MSRI
Mathematical Sciences Research Institute

17 Gauss Way Berkeley, CA 94720-5070 p: 510.642.0143 f: 510.642.8609 www.msri.org

NOTETAKER CHECKLIST FORM

(Complete one for each talk.)

Name: Malgorzata Marciniak **Email/Phone:** mmarciniak@lagcc.cuny.edu 5734620411

Speaker's Name: Daniela Ushizima

Talk Title: Applications of convnets to microstructural description and material design

Date: 10 /01 /2018 **Time:** 3 : 30 am / pm **(circle one)**

Please summarize the lecture in 5 or fewer sentences:
Imagine a terabyte of data in a minute. Come up with computer program that can analyze this data. Strategies of approaching the large database of materials. Software that tackles detection, segmentation and classification of materials (carbon fibers, concrete, CMC, etc). One of the main challenges is how to couple increasing data rate experiments to new Data Science methods in support of more automated analytical tasks for scientific discovery.

CHECK LIST

(This is **NOT** optional, we will **not pay** for **incomplete** forms)

- ☑ Introduce yourself to the speaker prior to the talk. Tell them that you will be the note taker, and that you will need to make copies of their notes and materials, if any.
- ☑ Obtain ALL presentation materials from speaker. This can be done before the talk is to begin or after the talk; please make arrangements with the speaker as to when you can do this. You may scan and send materials as a .pdf to yourself using the scanner on the 3rd floor.
  - **Computer Presentations**: Obtain a copy of their presentation
  - **Overhead**: Obtain a copy or use the originals and scan them
  - **Blackboard**: Take blackboard notes in black or blue **PEN**. We will **NOT** accept notes in pencil or in colored ink other than black or blue.
  - **Handouts**: Obtain copies of and scan all handouts
- ☑ For each talk, all materials must be saved in a single .pdf and named according to the naming convention on the "Materials Received" check list. To do this, compile all materials for a specific talk into one stack with this completed sheet on top and insert face up into the tray on the top of the scanner. Proceed to scan and email the file to yourself. Do this for the materials from each talk.
- ☑ When you have emailed all files to yourself, please save and re-name each file according to the naming convention listed below the talk title on the "Materials Received" check list. (*YYYY.MM.DD.TIME.SpeakerLastName*)
- ☑ Email the re-named files to notes@msri.org with the workshop name and your name in the subject line.

Speaker: Daniela Ushizima

Title: **Applications of convnets to microstructural description and material design**

Note Taker: Malgorzata Marciniak

Imagine a terabyte of data in a minute. Come up with computer program that can analyze this data systematically.

CAMERA (Director James Sethian): Build the applied mathematics that can accelerate scientific discovery at DOE experimental facilities and deliver it as robust user-friendly software. Coordinated team of applied mathematicians, beam scientists, computational chemists, computer scientists, material scientists, statisticians, image and signal processors, …

OUTLINE:

1. Image analysis at Berkeley Lab
2. 2D
   a) Scattering Patterns and HipGISAXs
   b) Ie earchortng and ranking
3. 3D
   a) MicoCT with NASA
   b) Concrete with UC Engineering

Machine Learning, which use statistical methods to enable machines to improve with experiences is a subset of Artificial Intelligence (any technique which enables computers to mimic human behavior). Deep Learning is a subset of Machine Learning which make the computation of multi-layer neural network feasible.

In 2009 ImageNet began in Princeton.

Machine learning can be supervised (with labels) or unsupervised.

Tasks: Classification, Detection, Segmentation. Use-case: Scattering Patterns (decide whether bcc or hcp class).

Voronoi diagram (X metric space with distance d)

$$R_k = \{x \in X | d(x, P_k) \leq d(x, P_j) \; for \; all \; j \neq k\}$$

How to query image collections? By the content. Google searching system is not useful for scattered patterns. Easy to find popular items.

pyCBIR is based on Python, acronym Content Based Image Retrieval (https://bids.berkeley.edu/news/searchable-datasets-python-images-across-domains-experiments-algorithms-and-learning). Is a new visual search engine foe scientific image retrieval based on pictorial similarity. This tool is capable of retrieving relevant images using datasets across science domain in real time using compact data representation. Enable investigation of abstract patterns by leveraging historical data gathered by domain experts at a high cost. Improve researchers collaboration across scientific communities.

http://crd.lbl.gov/news-and-publications/news/2017/recognition-software-drives-matches-across-multiple-science-domains/

GISAX results: accuracy rates, times of retrieving the set,

“Convolutional Neural Network-based Screening in Crystallography” identification of Bragg spots from massive diffraction patterns datasets data driven deep learning for X-ray crystallography images obtained as LCLS. In serial crystallography a full dataset requires 10^2 to 10^5 diffraction patterns from several crystals, but only a fraction of theimages contain Bragg spots. Using a CNN similar to AlexNet images fall into 3 categories: “hit”, “miss” or “maybe”.

3D: nanoparticles, carbon textiles, ceramic matrix composites, geological samples.

Analyze microstructure of materials with applications to aviation, civil engineering, geology, etc.

Use-case: Microtomography

NASA’s Mars science laboratory mission: protect the rover during landing in a a woven carbon fiber fabric. MicroCT of the woven carbon fiber, challenges in carbon fiber analysis: segmentation (training, prediction, regularization), still preliminary results of segmentation algorithm performance.

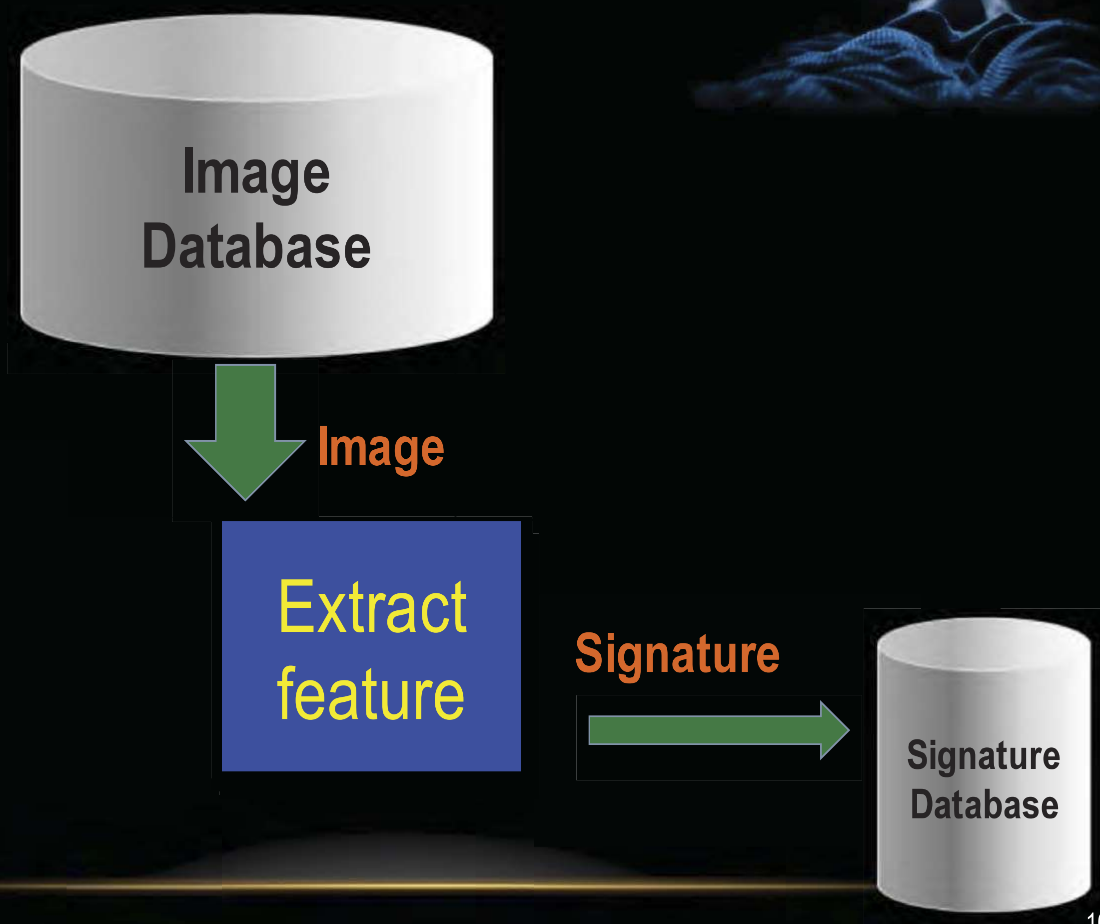
Other projects: supersonic parachute (https://www.youtube.com/watch?v=mTAbj8aRVvg),

WHY PYCBIR?

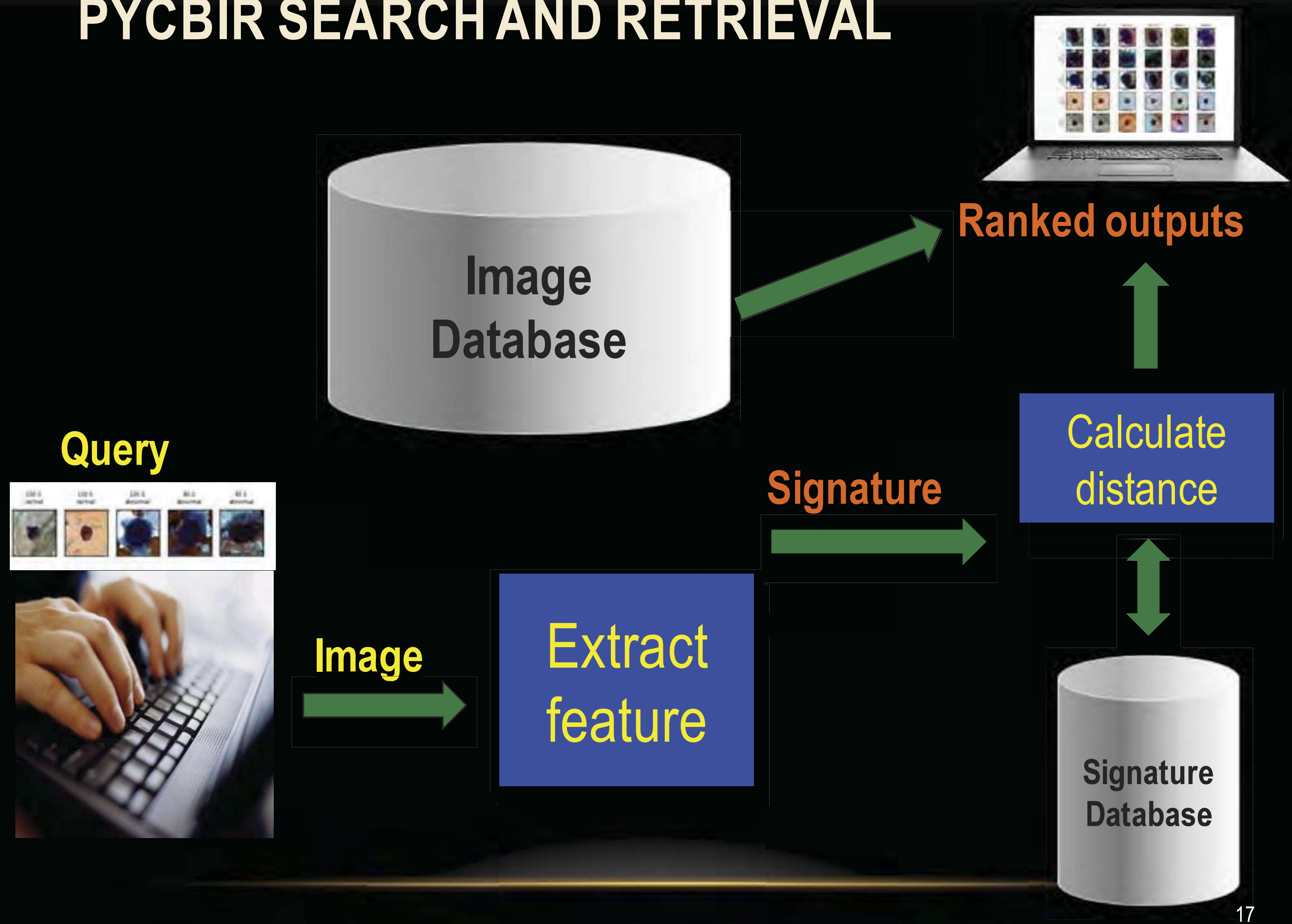
- New python tool for content-based image retrieval (CBIR);
- Query by example: capable of searching relevant items in large databases, given image samples;
- pyCBIR allows general purpose investigation across image domains;
- Our experiments: can we recover high-level abstraction from data using:

a. Color, texture, shape?
b. Learn signatures using CNN?
c. Similarity = distance?

PYCBIR TRAINING



PYCBIR SEARCH AND RETRIEVAL

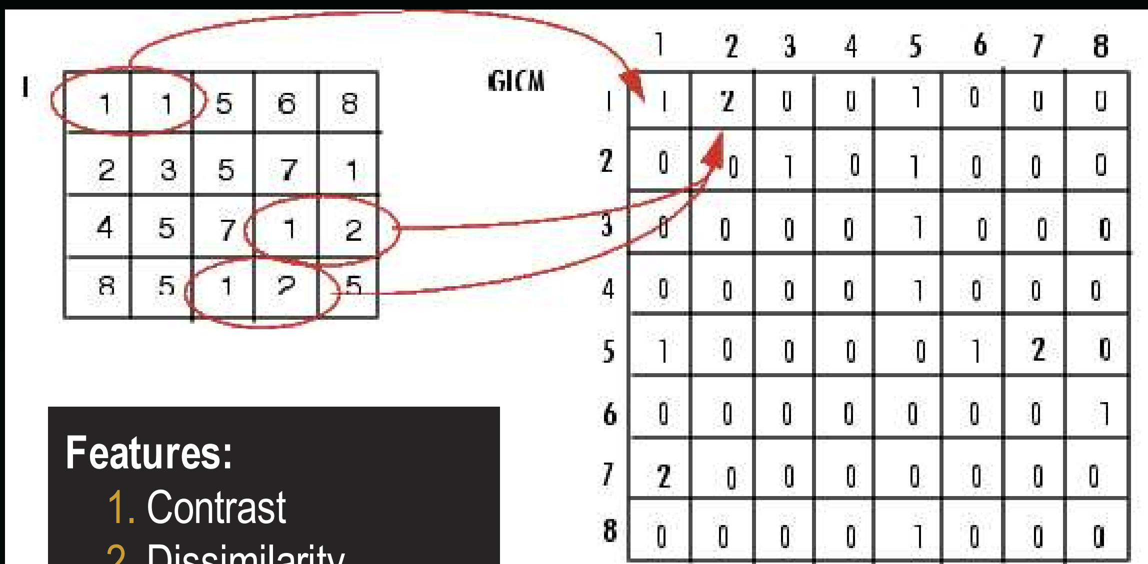


FEATURE EXTRACTION METHODS

- Signature = index = feature vector = descriptors;

1. Gray Level Co-Occurrence Matrix;
2. Histogram of Oriented Gradient;
3. First Order Texture Features;
4. Local Binary Pattern;
5. Convolutional Neural Network.

GRAY LEVEL CO-OCCURRENCE MATRIX (GLCM)



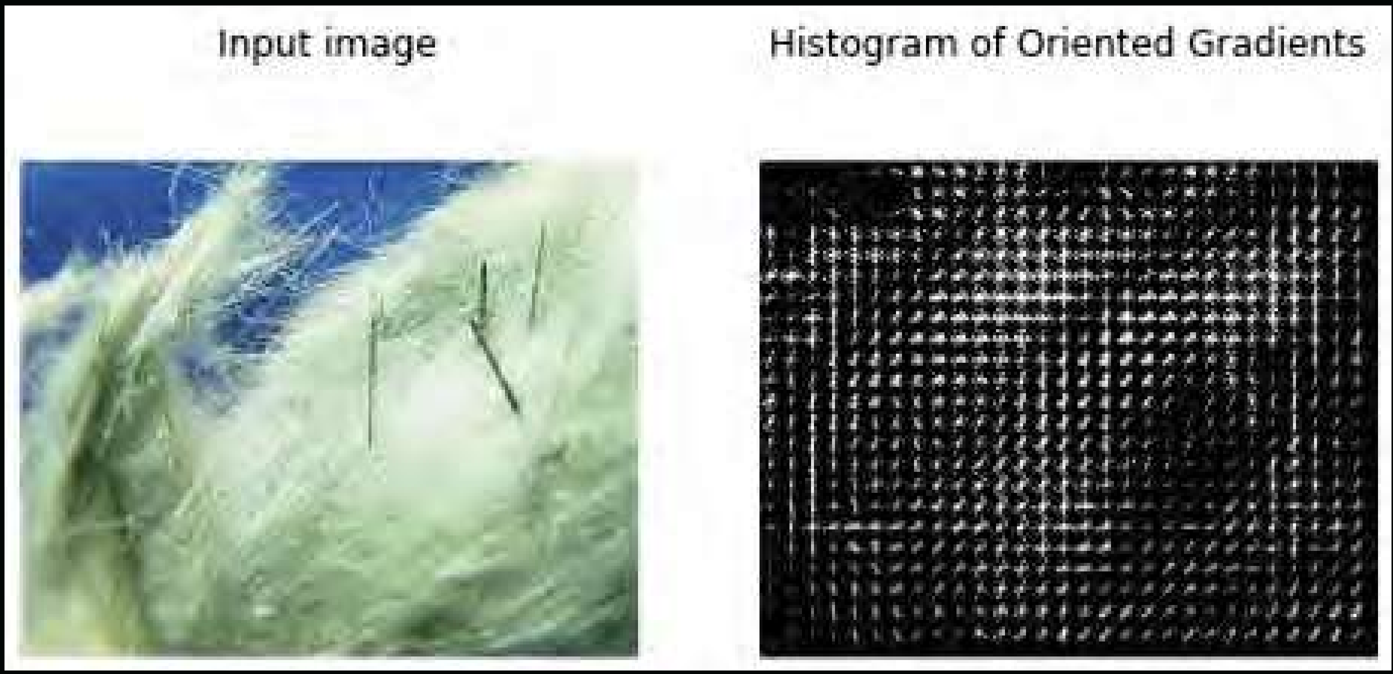
1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

	1	2	3	4	5	6	7	8
1	1	2	0	0	1	0	0	0
2	0	0	1	0	1	0	0	0
3	0	0	0	0	1	0	0	0
4	0	0	0	0	1	0	0	0
5	1	0	0	0	0	1	2	0
6	0	0	0	0	0	0	0	1
7	2	0	0	0	0	0	0	0
8	0	0	0	0	1	0	0	0

Features:

1. Contrast
2. Dissimilarity
3. Homogeneity
4. Energy
5. Correlation
6. ASM

HISTOGRAM OF ORIENTED GRADIENTS



Histogram of oriented gradients of a Describable Textures Dataset (DTD) image.

(a) (b)

*Source: dataset at https://www.robots.ox.ac.uk/~vgg/data/dtd/

FIRST ORDER TEXTURE FEATURES

- Mean

$$\mu = \sum_{i=0}^{G-1} i p(i)$$

- Variance

$$\sigma^2 = \sum_{i=0}^{G-1} (i - \mu)^2 p(i)$$

- Skewness

$$\mu_3 = \sigma^{-3} \sum_{i=0}^{G-1} (i - \mu)^3 p(i)$$

- Kurtosis

$$\mu_4 = \sigma^{-4} \sum_{i=0}^{G-1} (i - \mu)^4 p(i) - 3$$

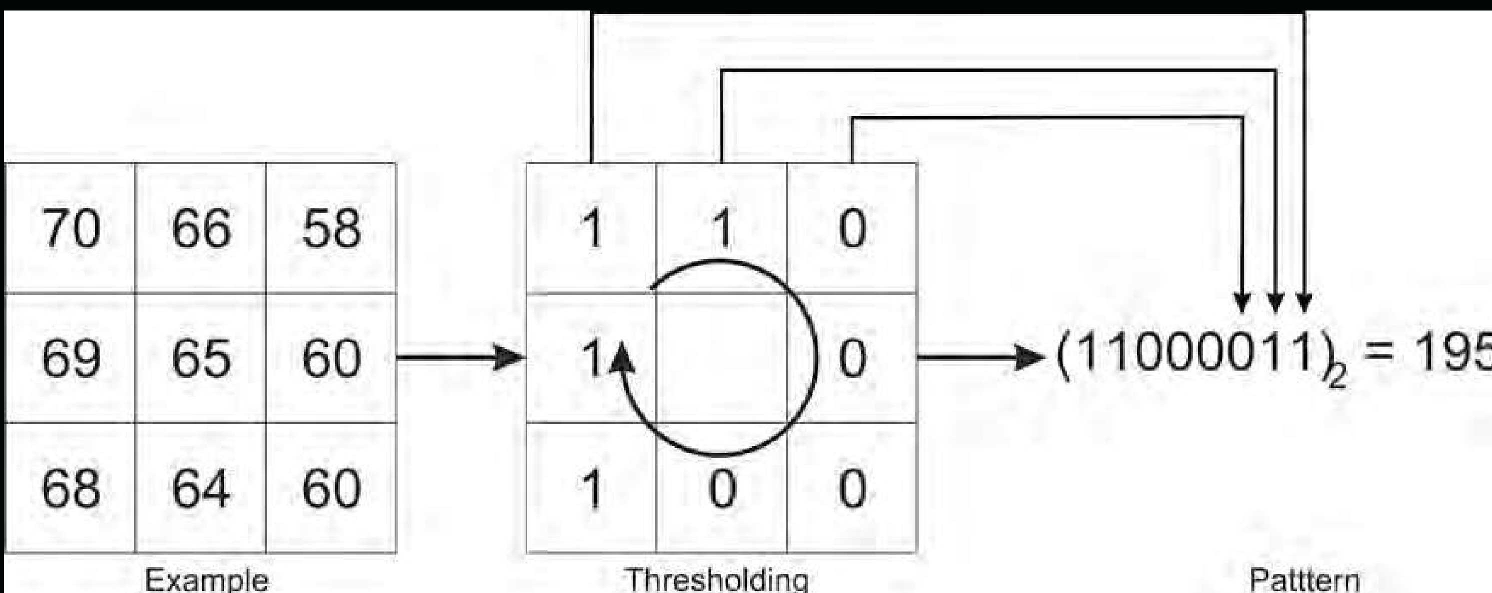
- Energy

$$H = -\sum_{i=0}^{G-1} p(i) \log_2 [p(i)]$$

- Entropy

$$E = \sum_{i=0}^{G-1} [p(i)]^2$$

LOCAL BINARY PATTERN (LBP)



CONVOLUTIONAL NEURAL NETWORK (CNN)

- We used the CNN in two different ways:
  1. Trained with the **same** database of the image retrieve;
     - 2 convolutional layers;
  2. Trained with the **imageNet** Database: CNN Inception*
     - 6 convolutional layers;

TensorFlow

theano

*Available in: http://arxiv.org/pdf/1602.07261v1.pdf

SIMILARITY

- IF Image = multidimensional vector,

  THEN similarity = distance!

1. Euclidean
2. Infinity
3. Cosine
4. Pearson
5. Chi-Square
6. Kullback-Liebler Divergence
7. Jeffrey Divergence
8. Kolmogorov-Smirnov Divergence
9. Cramer-von Mises Divergence
10. Cityblock Distance

DISTANCE METRICS

- Euclidean Distance

$$d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Infinity Distance

$$d(\mathbf{x},\mathbf{y}) = max_{i=1}^{n} |x_i - y_i|$$

- Cosine Similarity

- Pearson Correlation Coefficient

$$s(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

- Chi-Square Dissimilarity

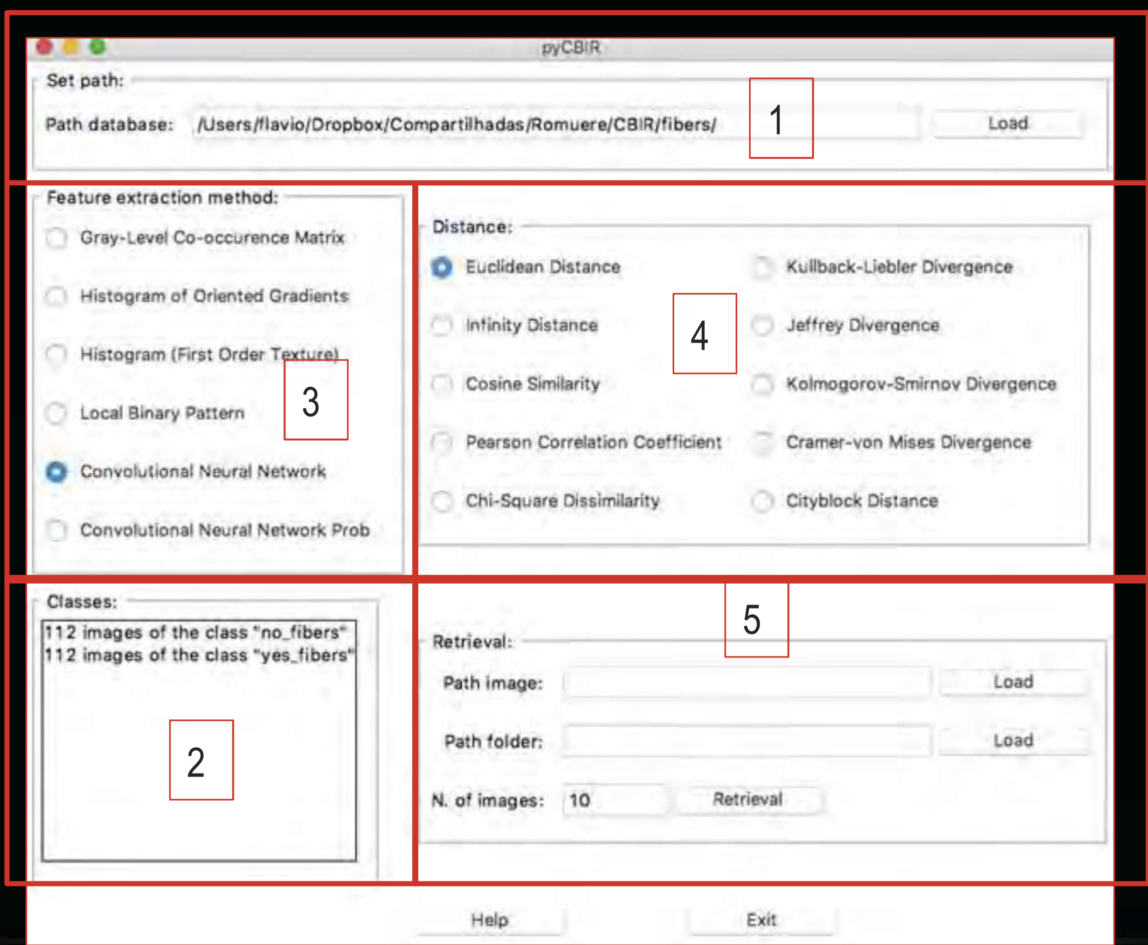
$$d(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

DISTANCE METRICS

- Kullback-Liebler Divergence $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{y_i}$
- Jeffrey Divergence $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i \log \frac{x_i}{\mu_i} + y_i \log \frac{y_i}{\mu_i}$
- Kolmogorov-Smirnov Divergence $d(\mathbf{x}, \mathbf{y}) = max_{i=1}^{n} |X_i - Y_i|$
- Cramer-von Mises Divergence $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} (X_i - Y_i)^2$
- Cityblock Distance $(L_1)\ d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i|$

HOW TO ORGANIZE THE DATABASE?

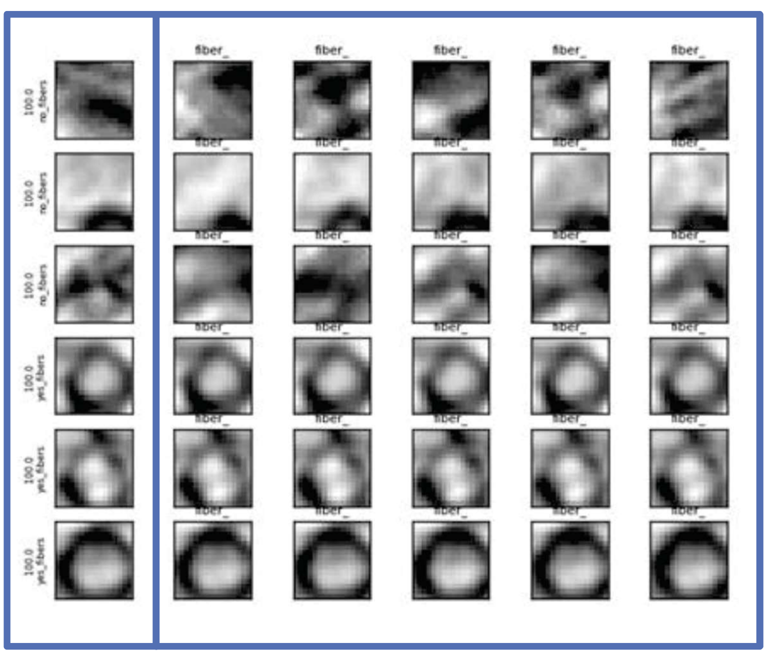
GRAPHICAL USER INTERFACE



EXPERIMENTS - FIBERS DATASET

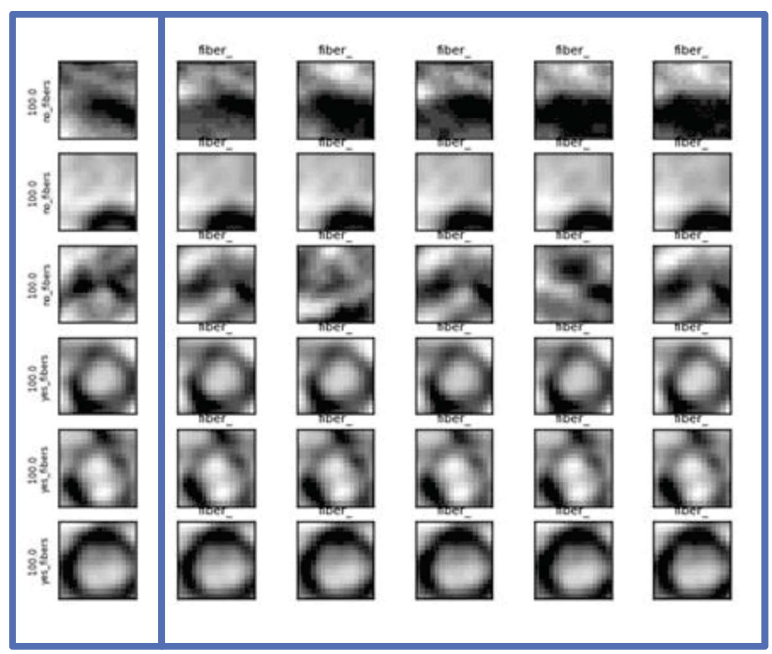
4,000 images of 16 X 16 for two balanced classes.

Query | Top 5 retrieved



Result obtained using the CNN trained with the same dataset.

Query | Top 5 retrieved



Result obtained using the inception network.

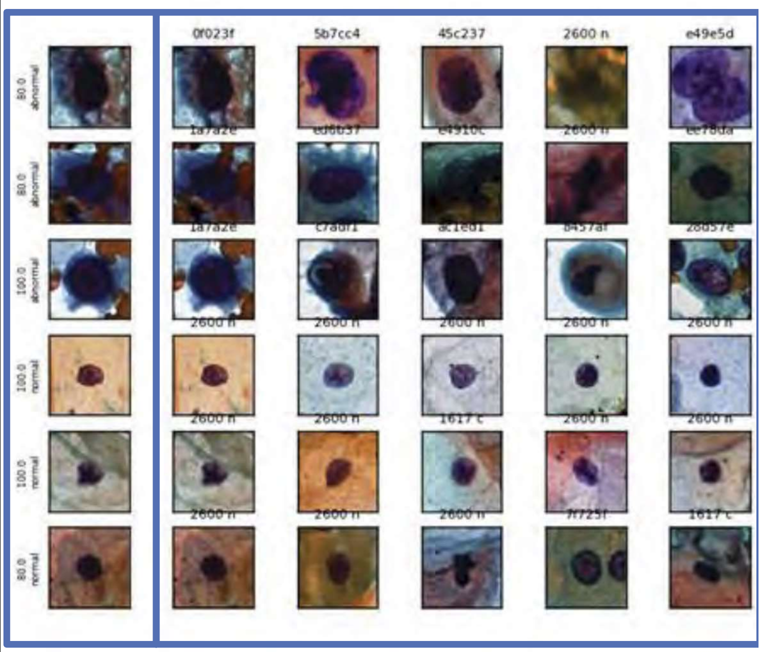
EXPERIMENTS - TIME

	Training	Extraction of features for the whole database	Top 10 retrieved for a query image
Approach 1 (same DB)	3.4 minutes	9 seconds	4 seconds
Approach 2 (inception)	-	29 minutes	15 seconds

EXPERIMENTS - CELLS DATASET

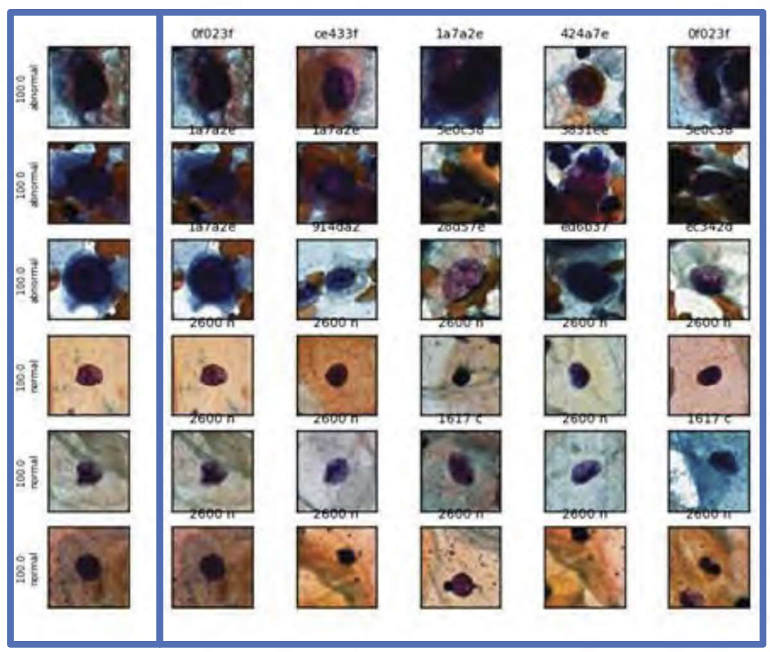
1,886 images of normal cells and 1,509 of abnormal - 100 X 100 pel;

Query | Top 5 retrieved



Result obtained using the CNN trained with the same datase.

Query | Top 5 retrieved



Result obtained using the inception network.

EXPERIMENTS - TIME

	Training	Extraction of features for the whole database	Top 10 retrieved for a query image
Approach 1	94 minutes	48 seconds	5 seconds
Approach 2	-	23 minutes	12 seconds

CONCLUSIONS

Approach 1 (same DB)	Approach 2 (inception)
Advantages	
Feature extraction faster after training	Doesn't need training
Training done only once	
Each image = 256 features	
Disadvantages	
For datasets with big images and a lot of classes the training is slow	Feature extraction is slow
	Each image = 2,048 features