BIGDIG 2017

Argonne NATIONAL LABORATORY

DESIGNING A HIGH-THROUGHPUT PIPELINE FOR DIGITIZING PINNED INSECTS

MARK HERELD
Senior Experimental Systems Engineer

October 24, 2017
Auckland, New Zealand

THE PROBLEM

Argonne NATIONAL LABORATORY

WHAT ARE WE TRYING TO DO?

- Overview
  - Collections
  - Size
  - Goals
- Role of computation
- Overwhelming variety
- Complexity of imaging problem
- Need for speed


EXIT

THE COLLECTION

Overview in Round Numbers

- 4.5 million insects
  - As small as a millimeter
  - As large as 500 mm
- 15 thousand drawers
  - Filled with standard unit trays
- 1.6 seconds per specimen on average
  - 1 year, 2000 hours


Row 58
COLEOPTERA: Polyphaga
Staphyliniformia
Staphylinoidea
Staphylinidae
Trigonurinae
Apateticinae
Scaphidiinae
Piestinae
Osorinae
Oxytelinae

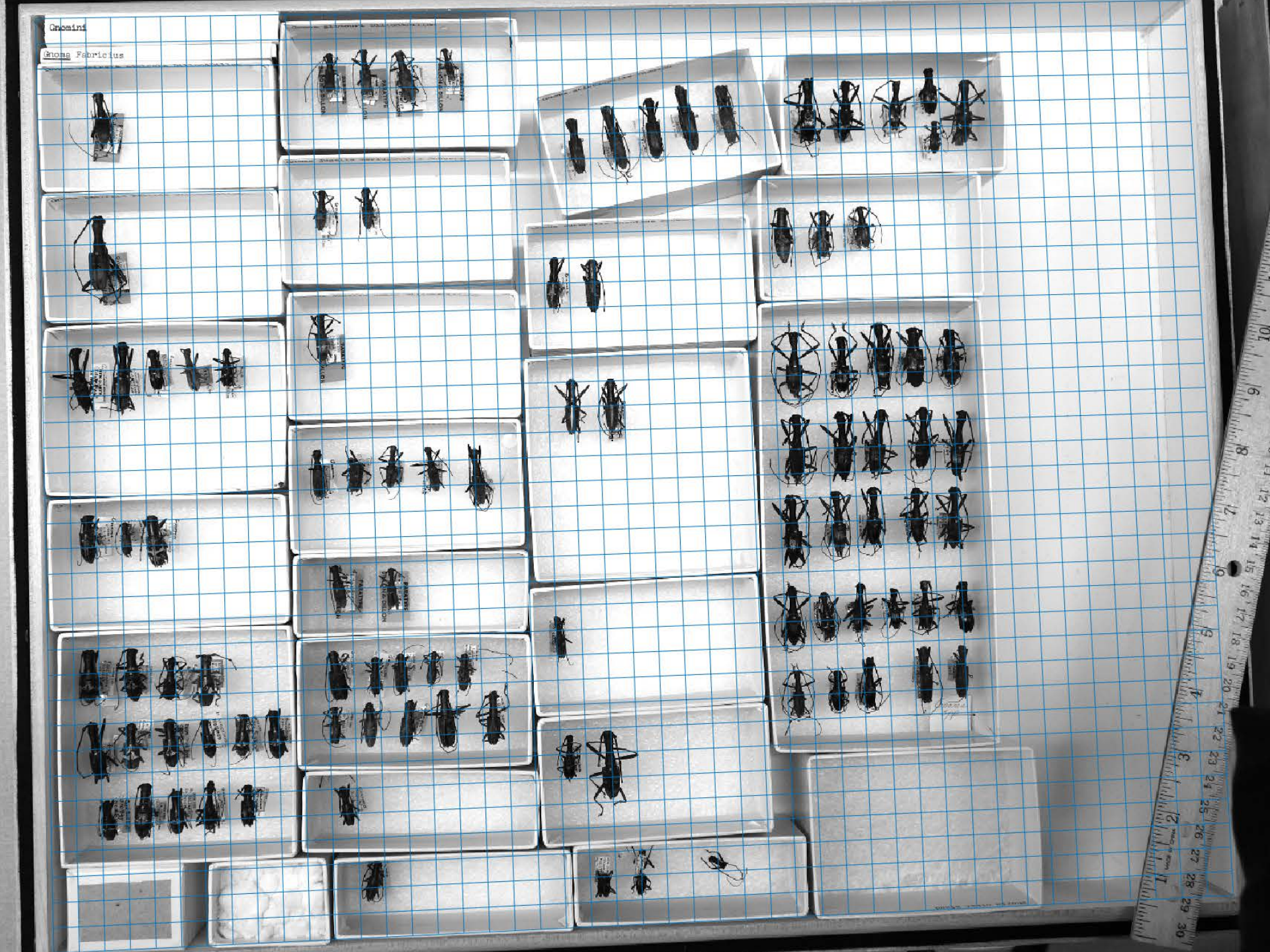
RECONNAISSANCE

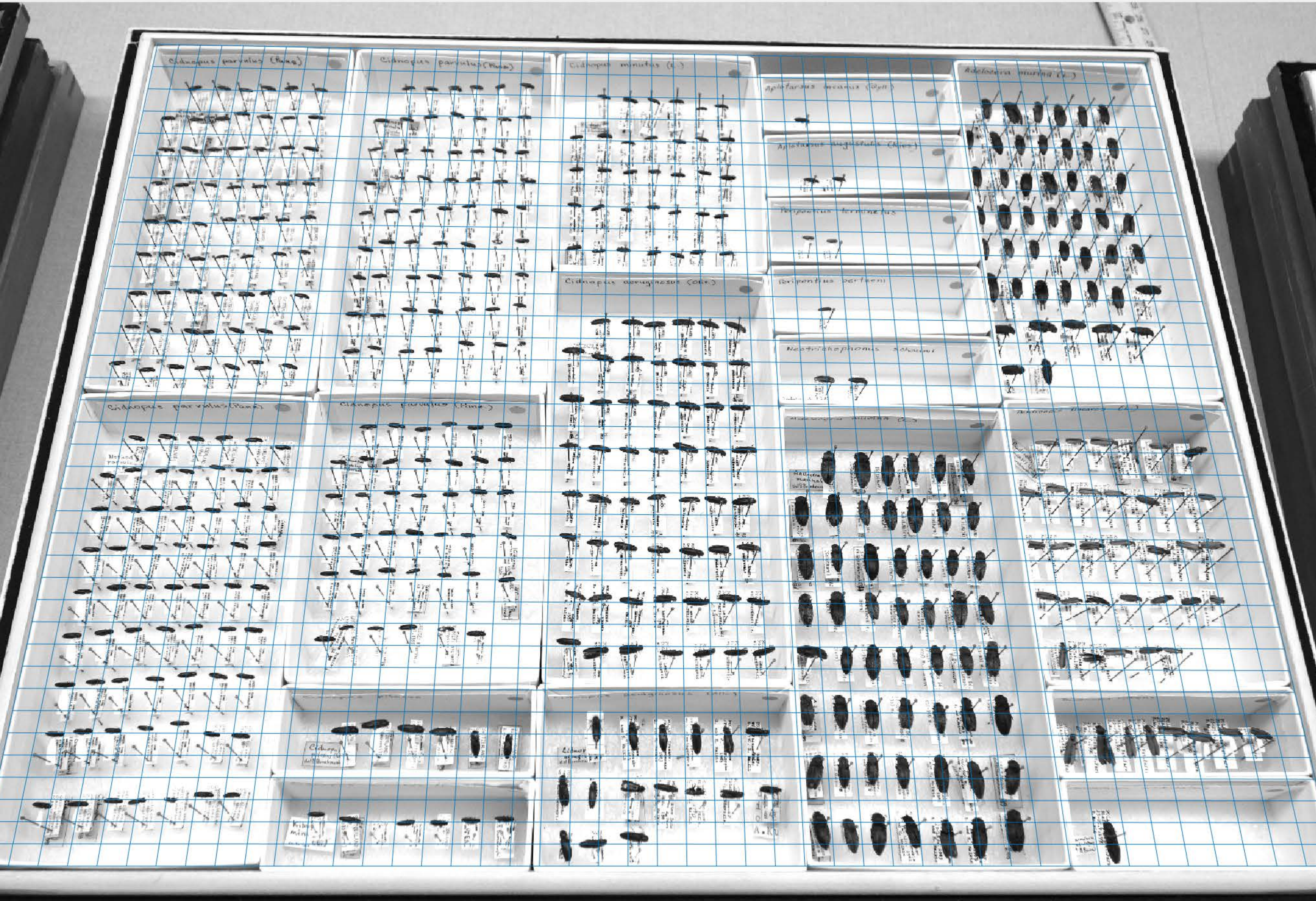
- How many insects per drawer?
- How are they organized?
- What is the range of insect sizes?
- How many labels per pin?
- Handwritten or typed?
- How close are labels to one another? … to the insect body?
- How often are there additional labels in unit trays?

To understand:

- Sample volume requirements for instrument design
- Handling needs
- Project priorities

OVERWHELMING VARIETY


Cnemini
Gnoma Fabricius


Cidnopus parvulus (Panz.)
Cidnopus parvulus (Panz.)
Cidnopus minutus (L.)
Aplotarsus incanus (Gyll.)
Cidnopus aeruginosus (Ol.)
Neotrichophorus schaumi
Cidnopus parvulus (Panz.)
Cidnopus parvulus (Panz.)


Eutin
Brandenburg
Palaearctic Colln.

Volo,marsh S of
Sayer Bog 24-x-80
Watrous grassy l.
WISCONSIN: Kenosha
Co.: Chiwaukee Prairie,
Carol Beach. Leg. W.
Suter. 10:V:1975.
19:IV:1976. leg. W. Suter.
WS#76-11:b,c,d,e
Harris Tract
1980 LEWatrous
leaf litter
vernal pond
12-iii-1981
FM(HD)#81-129
L.E. Watrous
WISC.; Kenosha Co.;
2.V.1965
9.VII.1959
25 March 1968
Cook Co.,
W. Suter leg.
Palos Hills, 25-IV-
1981,
Co., Silver Springs
St. Pk., 26-IV-1981,
Watrous

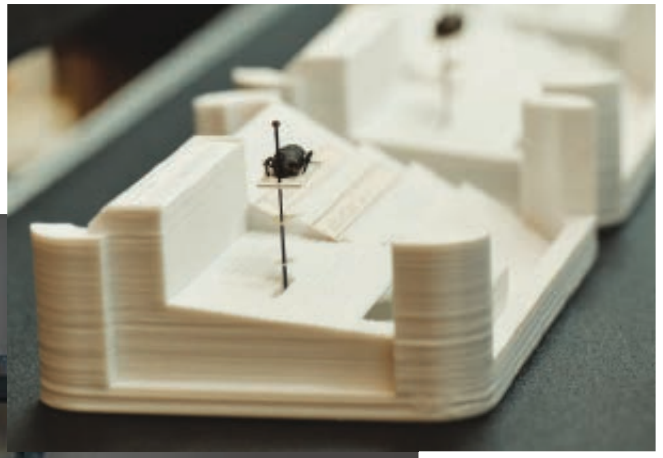

2.5 mi.SW of
Forestburg, Mont.
TEX
Col. & pres. by
York Harbor

NEED FOR SPEED

RELATED WORKS

AUTOMATION: DIGITARIUM.FI

Pinned insect imaging and label digitization



"Using the insect line two operators can image in one working day about 500 pinned insects and their labels."

Tegelberg, Riitta, Tero Mononen, and Hannu Saarenmaa. "High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens." *Taxon* 63.6 (2014): 1307-1313.

Argonne NATIONAL LABORATORY

Nguyen, Chuong V., et al. "Capturing natural-colour 3D models of insects for species discovery and diagnostics." *PloS one* 9.4 (2014): e94346.

Argonne NATIONAL LABORATORY

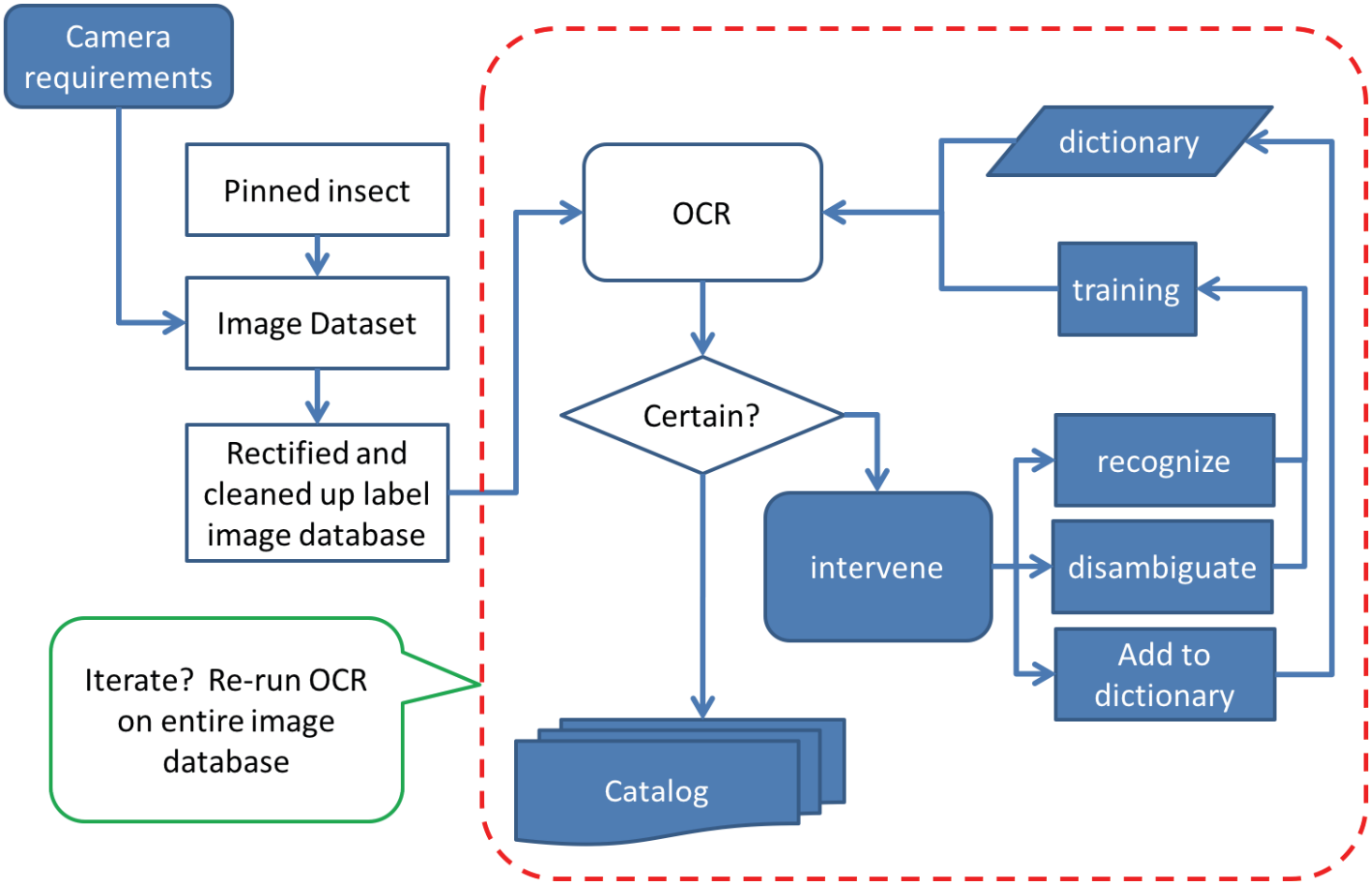
OUR APPROACH

CHARACTERIZE THE APPROACH

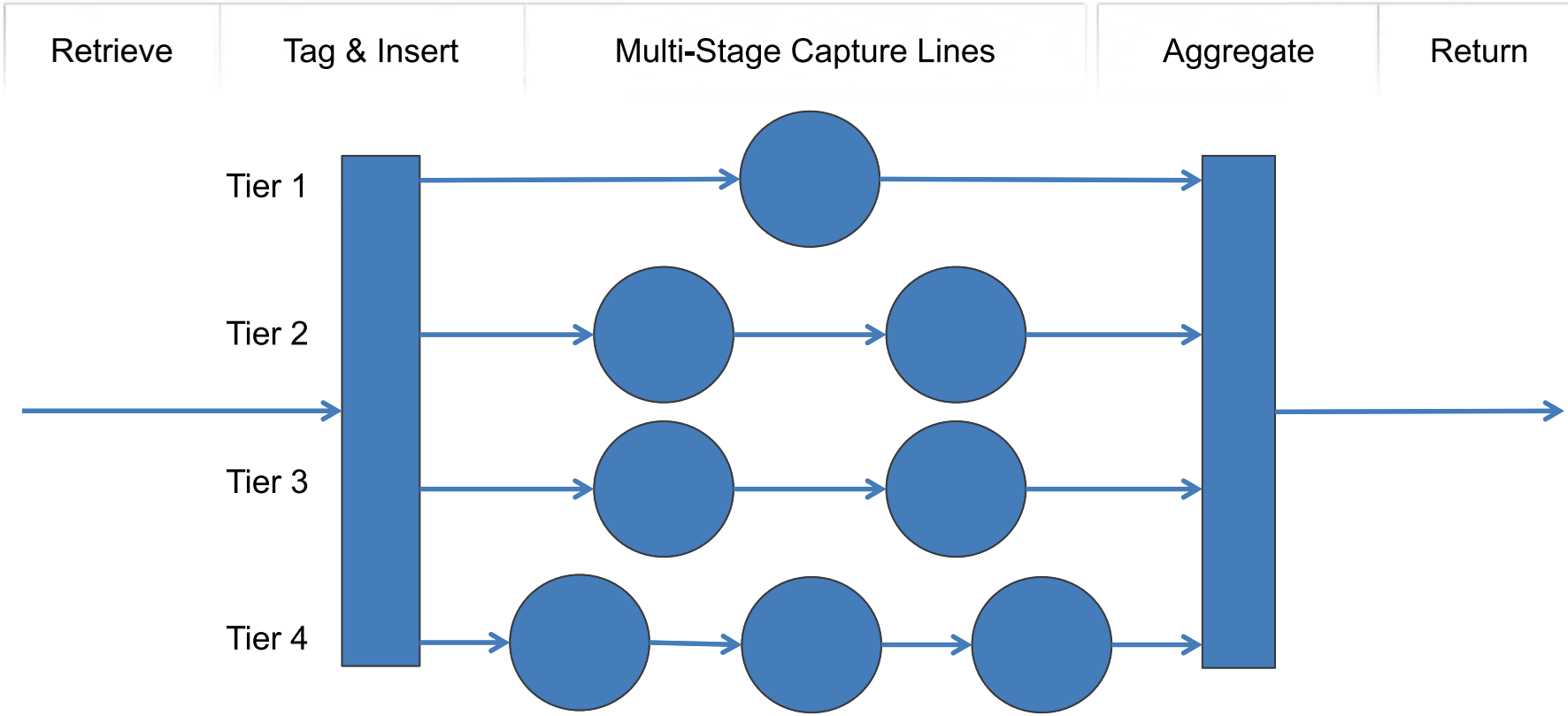
- Curatorial priorities
  - Tag each object with unique identifying code
  - Capture imagery for label data: full coverage
  - Capture reference imagery for object
- Practical priorities
  - Speed
  - One pass
    - 3D model data
- Assumptions
  - Computing and data space essentially free
  - Data can be reanalyzed at will
  - Few critical computing functions are required during collection, among them are QA, tracking, support for loading and unloading samples, error handling, data capture

ITERATIVE REFINEMENT

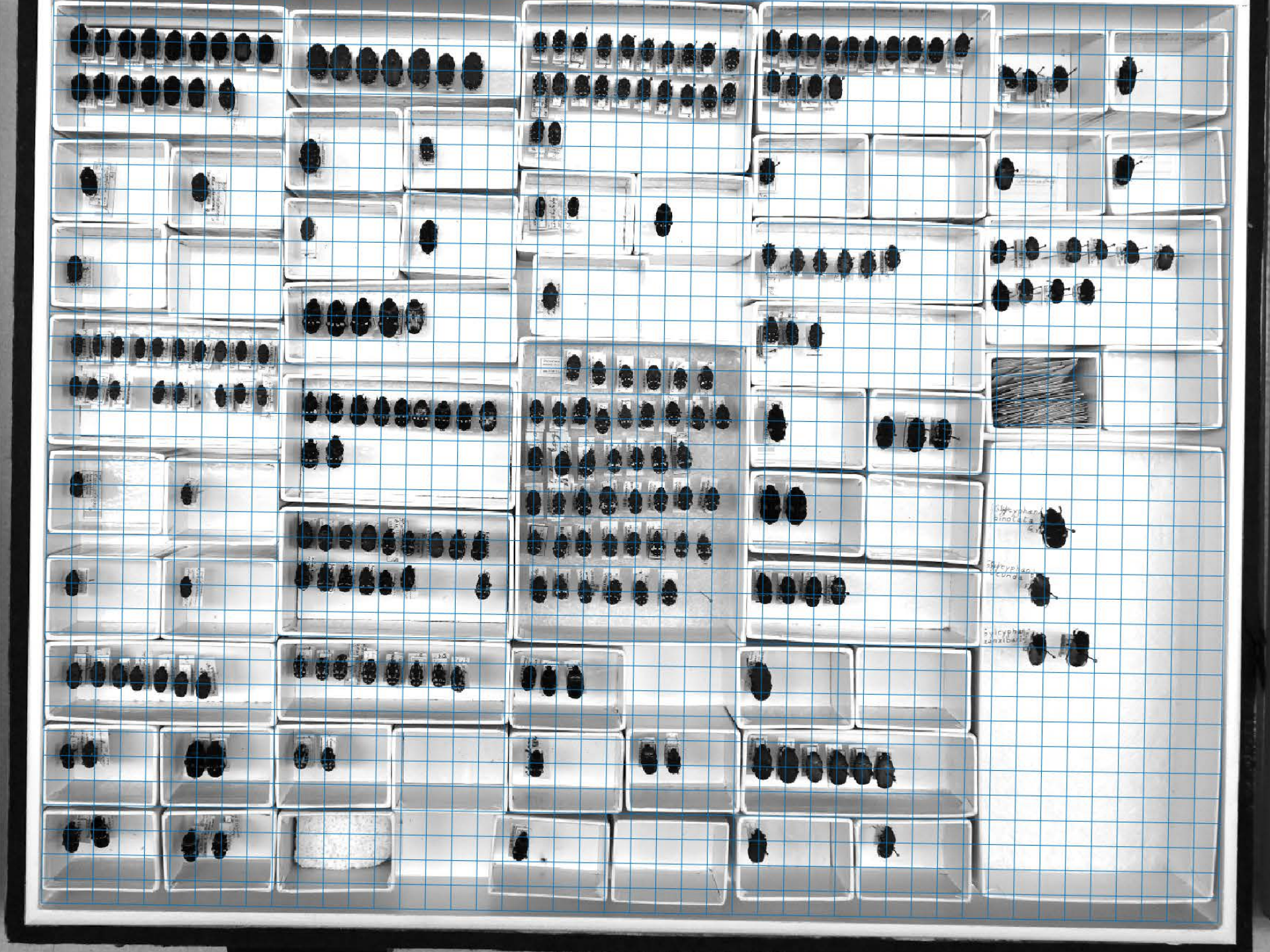
more data, better dictionaries, more time, better algorithms



PARALLEL PIPELINE DESIGN



- Time requirements
- Parallelism
- Interface between Human and Computer activities



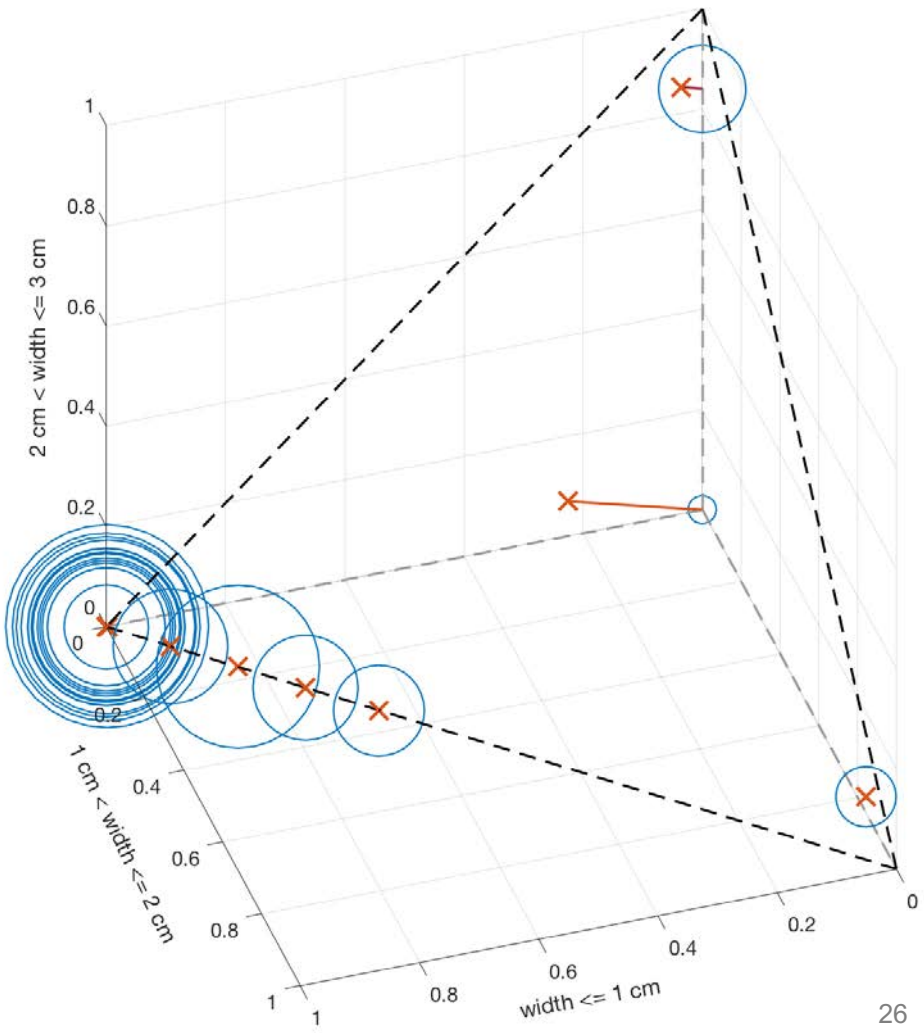
MANUAL TABULATION OF BUG SIZES IN A RANDOM SAMPLING OF THE DRAWERS

21 drawers out of the 15 thousand in the collection

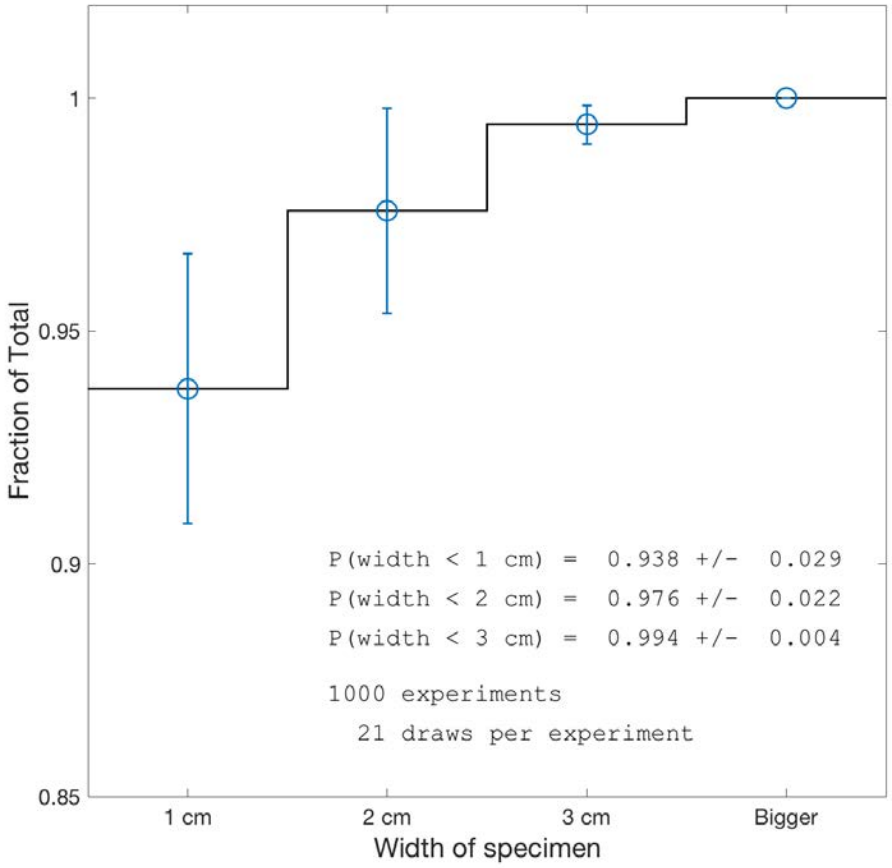
		WIDTH				LENGTH				
Drawer Index	Image Number	(0,1] cm	(1,2] cm	(2,3] cm	(3,inf) cm	(0,1] cm	(1,2] cm	(2,3] cm	(3,inf) cm	N_BUGS
1	00952	268	0	0	0	268	0	0	0	268
2	00958	0	0	0	11	0	0		11	11
3	00968	100	0	0	0	0	67	33		100
4	00990	303	0	0	0	303	0	0	0	303
5	01033	170	15	0	0	25	140	6	14	185
6	01051	1	46	4	0	0	0	1	50	51
7	01070	424	1	0	0	378	42	5	0	425
8	01086	311	62	0	0	158	192	23	0	373
9	01127	195	0	0	0	173	22	0	0	195
10	01149	577	0	0	1	532	45	0	1	578
11	01188	317	0	0	0	310	7	0	0	317
12	01221	231	0	0	0	53	178	0	0	231
13	01264	495	0	0	0	412	83	0	0	495
14	01290	249	0	0	0	249	0	0	0	249
15	01312	116	39	0	0	60	35	21	39	155
16	01333	76	40	0	0	0	50	56	10	116
17	01353	196	0	0	0	169	27	0	0	196
18	01363	345	0	0	0	345	0	0	0	345
19	01375	349	0	0	0	349	0	0	0	349
20	01395	455	0	0	0	455	0	0	0	455
21	01436	0	0	90	17	0	74	32	1	107
		5178	203	94	29	4239	962	177	126	5504
		0.941	**0.037**	**0.017**	**0.005**	0.770	0.175	0.032	0.023	

PROBABILITY DISTRIBUTION FUNCTION

From drawers to bugs



Tier: 1 2 3 4



Argonne NATIONAL LABORATORY

ESTIMATING THE SIZE DISTRIBUTION

Why does it matter?

- "Size" as a stand-in for "difficulty" to enable clustering objects into tiers
- Feeds pipeline design
- Informs R&D development priorities

- Some implications provided by the analysis:

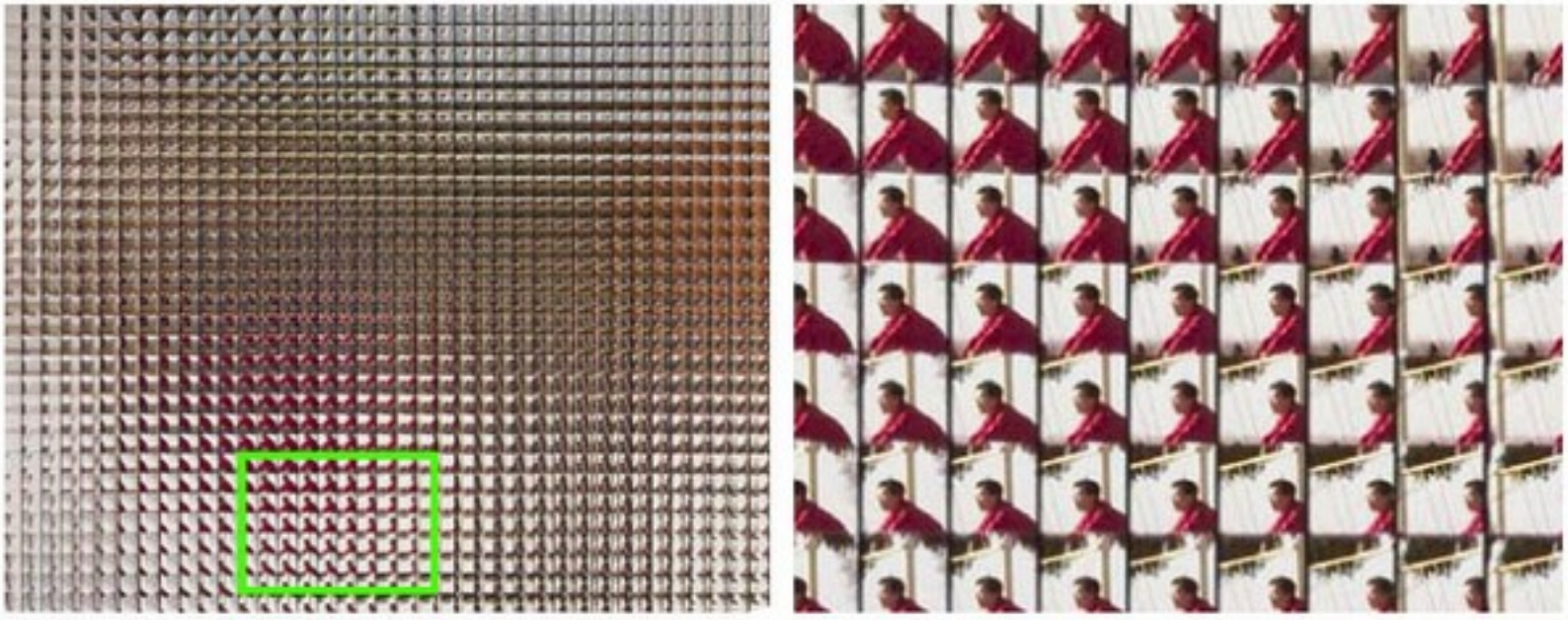
			Number of specimens in category – out of 4.5 million		Time budget per specimen (seconds) – assume 1 year		
Category	Fraction	Sigma	Smallest	Largest	Most lenient	Tightest	Mean
Tier 1 (< 1 cm)	0.938	0.028	4.1E+6	4.3E+6	1.76	1.66	1.71
Tier 2 (< 2 cm)	0.038	0.018	90.0E+3	252.0E+3	80.00	28.57	42.11
Tier 3 (< 3 cm)	0.018	0.017	4.5E+3	157.5E+3	1,600.00	45.71	88.89
Tier 4 (bigger)	0.006	0.004	9.0E+3	45.0E+3	800.00	160.00	266.67

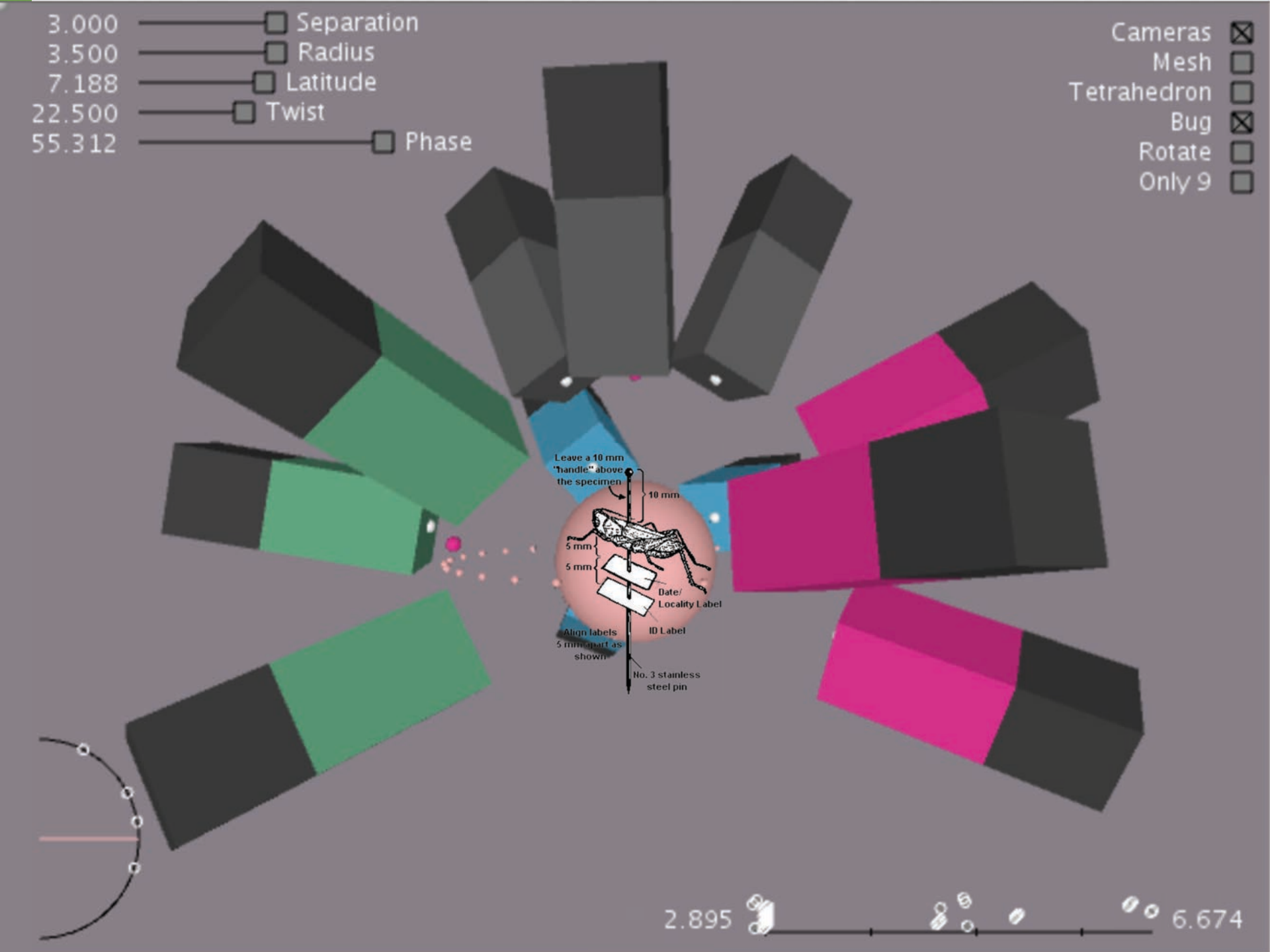
MODULAR CAMERA HEAD DESIGN

PLENOPTIC CAMERAS

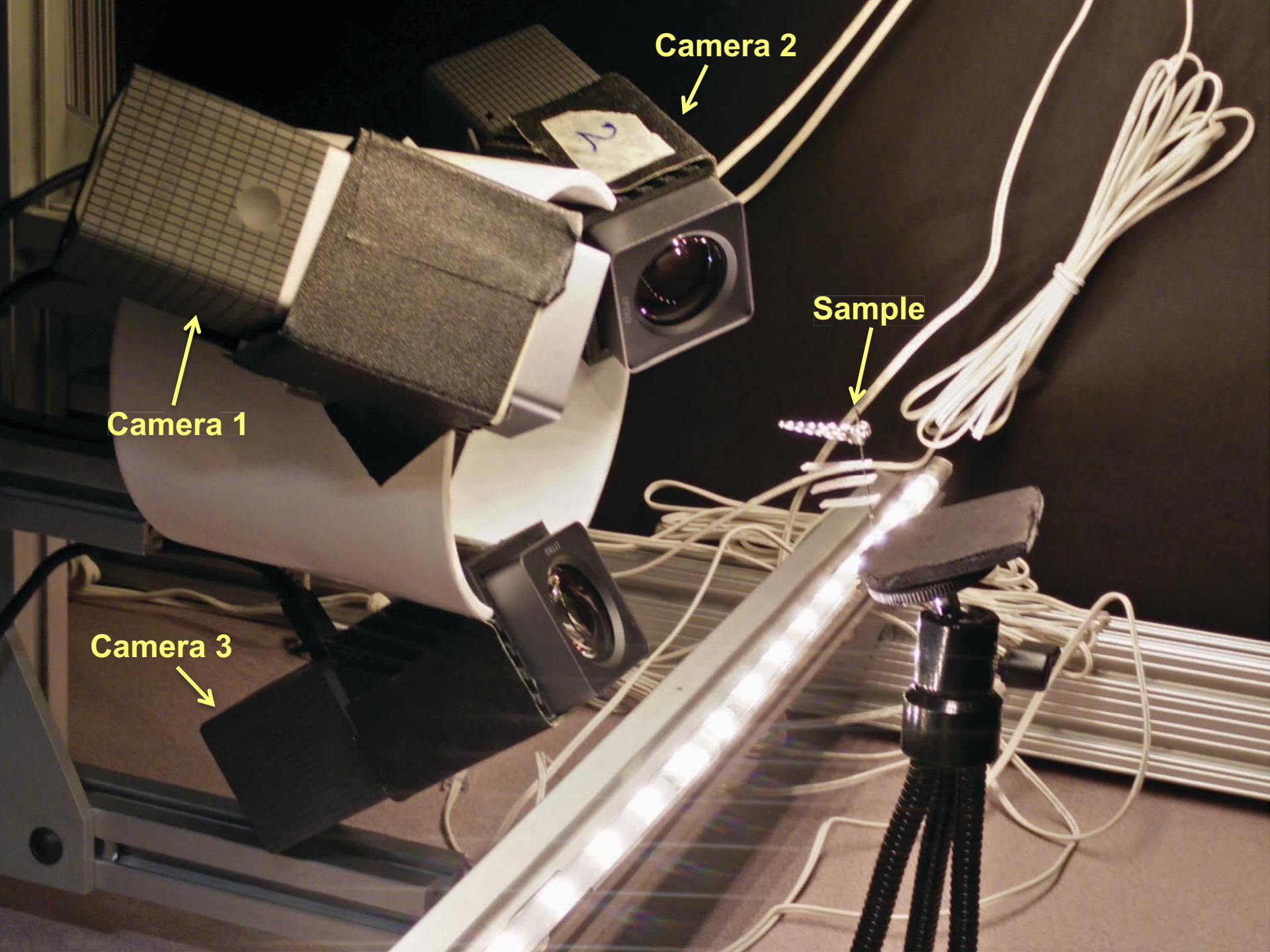
Light field camera = fine grain multi-view camera

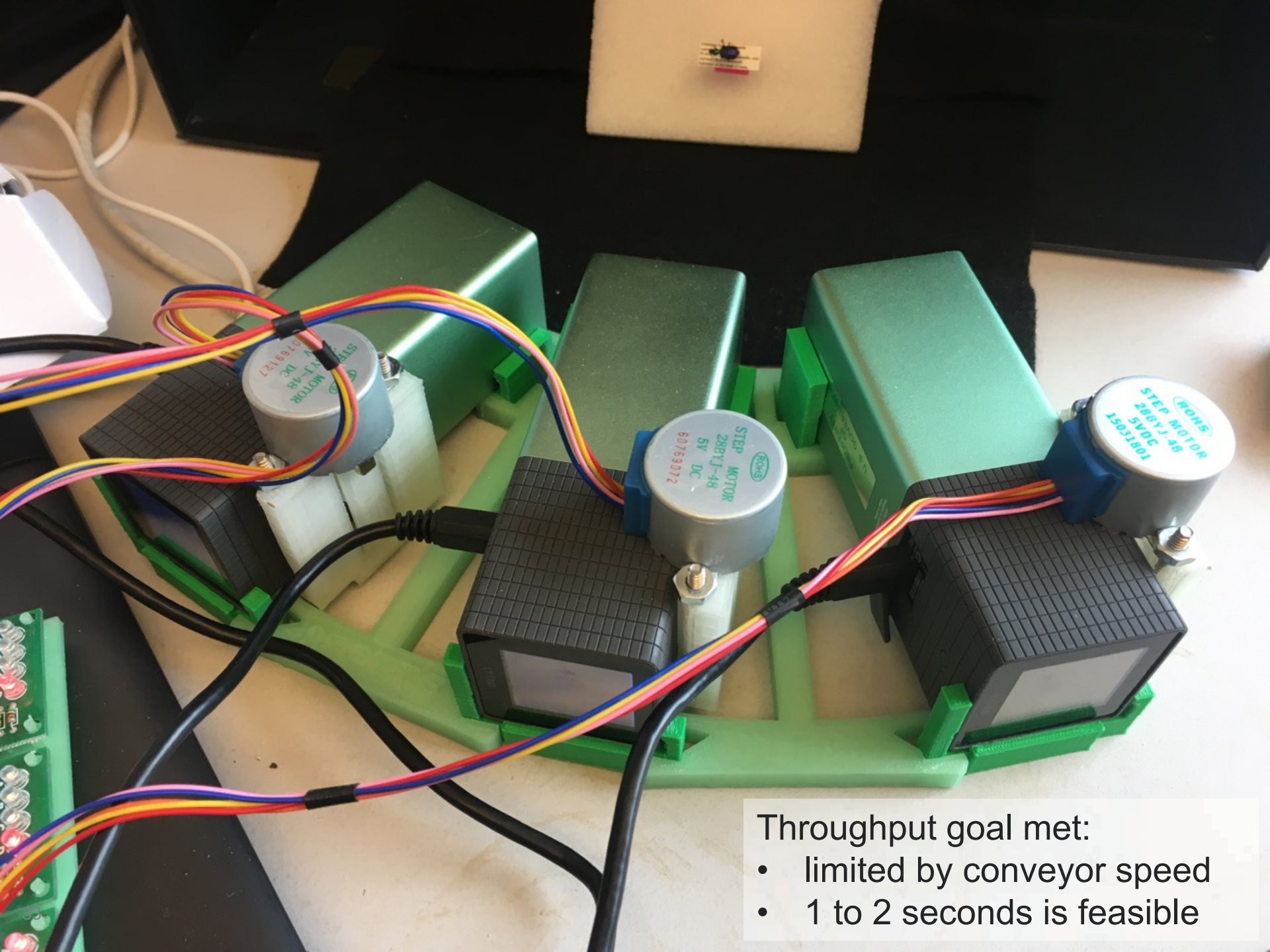
- 2 apertures enables stereoscopic 3D
- Light field camera provides large array of fine grain apertures
- Array of light field cameras provides 4π angular coverage of sample volume




3.000 Separation
3.500 Radius
7.188 Latitude
22.500 Twist
55.312 Phase
Cameras
Mesh
Tetrahedron
Bug
Rotate
Only 9
Leave a 10 mm "handle" above the specimen
10 mm
5 mm
5 mm
Date/ Locality Label
ID Label
Align labels 5 mm apart as shown
No. 3 stainless steel pin
2.895
6.674


The Benson


Camera 2
Sample
Camera 1
Camera 3



Throughput goal met:

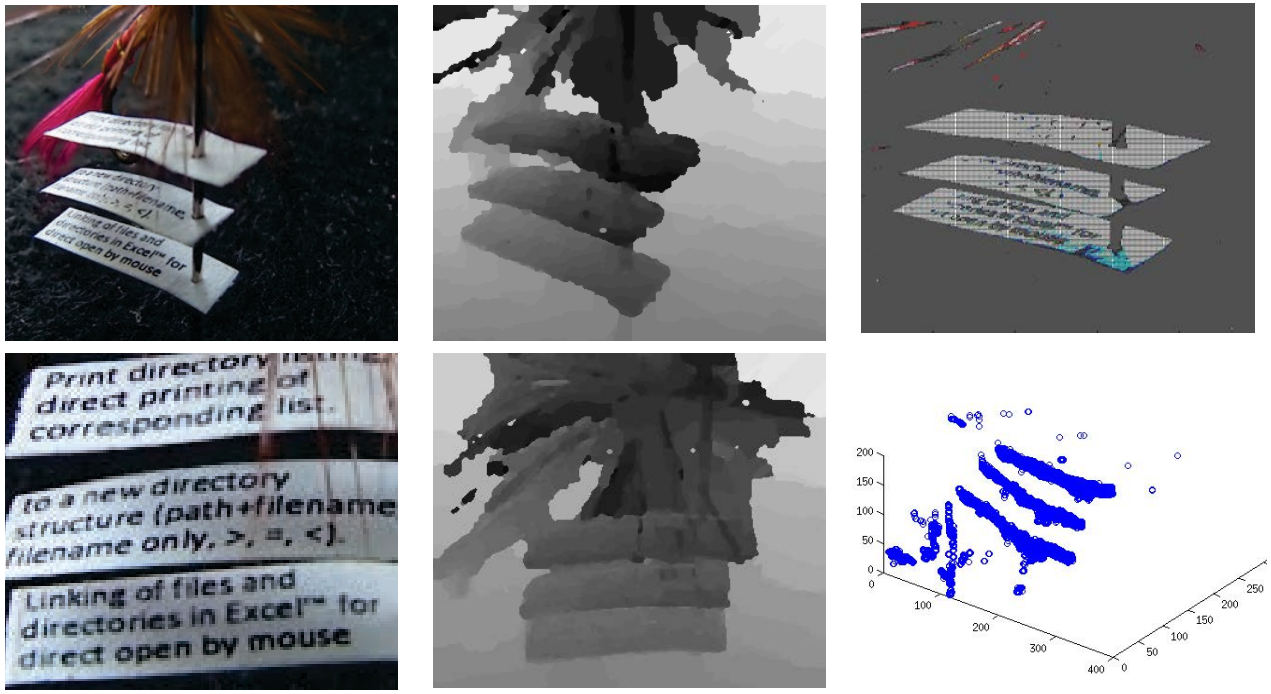
- limited by conveyor speed
- 1 to 2 seconds is feasible

ANALYZING THE IMAGES

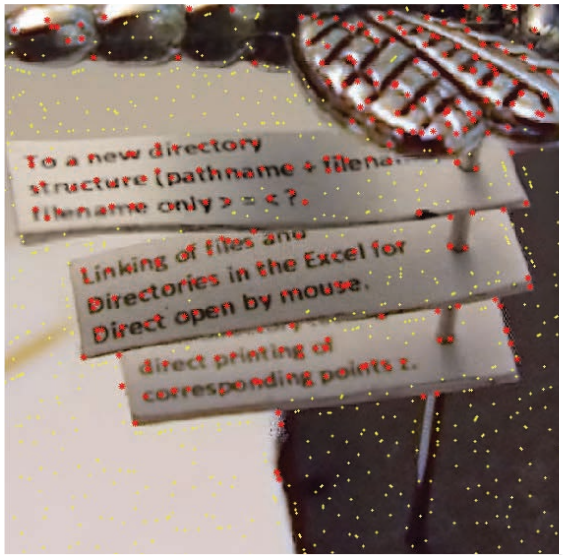
DEPTH FROM LIGHT FIELD

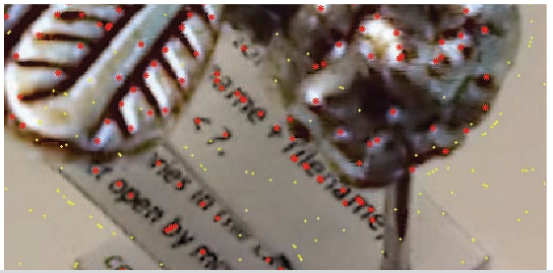
- RGB-D obtained from light field data
  - Identify labels (3D segmentation)
  - Infinite focus reconstruction

Image **Depth** **Point cloud**



DEPTH FROM MULTI-POSE RIG

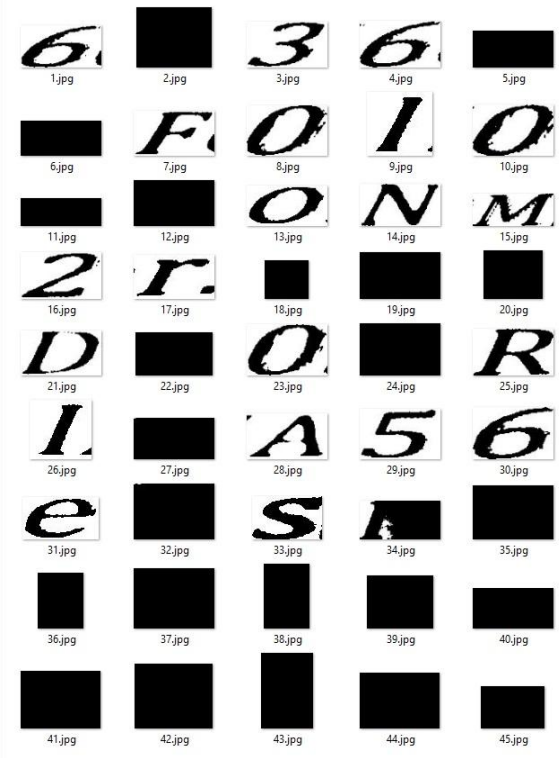




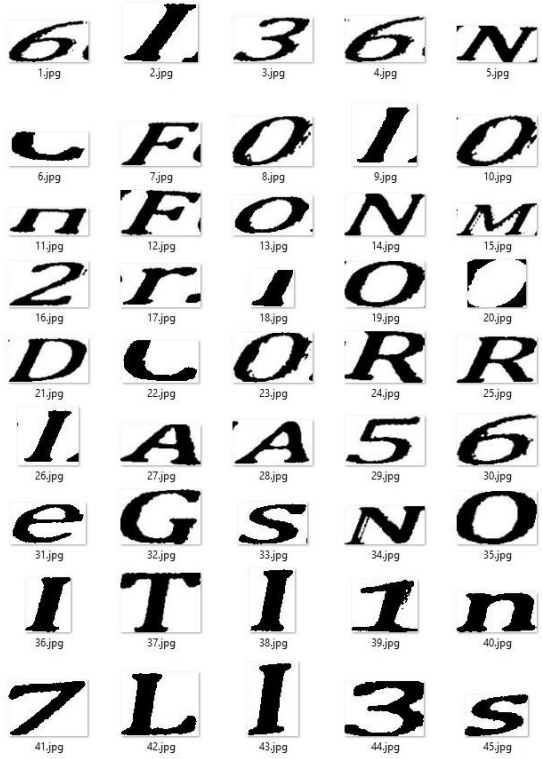
- Feature detection
- Matching
- Bundle adjustment
- Sparse reconstruction
- Dense reconstruction

CHARACTER CLEANING

ORIGINAL TILE

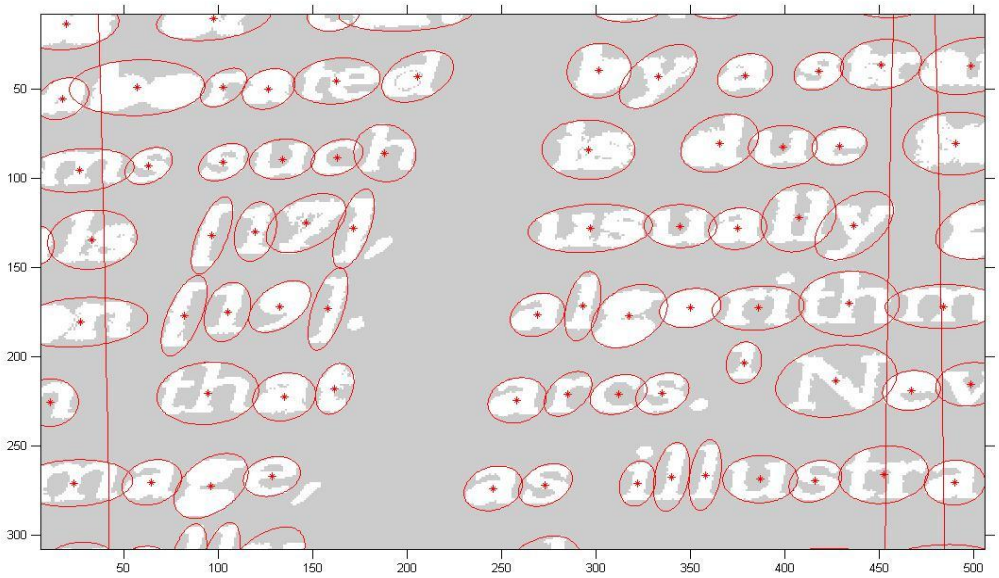


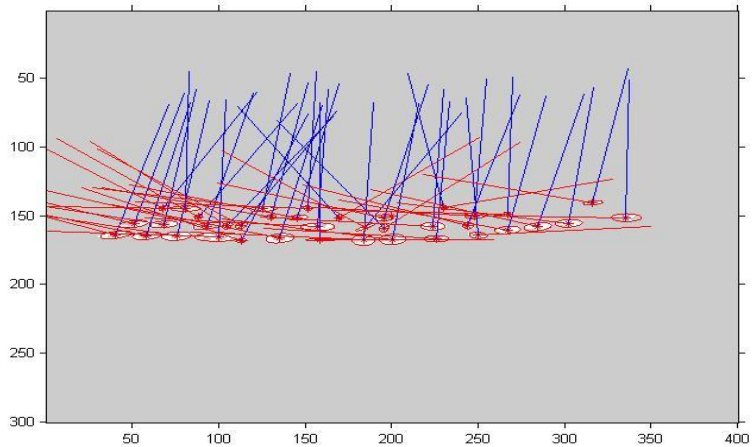
NOMINAL BINARIZATION



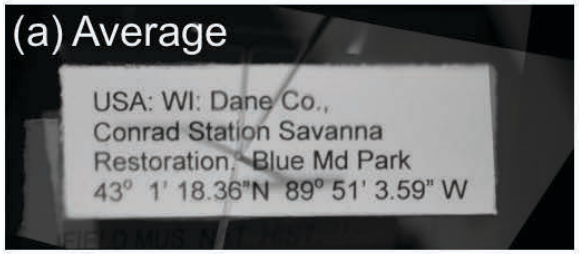
ADAPTIVE BINARIZATION

CHARACTER (BLOB) METHODS FOR ALIGNMENT, REGISTRATION, COMPOSITING



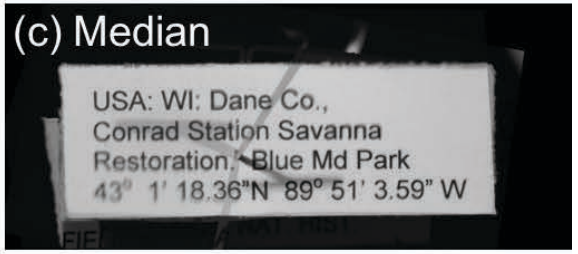


VIEW COMBINING



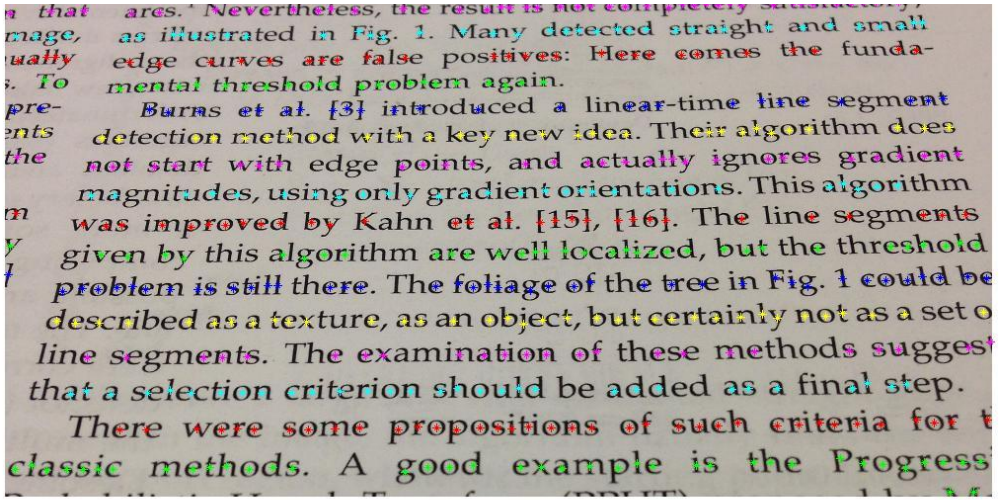
USA: WI: Dane Co.,
Conrad Station Savanna
Restoratior[ib]N3lue Md Park
43⁰ 1' 18.36"N 89° 51' 3.59" W

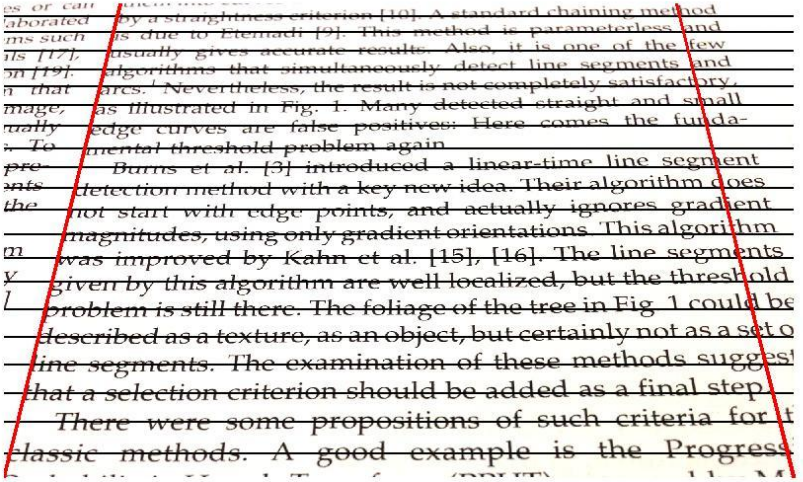
USA: WI: Dane Co●
Conrad Station Savanna
Restoration! Blue Md Park
43⁰ **1' 18_36"N 89" 51' 3.59" W**

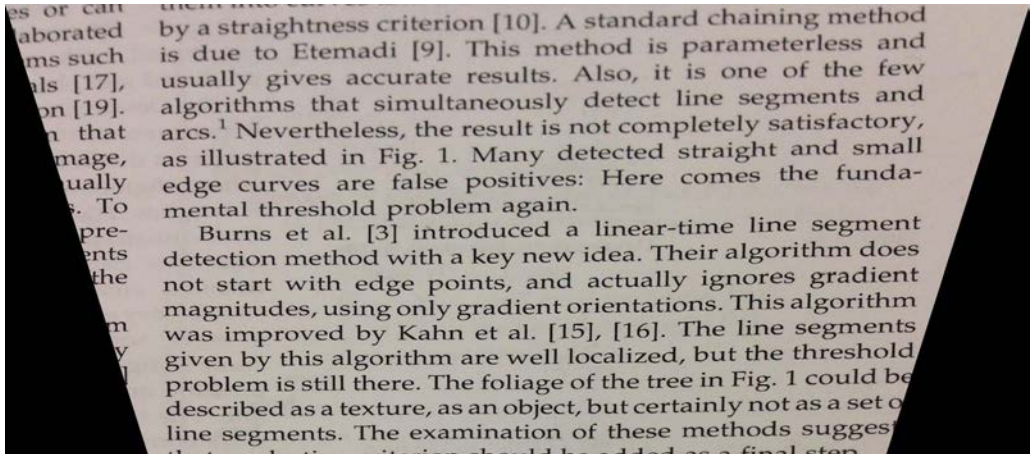


USA: WI: Dane Ca,
Conrad Station Savanna
Restoration.431ue Md Park
43⁰ 1' 18.36"N 89° 51' 3.59“W

RECTIFICATION



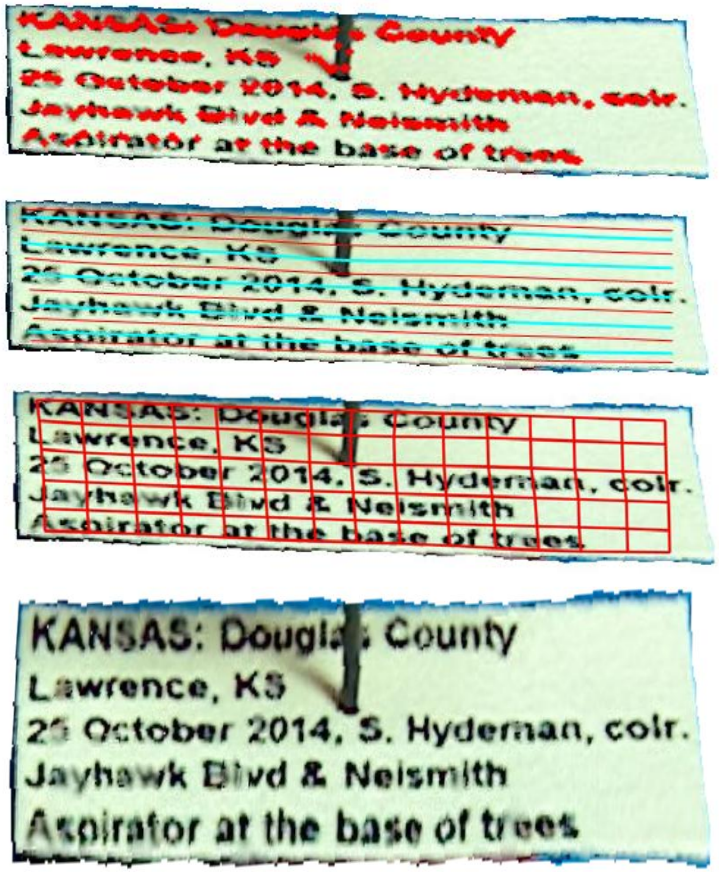




algorithms that simultaneously detect line segments and arcs.[1] Nevertheless, the result is not completely satisfactory, as illustrated in Fig. 1. Many detected straight and small edge curves are false positives: Here comes the fundamental threshold problem again.

Burns et al. [3] introduced a linear-time line segment detection method with a key new idea. Their algorithm does not start with edge points, and actually ignores gradient magnitudes, using only gradient orientations. This algorithm was improved by Kahn et al. [15], [16]. The line segments

RECTIFICATION



Points

Lines

Warped Grid

Unwarp

IN CLOSING

SUMMARY

High throughput pipeline for digitizing large collections of objects

- Immediate goals: label images for transcription, 3D model of each specimen
  - Target time frame: 1 year of data collection
- Computing is deferred unless it is critical to reliable operation of the pipeline
  - High performance as necessary
  - Iterative analysis for dictionary building, algorithm devel. & optimization
- Developed new hardware platforms for rapid “snapshot” 3D capture
  - Light field plenoptic camera configuration allows single shot, no focusing
  - Multi-camera rig provides robust snapshot coverage of entire object
  - Able to meet Tier 1 capture time budget
- Parameters for parallel pipelined digitization workflow
  - Multi-tier process to handle range of contingencies
- Built up a suite of methods for offline analysis of imagery to enable automated label transcription
- Connections with the larger collection digitization community

THE TEAM

- Mark Hereld, MCS
- Nicola J. Ferrier, MCS

STUDENTS

- Nitin Agarwal, Research Aide 2016
- Juliana Kim, SULI 2015
- Joshua Koblich, SULI 2015
- Andi Zang, Research Aide 2014
- Bryan Dalle-Molle, SULI 2014
- Daniel Shiff, UC MS Practicum 2014

COLLABORATORS

- Petra Sierwald, FMNH
- Rudiger Bieler, FMNH
- Crystal Maeir, FMNH

THANK YOU !

This material is based upon work supported by Laboratory Directed Research and Development (LDRD) funding from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under contract DE-AC02-06CH11357.