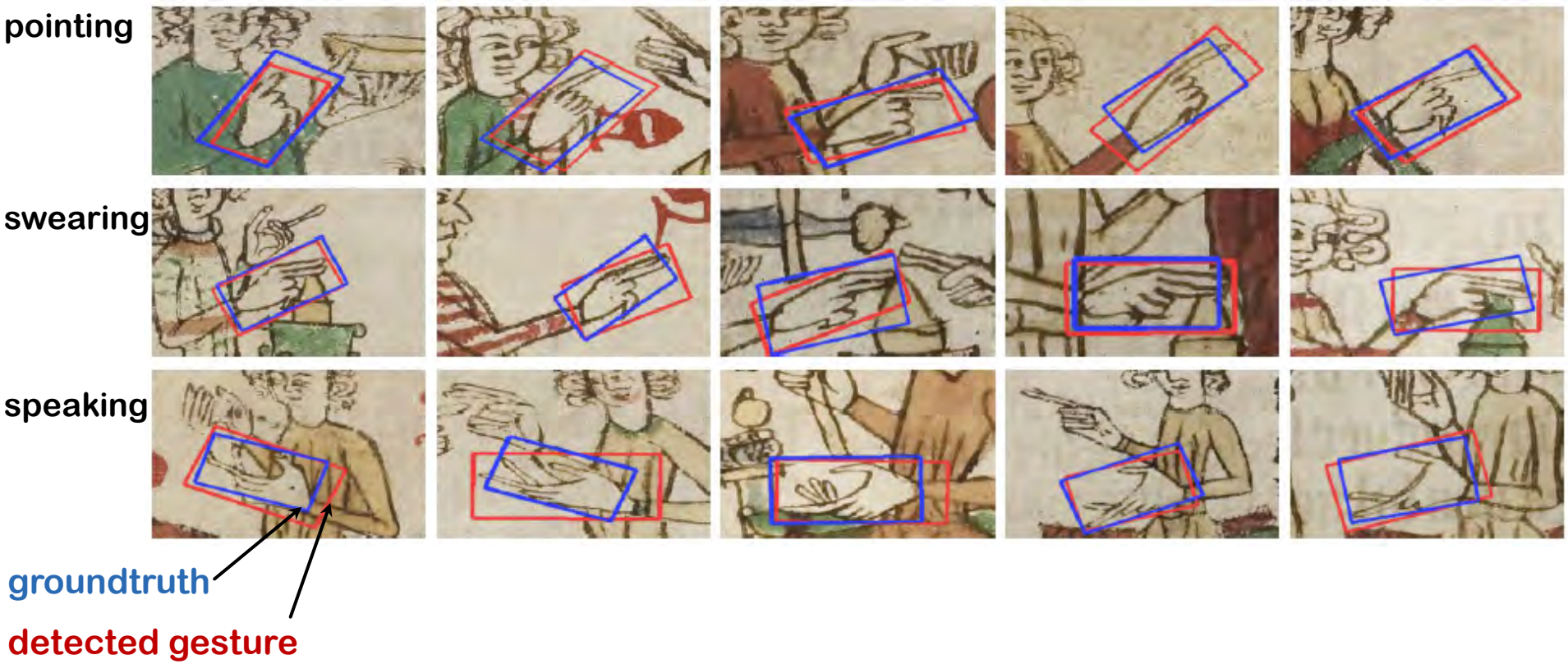
5. Gesture recognition [Sachsenspiegel]

6. Iconographic analysis [Chinese Revolution Comics]



7. Analysis of ancient script [Cuneiform inscriptions]

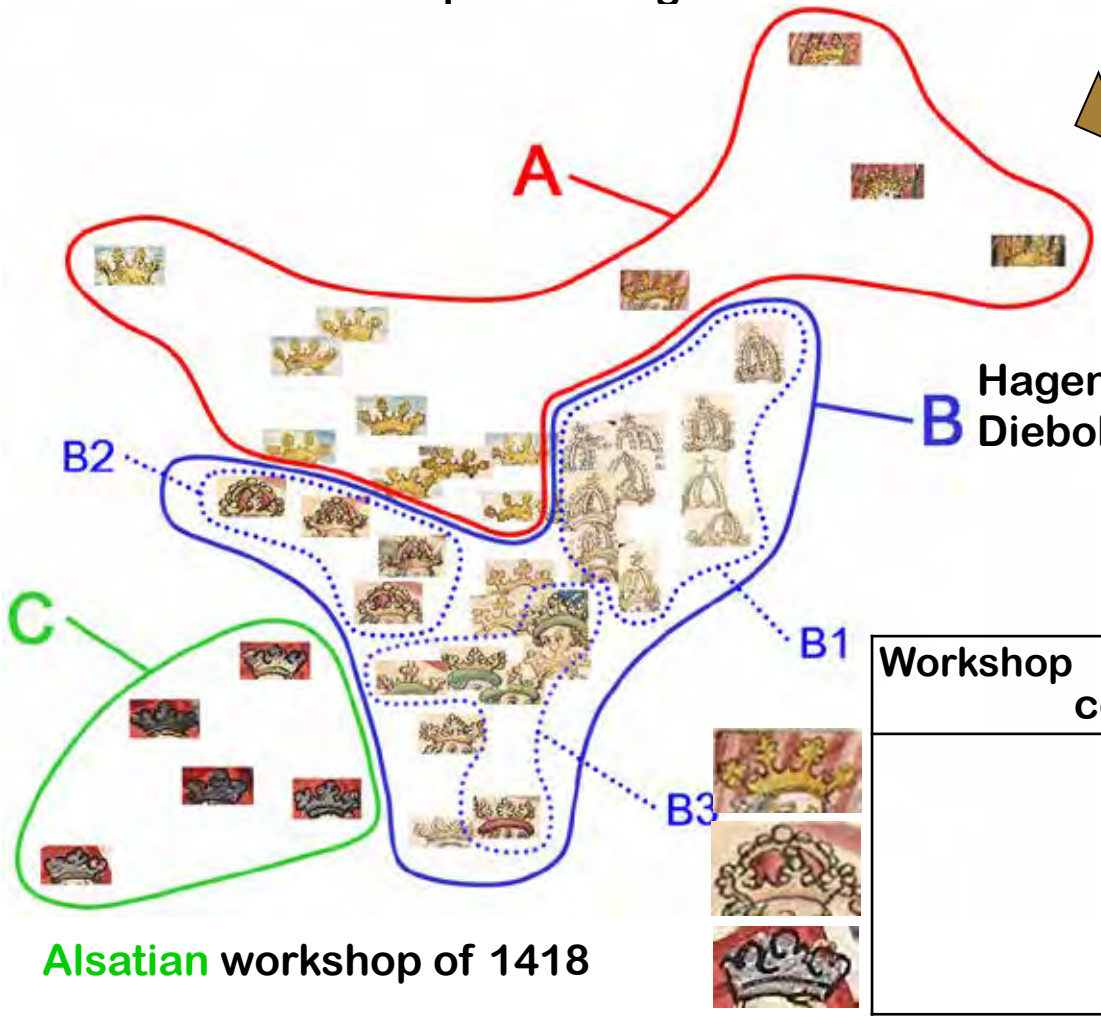
Gesture Recognition - Sachsenspiegel



[Schlecht, Carque, Ommer, ICIP'11]

Intra-Category Variability of Crowns

Swabian workshop of Ludwig Henfflin



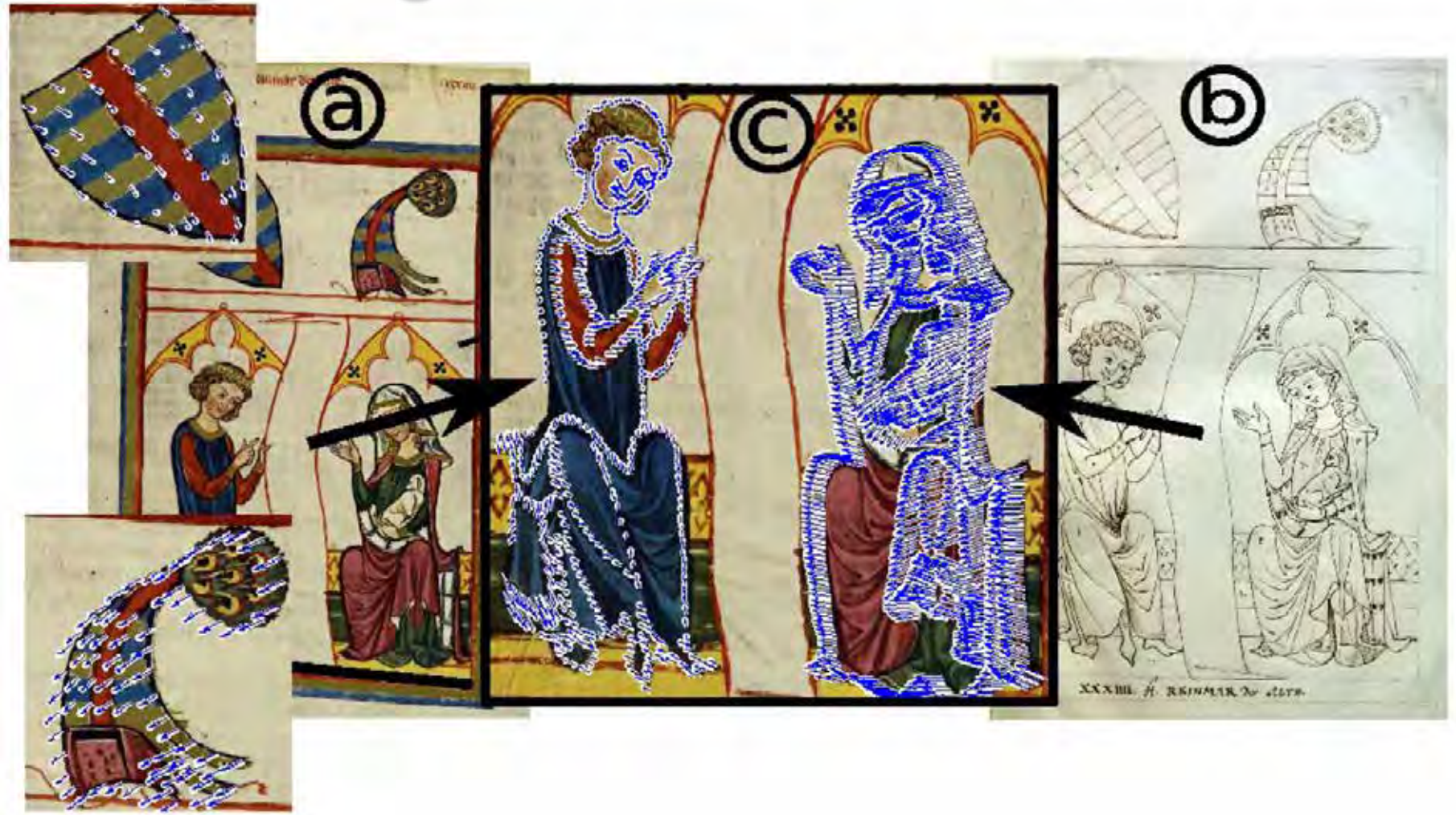
Hagenau workshop of Diebold Lauber.

Alsatian workshop of 1418

Workshop pred: / corr.:	Henfflin	Lauber	Alsatian
Henfflin	98.4%	1.6%	0%
Lauber	3.7%	96.3%	0%
Alsatian	0.8%	0.8%	98.3%

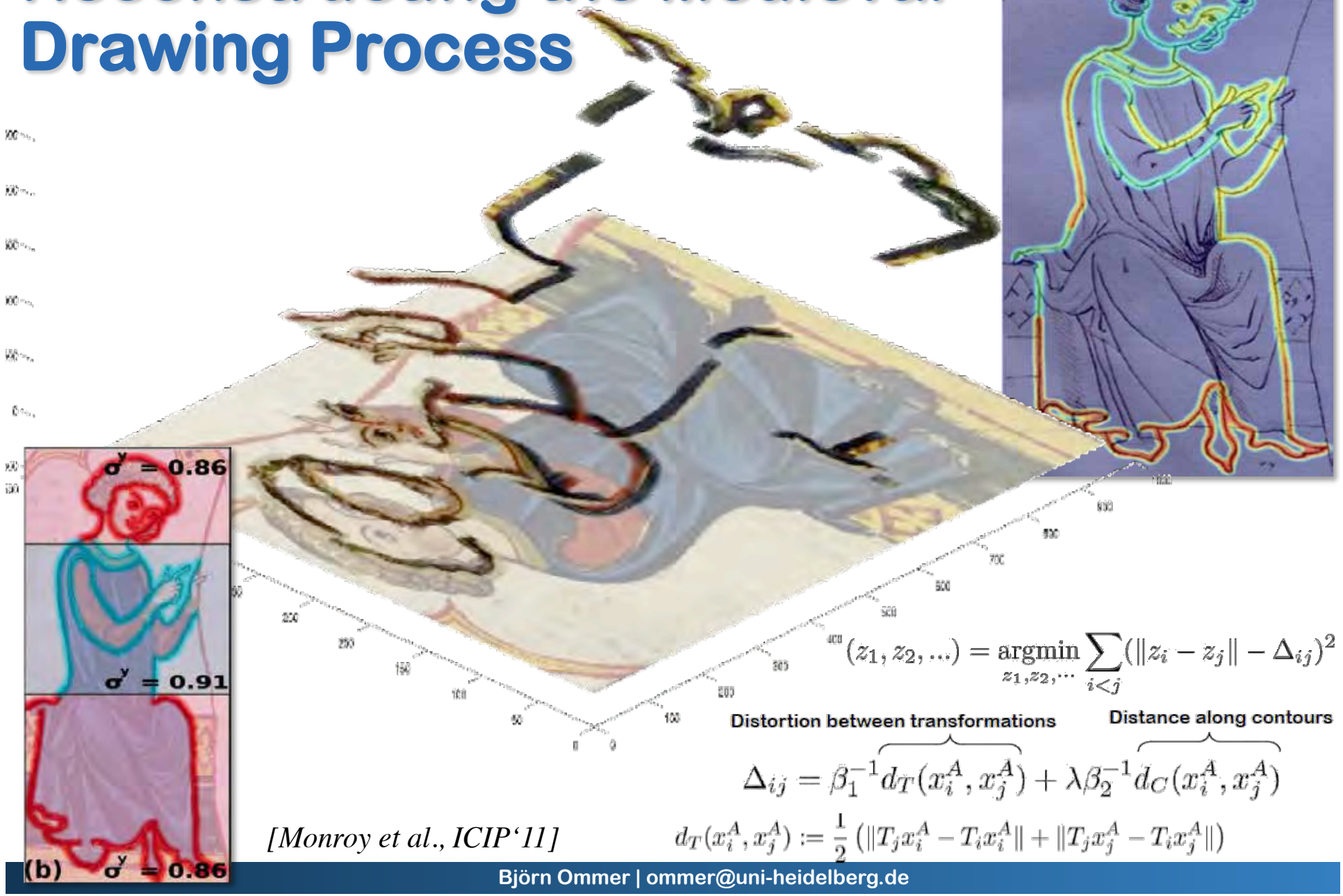
[Yarlagadda et al., ACCV'10 (eHeritage)]

Image Registration – Codex Manesse



[Monroy, Carque, Ommer, ICIP'11]

Reconstructing the Medieval Drawing Process



$$(z_1, z_2, ...) = \underset{z_1, z_2, \cdots}{\operatorname{argmin}} \sum_{i<j} (\|z_i - z_j\| - \Delta_{ij})^2$$

Distortion between transformations — Distance along contours

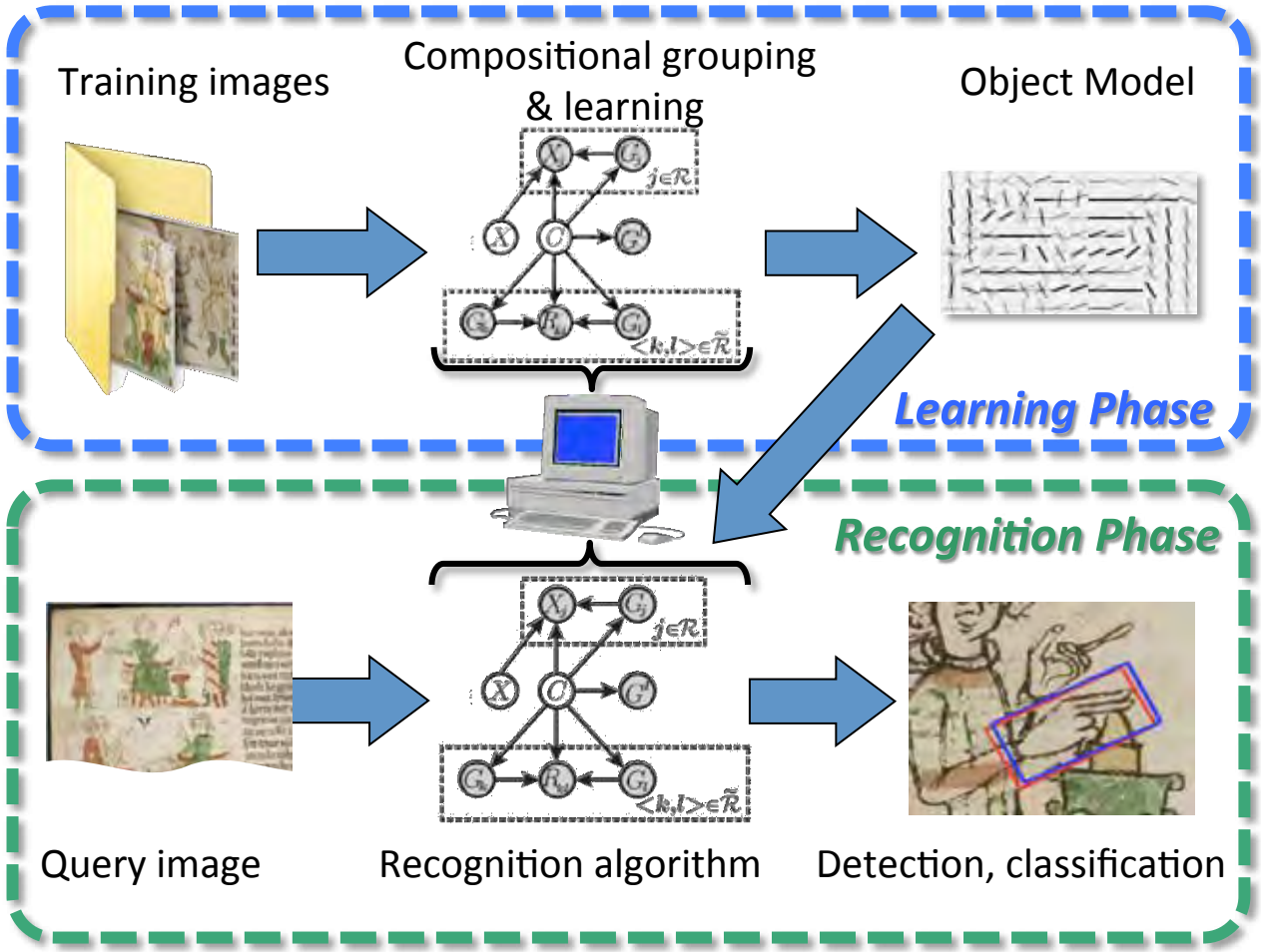
$$\Delta_{ij} = \beta_1^{-1} \overbrace{d_T(x_i^A, x_j^A)} + \lambda \beta_2^{-1} \overbrace{d_C(x_i^A, x_j^A)}$$

$$d_T(x_i^A, x_j^A) := \frac{1}{2} \left(\|T_j x_i^A - T_i x_i^A\| + \|T_j x_j^A - T_i x_j^A\| \right)$$

[Monroy et al., ICIP'11]

Compositional Object Recognition – Learning Object Structure from Samples

Our (visual) world is highly structured:



[Ommer & Buhmann, PAMI'10]

Dataset Annotation vs. Model Learning

- *Give a man a fish and you feed him for a day.*
  Teach a man to fish and you feed him for a lifetime.

➪ Annotate subset of data to train recognition alg. and have computer label additional data

- Benefit of training recognition algorithm:
  - Automatic generalization of training labels to whole dataset ➪efficiency
  - Learned object models yield an abstraction that can be verified on novel data ➪label consistency
  - Generalizes to non-categorial annotations (relational data)
  - …

Learning Object Models

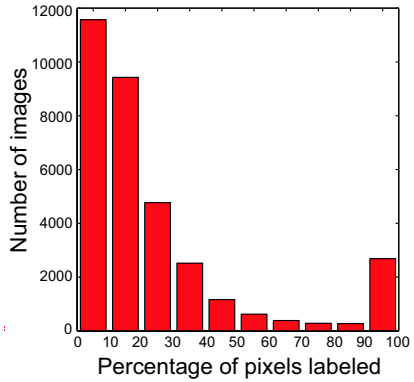
- Level of supervision

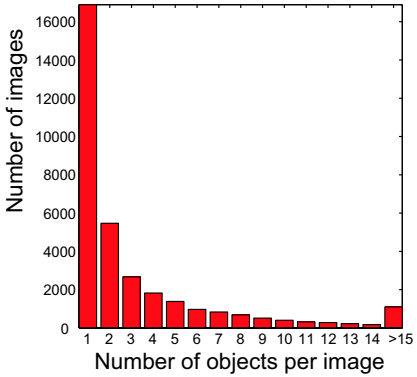



- Number of training samples


Google images

Crowdsourcing





- **Many independent annotators:**
  - **Effective means for obtaining large amounts of training data**
  - **Inherent check for consistency**

Combining Crowdsourcing with Object Recognition

- Recognition alg. supports annotators:
  - Suggests related images, object localization, labels, …
  - Generalizes annotation to novel images
  - Alleviate simple tasks ➪ automation
- Continuous annotation supports recognition alg.:
  - Previously learned models are verified in consecutive rounds
  - Provides large amounts of training data
- Combination of man/machine enables novel games:
  - Have users deal with large numbers of images simultaneously, e.g. relations btw. Images / artistic reproductions