CS 5604: Information Storage And Retrieval

Final Presentation

ETD Team

Instructor: Prof. Edward A. Fox
Department of Computer Science, Virginia Tech
Blacksburg, Virginia - 24061

December 9, 2020

Aarathi Raghuraman
Alex Nguyen
Javaid Manzoor
Jiahui Fan
Nicolas Hardy
Sam Furman

Grant: IMLS LG-37-19-0078-19

AGENDA

1. Summary Statistics
2. ETD Architecture
3. ETD Services
    a. Validate Input
    b. Figure/Table Extraction
    c. Chapter Segmentation
    d. Text Extraction
    e. Classification
    f. Metadata Ingestion (ELS)

Summary Statistics

- Collection contains 19,779 theses and 14,691 dissertations.
  - Subsets contain 410 theses and 282 dissertations
- 19% and 32% of the 31 attributes are empty for theses & dissertations respectively
- We add attributes to ETD Metadata and create metadata for chapters

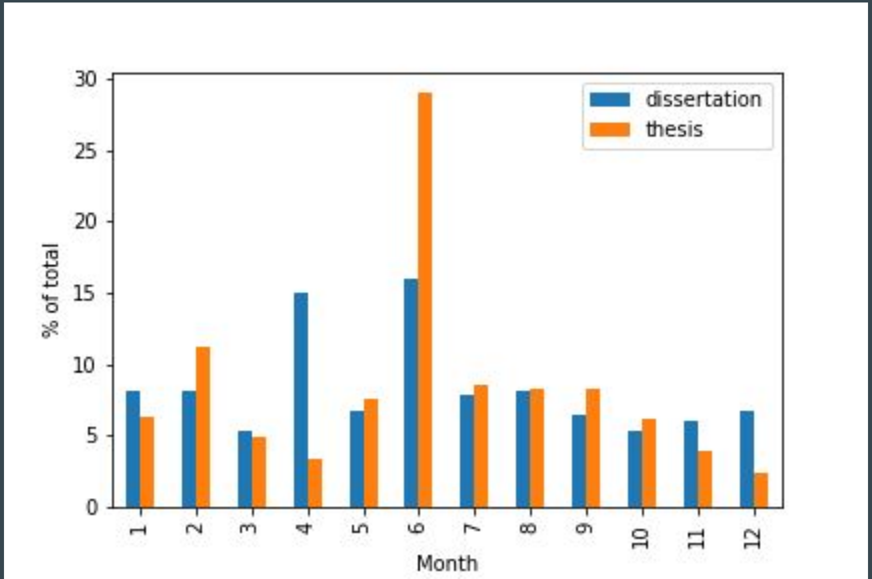
Field	Description
contributor_author	main author of the document
contributor_committeechair	chair(s) of the author's advisory committee
contributor_committeecochair	co-chair(s) of the author's advisory committee
contributor_comm	committee
contributor_d	or belongs to
date_acces	e collection
data_ava	for viewing
date_iss	
date_ac	
date_sc	
date_r	
degree_dis	
degree_gr	ree
degree_l	r "masters"
degree_n	or "MS"
description_	
description_	"PHD"
description_p	at are noteworthy
description_sp	ed the research
format_m	TD"
handl	ents
identifier_	re not the handle
identifier	ent
publish	rce available
relatio	t(s)
relation_h	ocument
rights	nformation
subjec	the document
subject_lcc	Library of Congress categories
subject_lcsh	Library of Congress subject headers
title	the title of the document

- ★ etd_filename
- ★ etd_uri
- ★ figures_folder_uri
- ★ figures_total_count
- ★ tables_folder_uri
- ★ tables_total_count
- ★ subject [append unique]

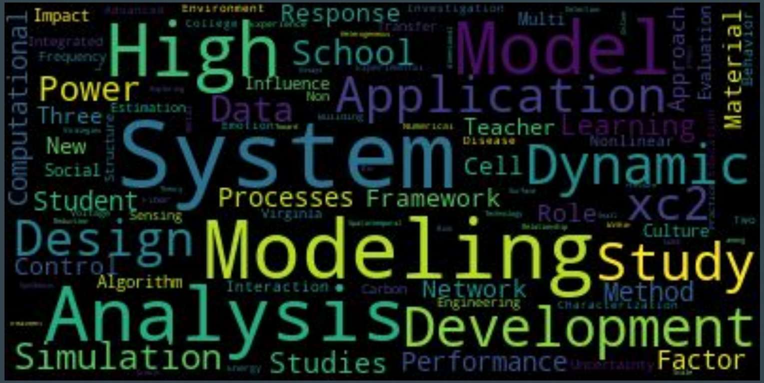
- ★ chapter_number
- ★ chapter_uri
- ★ subject
- ★ chapter_fulltext_uri
- ★ pages
- ★ page_number

Summary Statistics

- ❖ Collection contains 19,779 theses and 14,691 dissertations.
  - ➢ Subsets contain 410 theses and 282 dissertations
- ❖ 19% and 32% of the 31 attributes are empty for thesis & dissertations respectively
- ❖ We add attributes to ETD Metadata and create metadata for chapters
- ❖ Fact Check: Issue Date highest in June



Summary Statistics



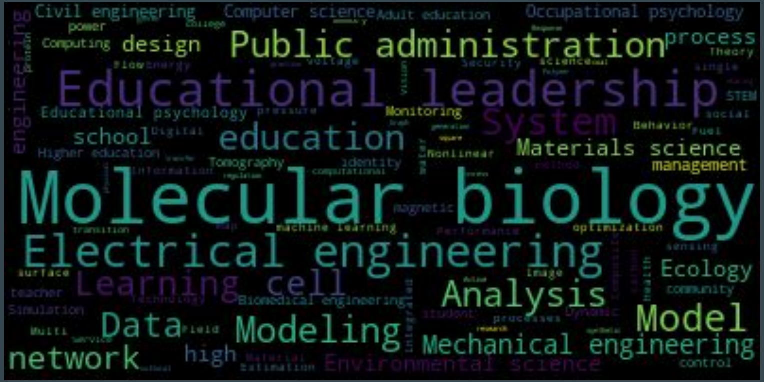
Dissertations

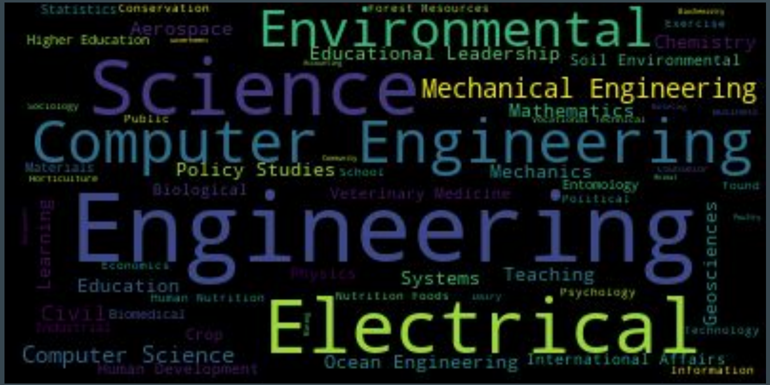
Theses

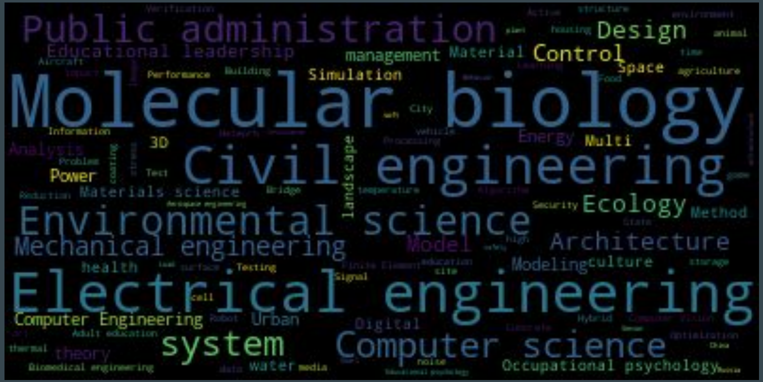
ABSTRACTS

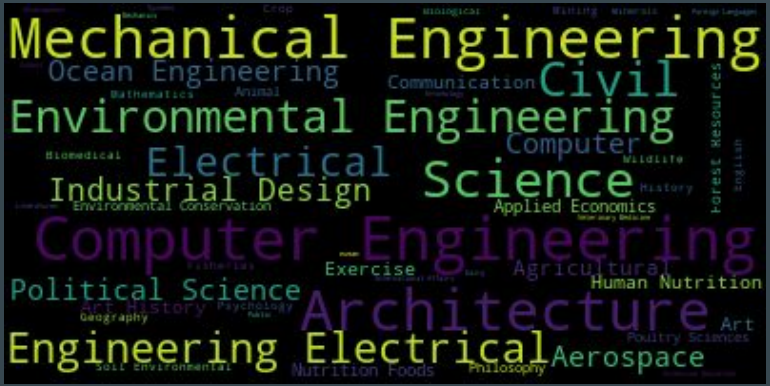
TITLES

Summary Statistics





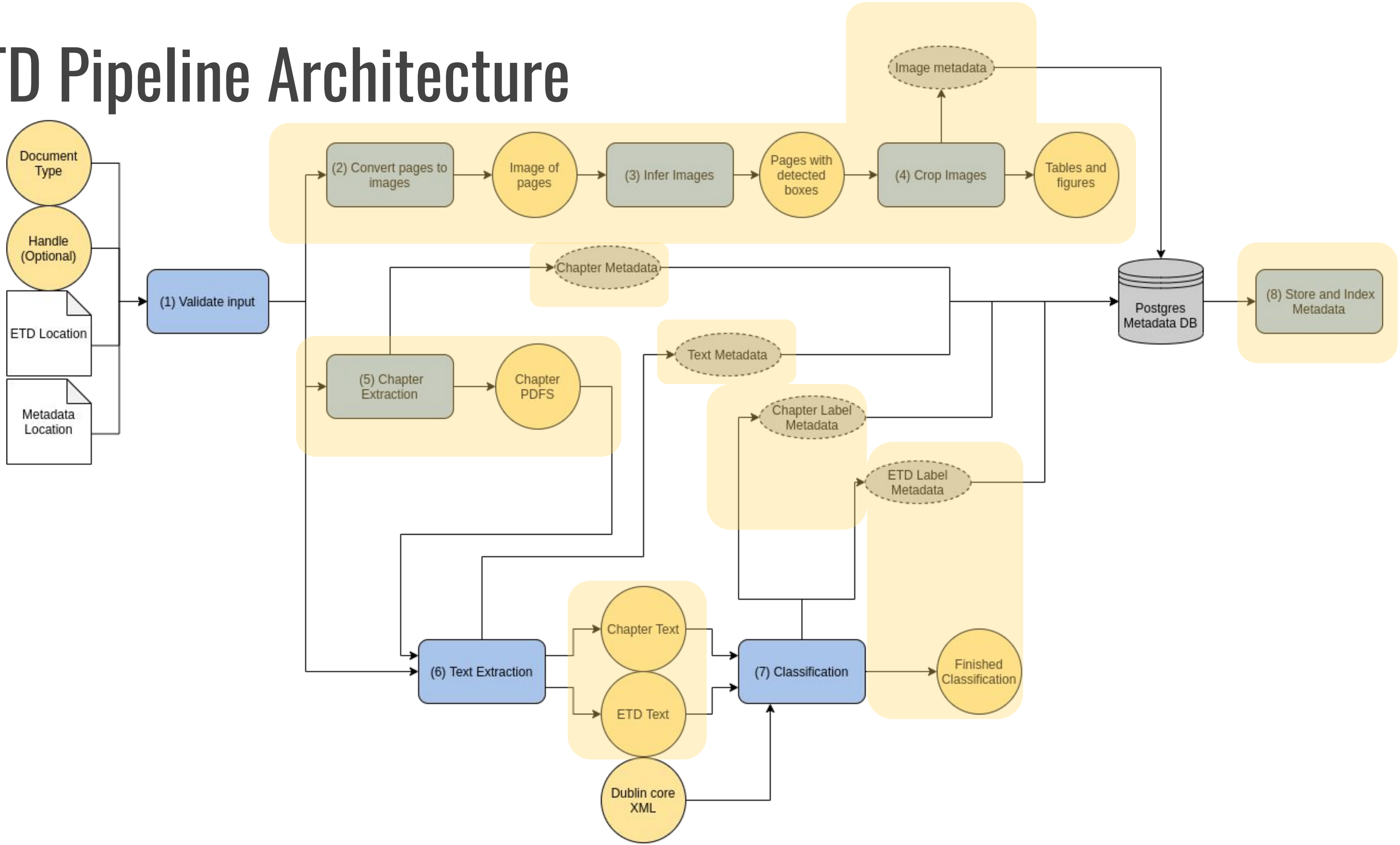




SUBJECTS

AUTHOR DEPARTMENT

ETD Pipeline Architecture



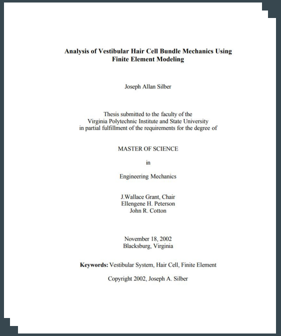
Validate Input

```
{handle: 9704}
```

handle

```
{type: Thesis}
```

doc_type


Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

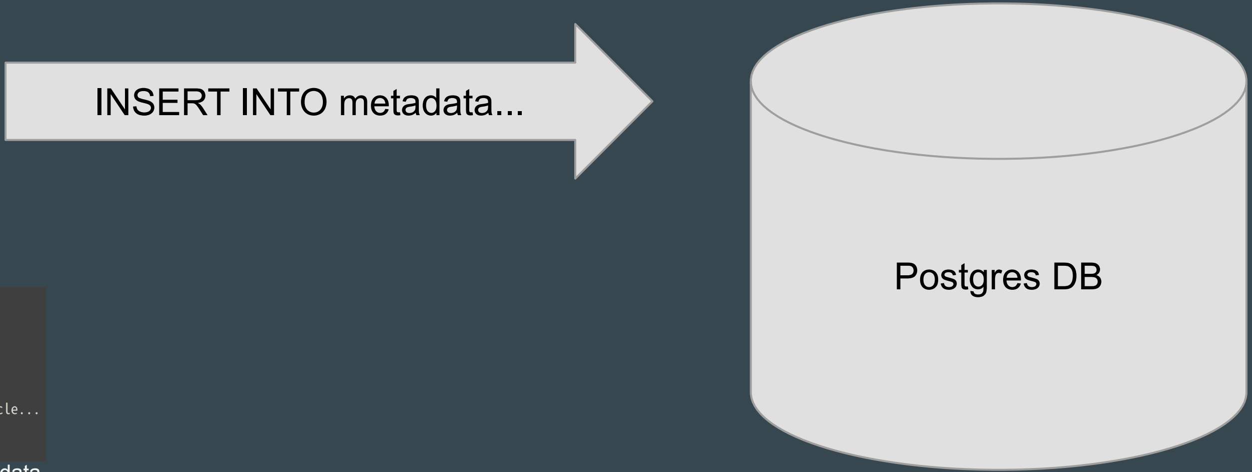
Copyright 2002, Joseph A. Silber

ETD

```
contributor-author: Silber, Joseph Allan
date-accessioned: 2011-08-06T14:45:39Z
date-available: 2011-08-06T14:45:39Z
date-issued: 2002-11-18
identifier-other: etd-12012002-165307
identifier-uri: http://hdl.handle.net/10919/9704
description-abstract: The vestibular system of vertebrates consists of the utricle...
...
...
```

metadata

Creates a new row in the database for the incoming document.



Validate Input

Tested & Containerized

{handle: 9704}

handle

{type: Thesis}

doc_type


Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

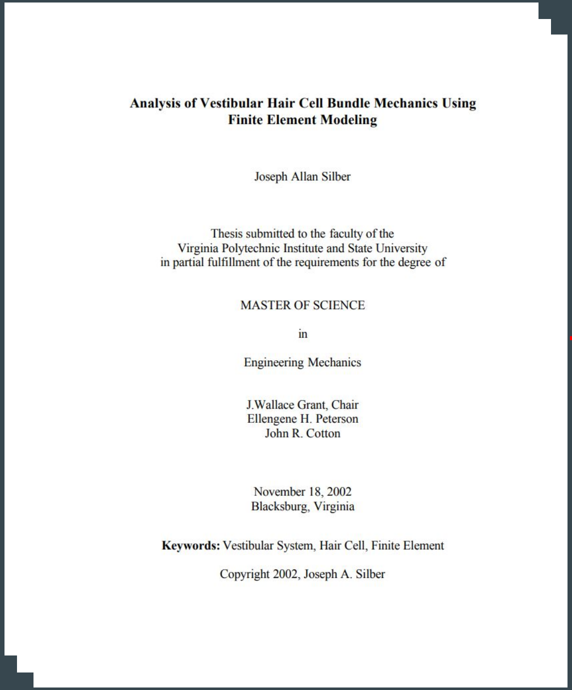
Creates a new row in the database for the incoming document.

Future works

★ Add support for more forms of ETD metadata (e.g. JSON, XML)

Table/Figure Extraction: Image to PDF

- First we convert pdfs to an image format that is amenable to inference


Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber

ETD

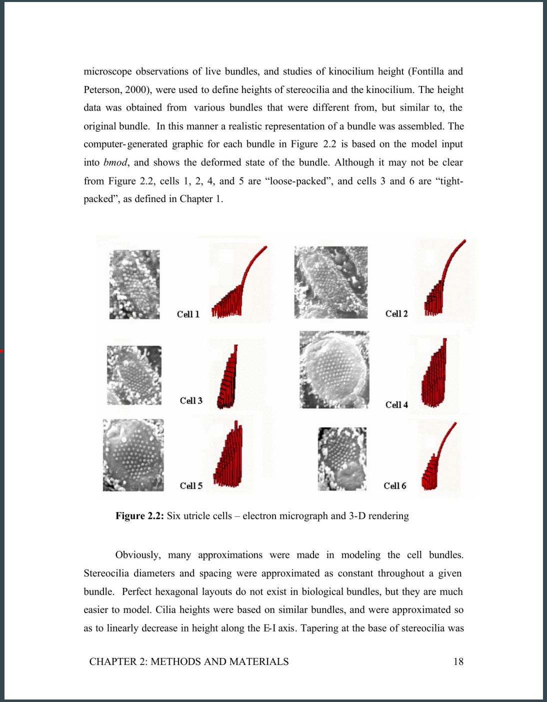

microscope observations of live bundles, and studies of kinocilium height (Fontilla and Peterson, 2000), were used to define heights of stereocilia and the kinocilium. The height data was obtained from various bundles that were different from, but similar to, the original bundle. In this manner a realistic representation of a bundle was assembled. The computer-generated graphic for each bundle in Figure 2.2 is based on the model input into *bmod*, and shows the deformed state of the bundle. Although it may not be clear from Figure 2.2, cells 1, 2, 4, and 5 are "loose-packed", and cells 3 and 6 are "tight-packed", as defined in Chapter 1.

Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the E-I axis. Tapering at the base of stereocilia was

CHAPTER 2: METHODS AND MATERIALS 18

Page image(s)

Table/Figure Extraction: Image Inference

- Inference is accomplished via the best performing model trained by Sampanna and others

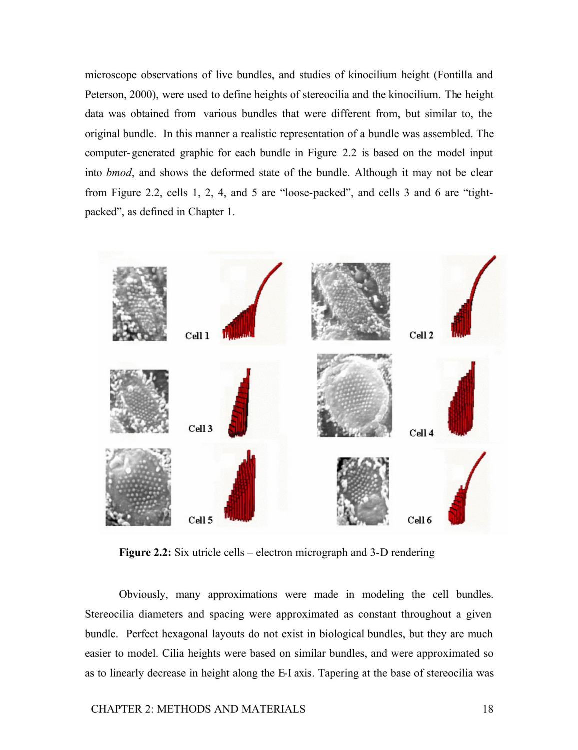
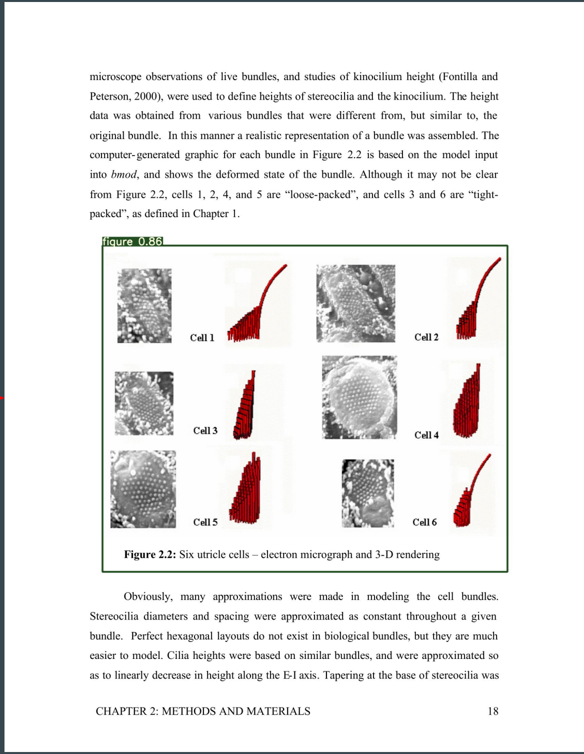

microscope observations of live bundles, and studies of kinocilium height (Fontilla and Peterson, 2000), were used to define heights of stereocilia and the kinocilium. The height data was obtained from various bundles that were different from, but similar to, the original bundle. In this manner a realistic representation of a bundle was assembled. The computer-generated graphic for each bundle in Figure 2.2 is based on the model input into *bmod*, and shows the deformed state of the bundle. Although it may not be clear from Figure 2.2, cells 1, 2, 4, and 5 are "loose-packed", and cells 3 and 6 are "tight-packed", as defined in Chapter 1.

Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the E-I axis. Tapering at the base of stereocilia was

CHAPTER 2: METHODS AND MATERIALS 18

Page image(s)



microscope observations of live bundles, and studies of kinocilium height (Fontilla and Peterson, 2000), were used to define heights of stereocilia and the kinocilium. The height data was obtained from various bundles that were different from, but similar to, the original bundle. In this manner a realistic representation of a bundle was assembled. The computer-generated graphic for each bundle in Figure 2.2 is based on the model input into *bmod*, and shows the deformed state of the bundle. Although it may not be clear from Figure 2.2, cells 1, 2, 4, and 5 are "loose-packed", and cells 3 and 6 are "tight-packed", as defined in Chapter 1.

Cell 1 Cell 2 Cell 3 Cell 4 Cell 5 Cell 6

Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the E-I axis. Tapering at the base of stereocilia was

CHAPTER 2: METHODS AND MATERIALS 18

Page image(s) with image bound information

Table/Figure Extraction: Image Cropping

- The model outputs bounding boxes that can be used to crop images

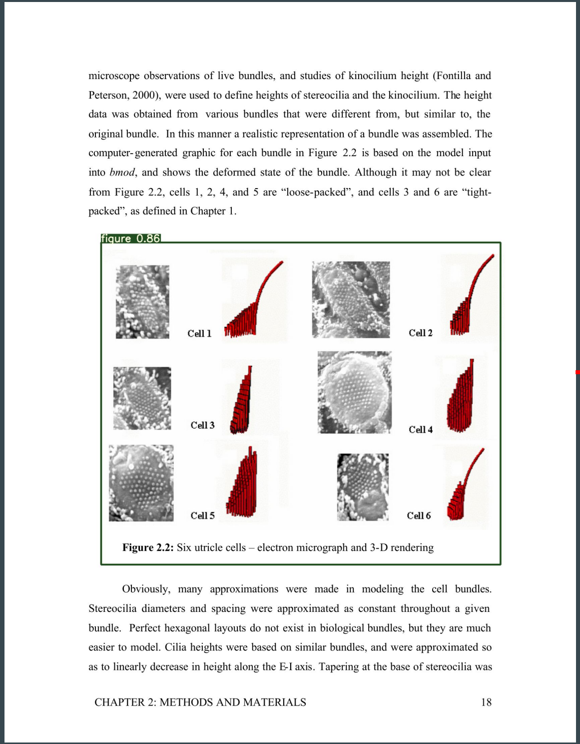

microscope observations of live bundles, and studies of kinocilium height (Fontilla and Peterson, 2000), were used to define heights of stereocilia and the kinocilium. The height data was obtained from various bundles that were different from, but similar to, the original bundle. In this manner a realistic representation of a bundle was assembled. The computer-generated graphic for each bundle in Figure 2.2 is based on the model input into *bmod*, and shows the deformed state of the bundle. Although it may not be clear from Figure 2.2, cells 1, 2, 4, and 5 are "loose-packed", and cells 3 and 6 are "tight-packed", as defined in Chapter 1.

figure 0.86

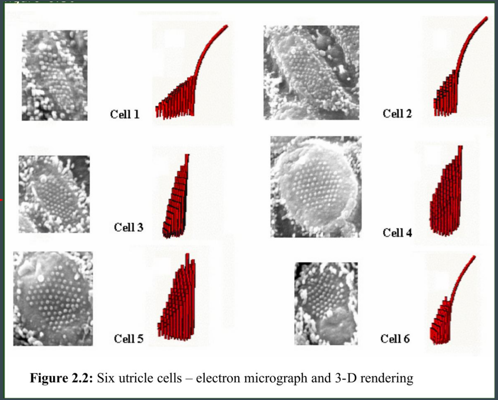
Cell 1 Cell 2 Cell 3 Cell 4 Cell 5 Cell 6

Figure 2.2: Six utricle cells – electron micrograph and 3-D rendering

Obviously, many approximations were made in modeling the cell bundles. Stereocilia diameters and spacing were approximated as constant throughout a given bundle. Perfect hexagonal layouts do not exist in biological bundles, but they are much easier to model. Cilia heights were based on similar bundles, and were approximated so as to linearly decrease in height along the E-I axis. Tapering at the base of stereocilia was

CHAPTER 2: METHODS AND MATERIALS 18

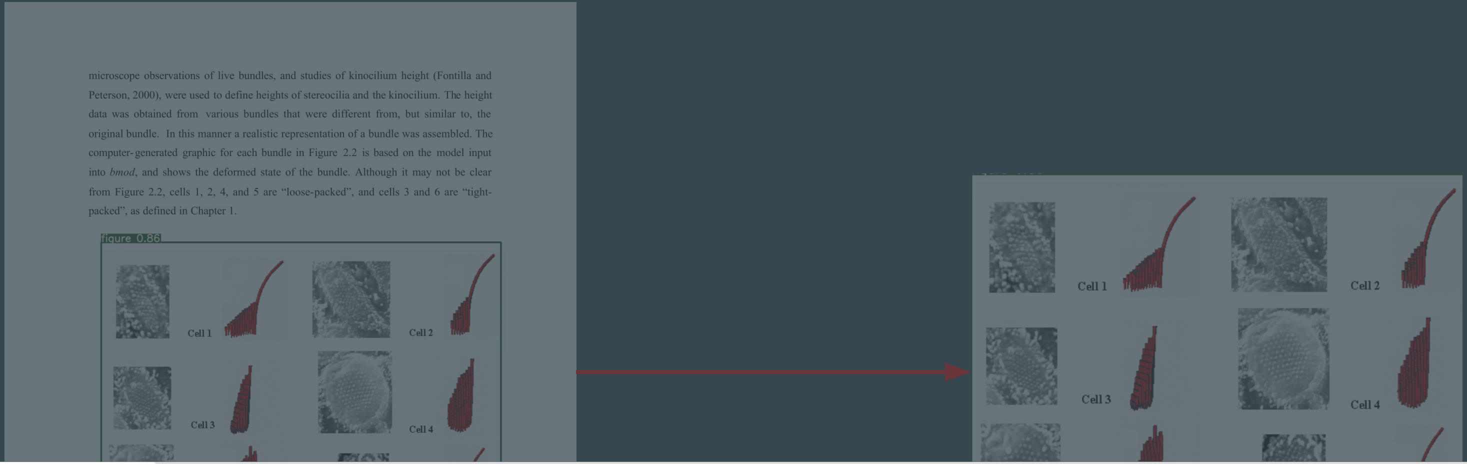
Page image(s) with image bound information



Cropped images

Table/Figure Extraction

- The model outputs bounding boxes that can be used to crop images



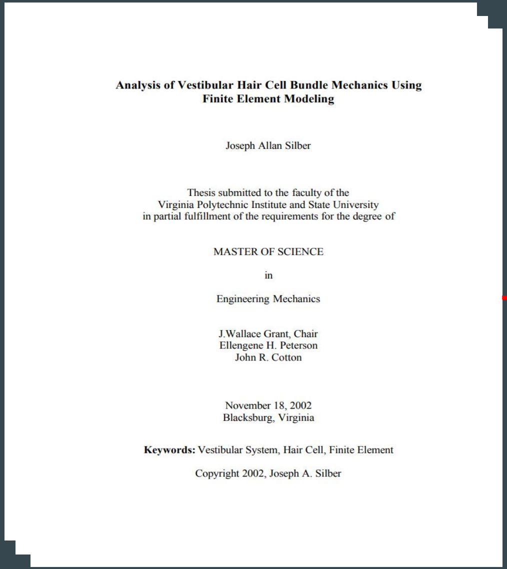
Future works

★ Adapt the full deepfigures-open pipeline to Airflow

★ Expose services to train and evaluate new figure extraction models

Chapter Segmentation

Median 7 Chapters
80% under 20 chapters
60% coverage of test data



Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber

ETD

`{handle: 9704}`

handle

`{type: Thesis}`

doc_type

CHAPTER 1: INTRODUCTION AND BACKGROUND

CHAPTER 2: METHODS AND MATERIALS

CHAPTER 3: THREE-DIMENSIONAL BUNDLE MECHANICS

CHAPTER 4: ION GATES

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

If one were to try and sum up the conclusions obtained from this research into one statement, perhaps the best summary would be to say that bundles are mechanically complex, and all details are important in accurately modeling them.

Accurate knowledge of the geometry of a bundle is crucial. Cilia diameters, numbers of and locations of cilia, and cilia heights all have significant effects on bundles stiffness, as elaborated on in chapter 3. Although not discussed in detail, even factors such as stereocilia base tapering, and tip link diameters can noticeably influence stiffness. Certainly, modeling a bundle as a simple row or column neglects a significant amount of information and can give incorrect results.

Equally important in accurate modeling are the material properties, such as elastic moduli and shear moduli. Of particular importance is the tip link elastic modulus, which is important both in affecting overall bundle stiffness, as well as influencing the behavior of the theorized ion gated channels.

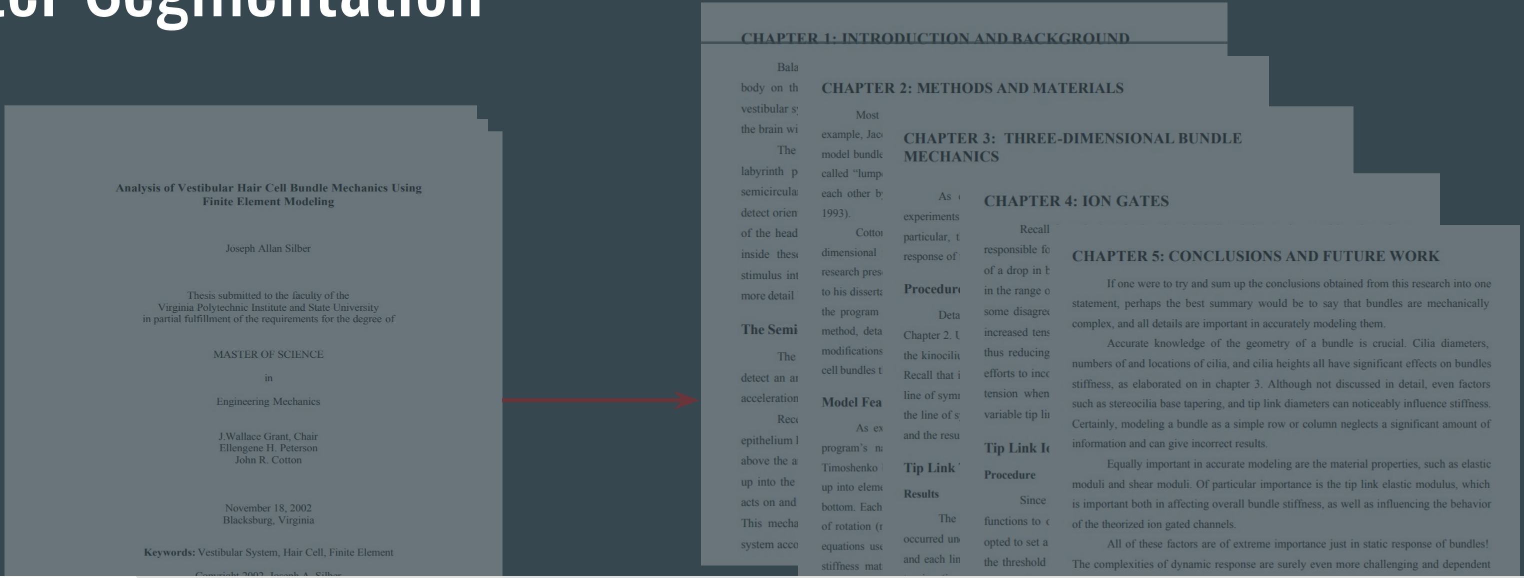
All of these factors are of extreme importance just in static response of bundles! The complexities of dynamic response are surely even more challenging and dependent on these (and other) factors.

The implications of these conclusions are three-fold. First, and unsurprisingly, better information about bundles is needed to improve modeling efforts. The material properties of tip and lateral links need to be known more precisely. Unfortunately, it is currently impossible to measure these properties directly; testing values in a model is presently the best possible way to determine these values. Geometric properties of individual bundles being modeled need to be measured more exactly. The details are important; rough estimates are insufficient. The importance of the stereocilia/kinocilium height ratio suggests that accurate height data is particularly crucial, but cilia diameters, taper ratios, and other values are also vital. Second, modeling needs to be as precise as possible. Lumped parameter models and simple 2-D row models are not sufficient. They

Chapter fulltext

Chapter Segmentation

Median 7 Chapters
80% under 20 chapters
60% coverage of test data

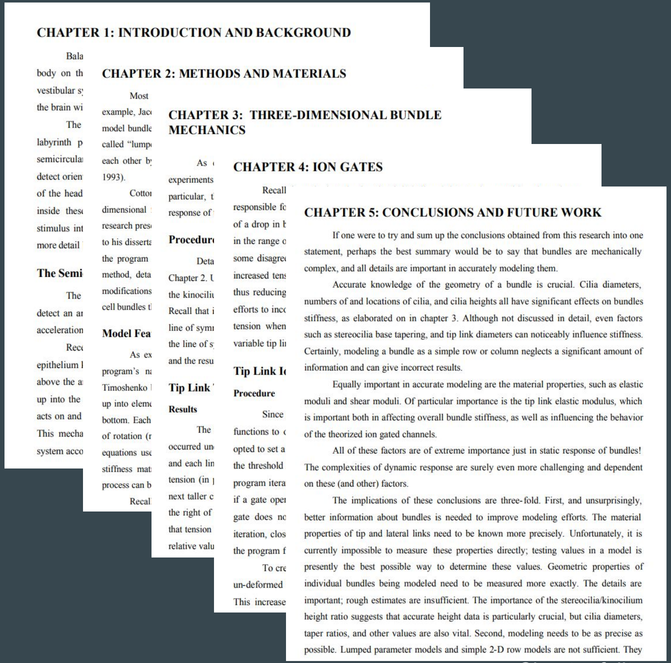


Future works

- ★ Further testing
- ★ Improve coverage and accuracy
- ★ PyMuPDF

Text Extraction

Tested (API version containerized)


CHAPTER 1: INTRODUCTION AND BACKGROUND

CHAPTER 2: METHODS AND MATERIALS

CHAPTER 3: THREE-DIMENSIONAL BUNDLE MECHANICS

CHAPTER 4: ION GATES

CHAPTER 5: CONCLUSIONS AND FUTURE WORK

If one were to try and sum up the conclusions obtained from this research into one statement, perhaps the best summary would be to say that bundles are mechanically complex, and all details are important in accurately modeling them.

Accurate knowledge of the geometry of a bundle is crucial. Cilia diameters, numbers of and locations of cilia, and cilia heights all have significant effects on bundles stiffness, as elaborated on in chapter 3. Although not discussed in detail, even factors such as stereocilia base tapering, and tip link diameters can noticeably influence stiffness. Certainly, modeling a bundle as a simple row or column neglects a significant amount of information and can give incorrect results.

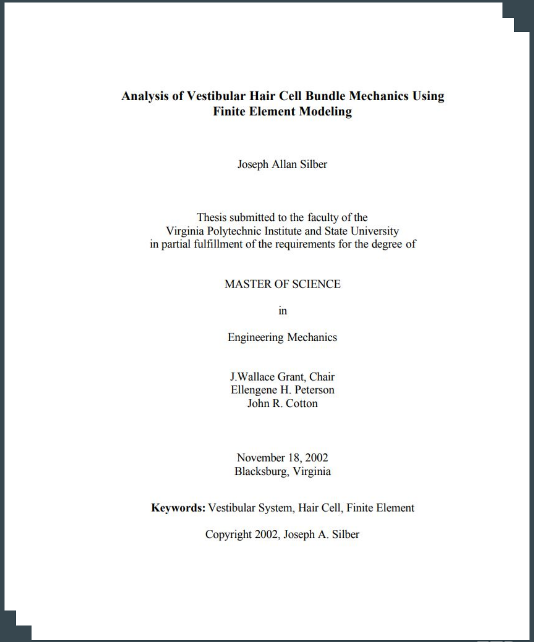
Equally important in accurate modeling are the material properties, such as elastic moduli and shear moduli. Of particular importance is the tip link elastic modulus, which is important both in affecting overall bundle stiffness, as well as influencing the behavior of the theorized ion gated channels.

All of these factors are of extreme importance just in static response of bundles! The complexities of dynamic response are surely even more challenging and dependent on these (and other) factors.

The implications of these conclusions are three-fold. First, and unsurprisingly, better information about bundles is needed to improve modeling efforts. The material properties of tip and lateral links need to be known more precisely. Unfortunately, it is currently impossible to measure these properties directly; testing values in a model is presently the best possible way to determine these values. Geometric properties of individual bundles being modeled need to be measured more exactly. The details are important; rough estimates are insufficient. The importance of the stereocilia/kinocilium height ratio suggests that accurate height data is particularly crucial, but cilia diameters, taper ratios, and other values are also vital. Second, modeling needs to be as precise as possible. Lumped parameter models and simple 2-D row models are not sufficient. They

Chapter fulltext

OR


Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber

ETD

{handle: 9704}

handle

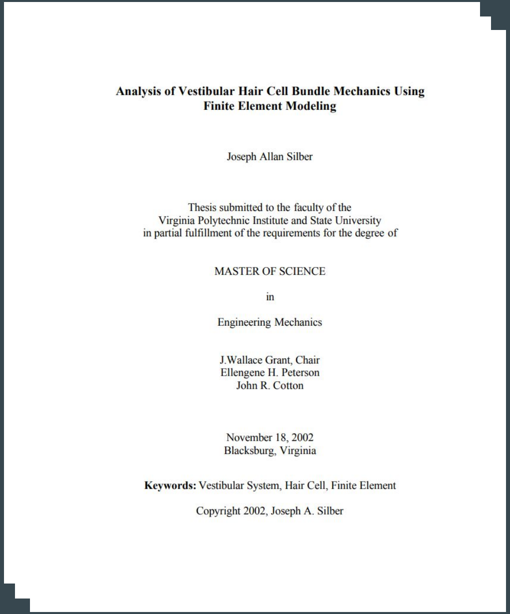
{type: Thesis}

doc_type

Tested (API version containerized)

Text Extraction

- Get one page from PDF file.



Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

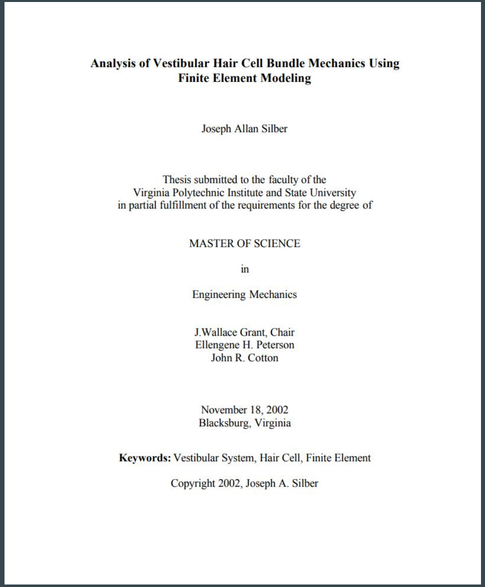
Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber



Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

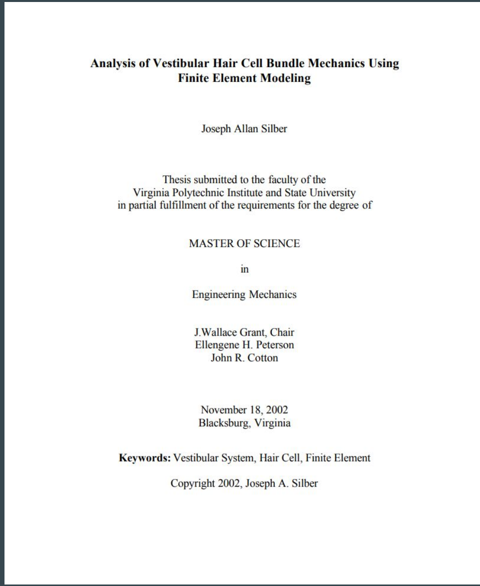
November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber

Text Extraction

- Extract text from this PDF page. And then, write the text into a .txt file. Repeat this steps until finish all PDF pages.



Analysis of Vestibular Hair Cell Bundle Mechanics Using Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

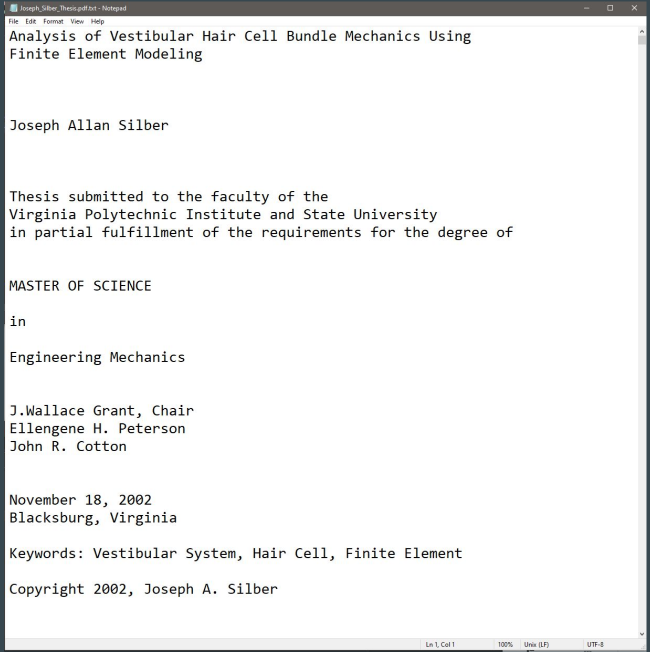
Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber



Analysis of Vestibular Hair Cell Bundle Mechanics Using
Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

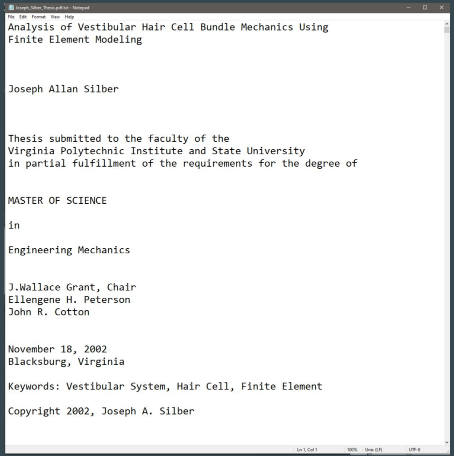
November 18, 2002
Blacksburg, Virginia

Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber

Extracted Text

Classification



Joseph_Silber_Thesis.pdf.txt - Notepad

Analysis of Vestibular Hair Cell Bundle Mechanics Using
Finite Element Modeling

Joseph Allan Silber

Thesis submitted to the faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Engineering Mechanics

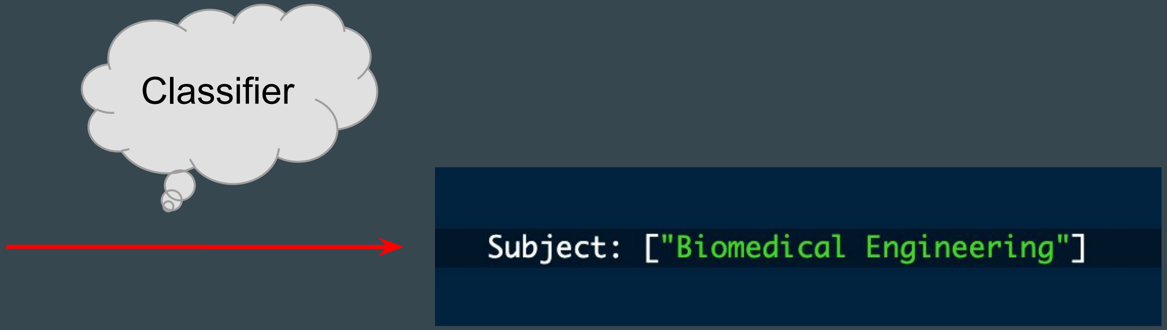
J.Wallace Grant, Chair
Ellengene H. Peterson
John R. Cotton

November 18, 2002
Blacksburg, Virginia

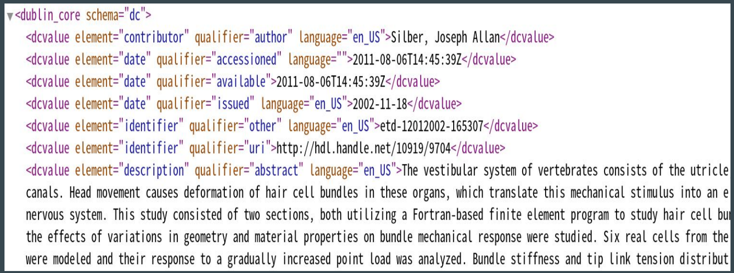
Keywords: Vestibular System, Hair Cell, Finite Element

Copyright 2002, Joseph A. Silber

Extracted Text



Labels for ETD

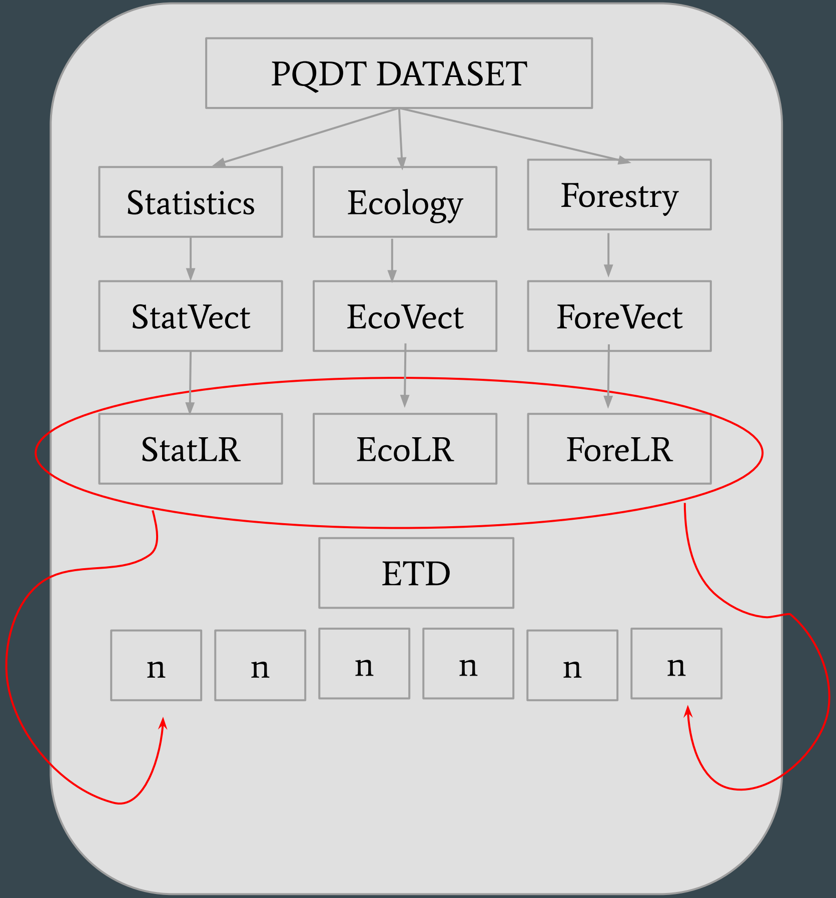



```
<dublin_core schema="dc">
  <dcvalue element="contributor" qualifier="author" language="en_US">Silber, Joseph Allan</dcvalue>
  <dcvalue element="date" qualifier="accessioned" language="">2011-08-06T14:45:39Z</dcvalue>
  <dcvalue element="date" qualifier="available">2011-08-06T14:45:39Z</dcvalue>
  <dcvalue element="date" qualifier="issued" language="en_US">2002-11-18</dcvalue>
  <dcvalue element="identifier" qualifier="other" language="en_US">etd-12012002-165307</dcvalue>
  <dcvalue element="identifier" qualifier="uri">http://hdl.handle.net/10919/9704</dcvalue>
  <dcvalue element="description" qualifier="abstract" language="en_US">The vestibular system of vertebrates consists of the utricle
  canals. Head movement causes deformation of hair cell bundles in these organs, which translate this mechanical stimulus into an e
  nervous system. This study consisted of two sections, both utilizing a Fortran-based finite element program to study hair cell bu
  the effects of variations in geometry and material properties on bundle mechanical response were studied. Six real cells from the
  were modeled and their response to a gradually increased point load was analyzed. Bundle stiffness and tip link tension distribut
```


Dublin Core XML

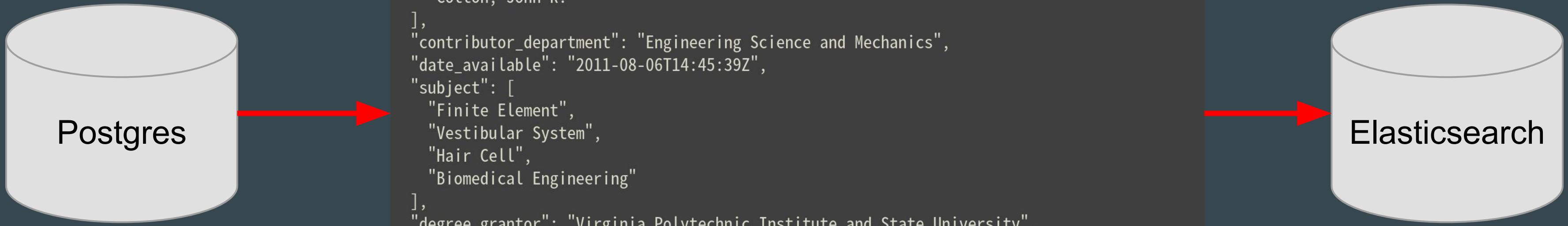
Classification Architecture

- Training Classifier
  - Vectorizers trained on each label
  - Focused Text uses: title, abstract, and keywords
  - Used Logistic Regression to generate label probabilities per *n* word segment
- Implementing Classifier on data
  - Segment ETD into *n* word blocks
  - Classify each segment
  - Cumulate per label probability over segments
  - Pick top *k* labels based on top 5 labels' variance



Ingestion

Read metadata from Postgres and index into Elasticsearch

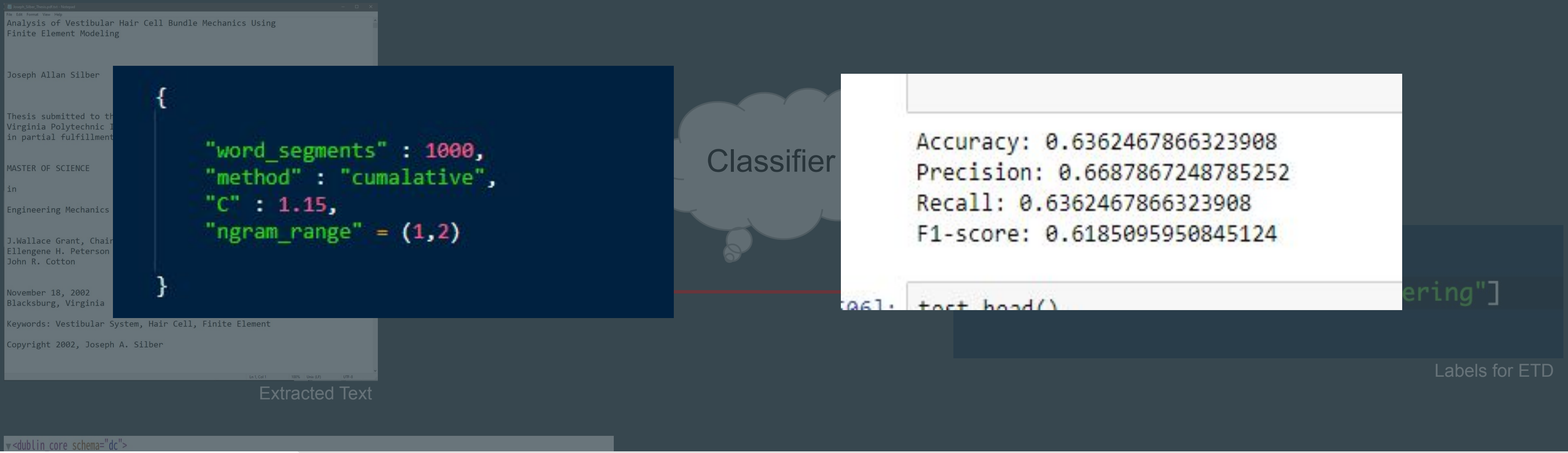



```
{
  "contributor_author": "Silber, Joseph Allan",
  "contributor_committeechair": "Grant, John Wallace",
  "contributor_committeemember": [
    "Peterson, Ellengene H.",
    "Cotton, John R."
  ],
  "contributor_department": "Engineering Science and Mechanics",
  "date_available": "2011-08-06T14:45:39Z",
  "subject": [
    "Finite Element",
    "Vestibular System",
    "Hair Cell",
    "Biomedical Engineering"
  ],
  "degree_grantor": "Virginia Polytechnic Institute and State University",
  "degree_level": "masters",
  "degree_name": "Master of Science",
  "description_abstract": "The vestibular system of vertebrates consists of the utricle, ...
  ...
}
```


Elasticsearch-ready record

Tested (API version containerized)

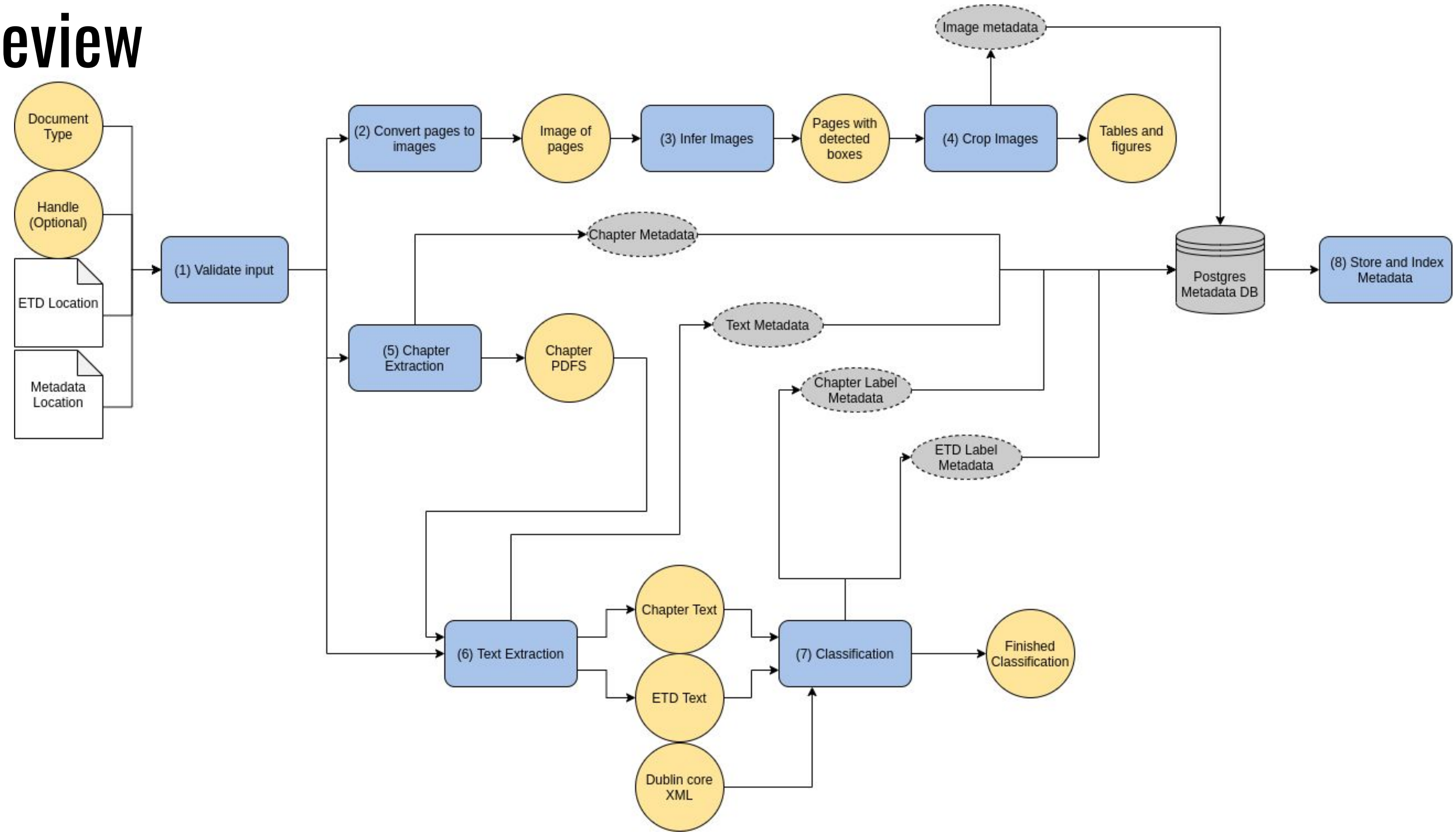
Classification Metrics



Future works

- ★ Test classifier with varied set of hyper-parameters
- ★ Implement zones for metadata weighting
- ★ Try different baseline models

Review



Document Type
Handle (Optional)
ETD Location
Metadata Location
(1) Validate input
(2) Convert pages to images
Image of pages
(3) Infer Images
Pages with detected boxes
(4) Crop Images
Image metadata
Tables and figures
Chapter Metadata
(5) Chapter Extraction
Chapter PDFS
Text Metadata
Chapter Label Metadata
ETD Label Metadata
Postgres Metadata DB
(8) Store and Index Metadata
(6) Text Extraction
Chapter Text
ETD Text
Dublin core XML
(7) Classification
Finished Classification

Lessons Learned

- Understand Intended Structure Early
- Microservices Workflow
- Cross-team Communication

Next Steps

- Expose more Airflow goals to provide further customization
- Expose new Airflow goals to allow for separate collections in Elasticsearch