Plankton Classification

Predict ocean health, one plankton at a time

Using Machine Learning Techniques

HARSH PARIKH 2011CS10240

SAHIL LOOMBA 2012CS10114

About the Challenge

- Plankton are critically important to our ecosystem.
- Traditional methods for measuring and monitoring plankton populations are time consuming. Improved approaches are needed.
  - One such approach is through the use of an underwater imagery sensor.
  - Need for automated algorithms to classify captured images.

Data Source: National Data Science Bowl

Getting acquainted with the Data

- Data is in the form of low-resolution grayscale images.
- Training data: 30,366 images
  - *121 classes: planktons(116) + unknown(3) + artifacts/junk(2).*

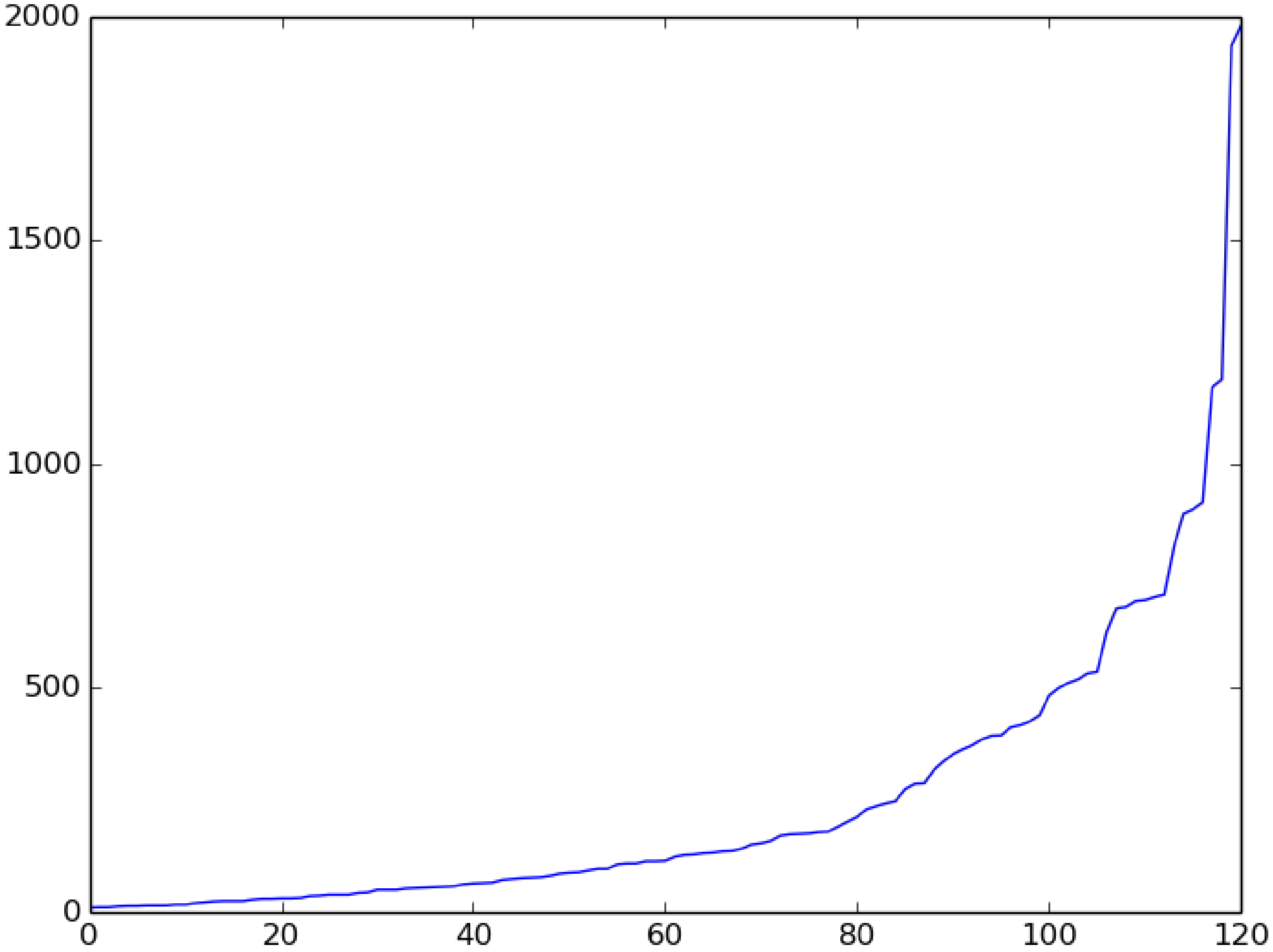
amphipod

unknown blob

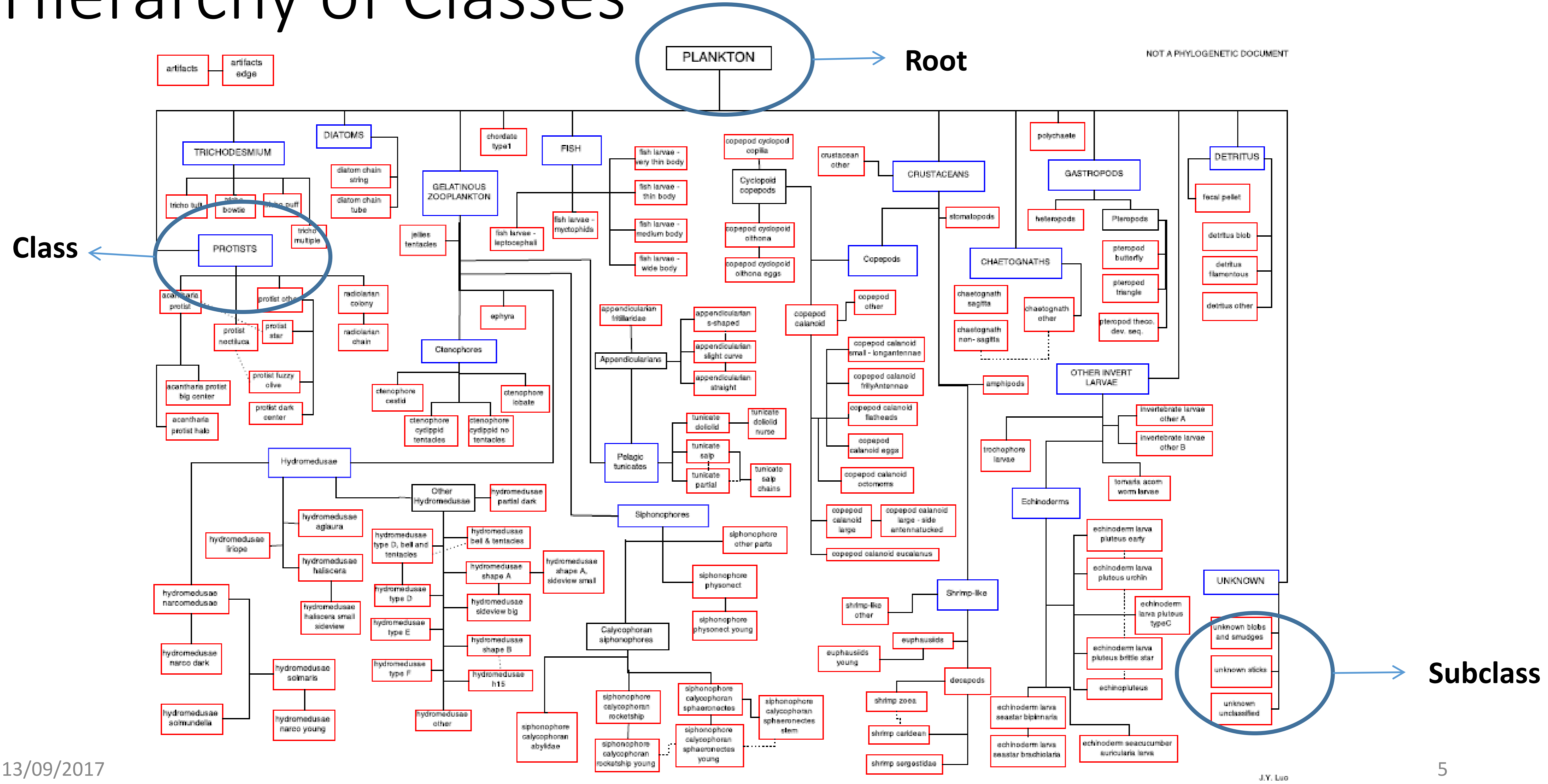
artifact

- Test data: 1,30,400 unlabelled images.
- Training data is skewed: *disproportionate number of images across classes*
  - *from as low as 9 to as high as 1979.*

Skewed Data Distribution across classes

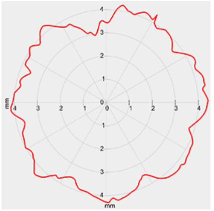


Hierarchy of Classes

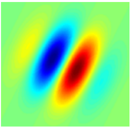


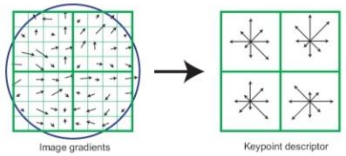
Data Features

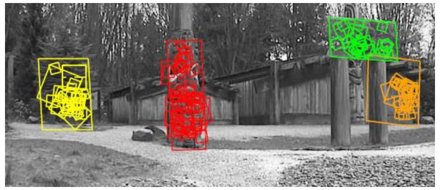
- Raw pixel data
  - Shape Descriptors
    - Circularity Constant
    - Hu Moments
  - Orientation Descriptors
    - Gabor Filtered Data
    - HoG Vectors
  - Feature Descriptors
    - SIFT
  - Information Descriptors
    - Fourier Analysis

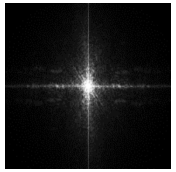


$$\phi_1 = \eta_{20} + \eta_{02}$$
$$\phi_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$
$$\phi_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$







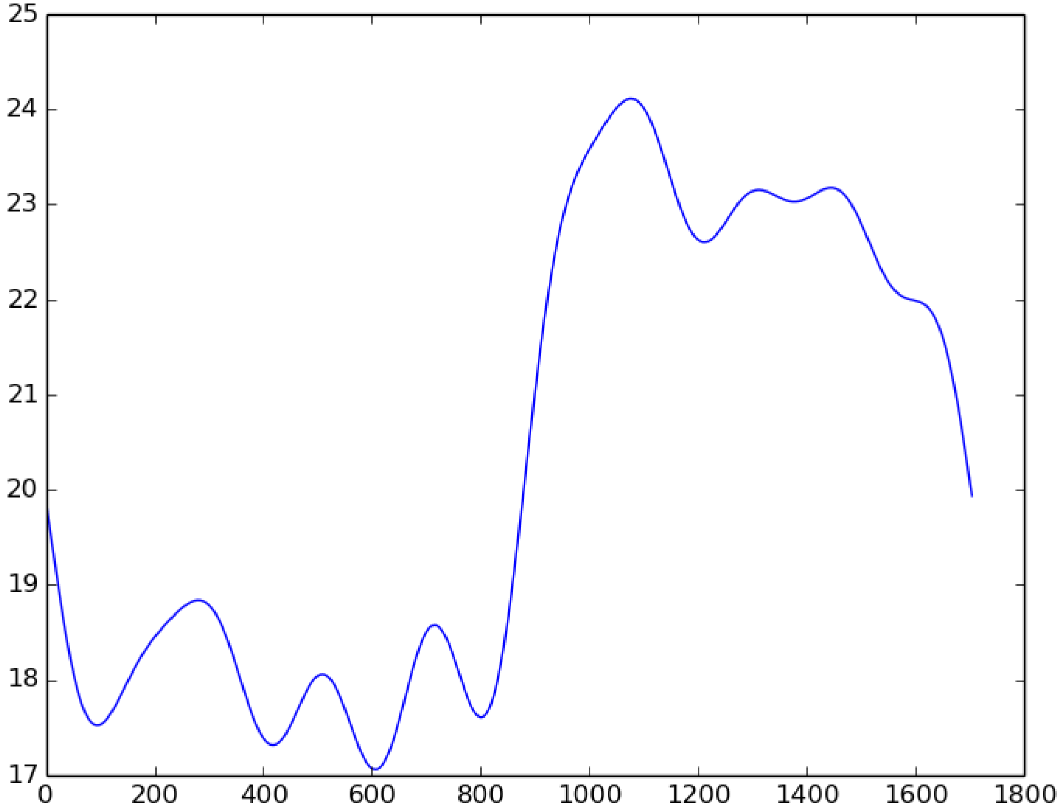


ML Techniques

- Classical Techniques:
  - Logistic Regression
  - Multiclass SVM with Gaussian Kernel
    - Holds well for Orientation and Feature Descriptors
  - Random Forests

- Deep Learning:
  - Artificial Neural Networks
  - Convolutional Neural Networks

Shape Descriptors Analysis – 1

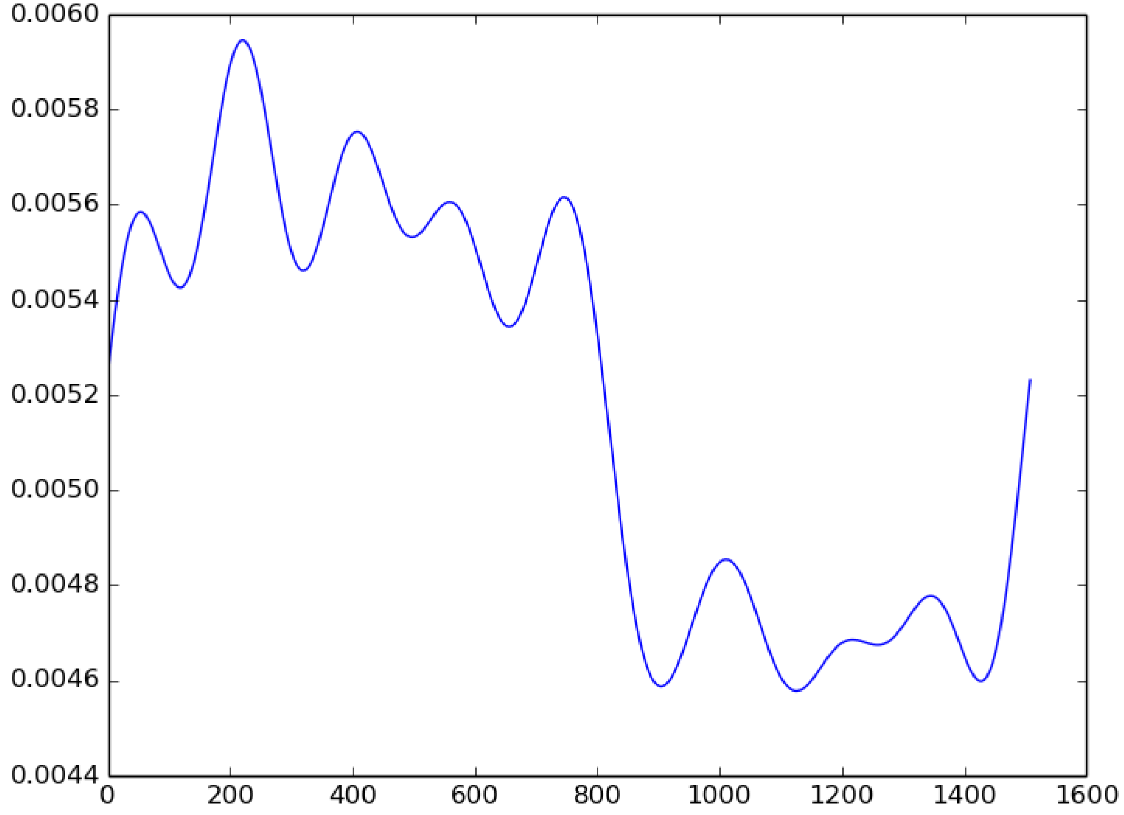


Circularity Constant for **acantharia protist** images and **chaetognath non sagitta** images

Acantharia protist

Chaetognath non sagitta

Shape Descriptors Analysis – 2



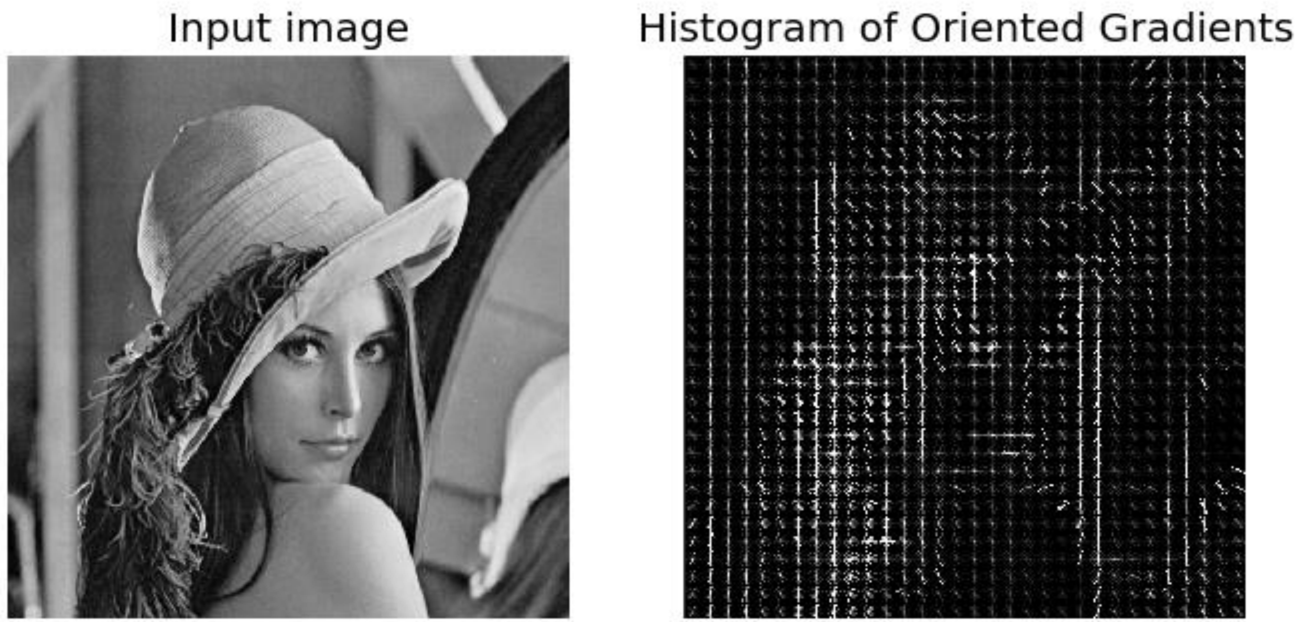
Hu Moment-1 for **chaetognath sagitta** images and **chaetognath non sagitta** images

Chaetognath sagitta

Chaetognath non sagitta

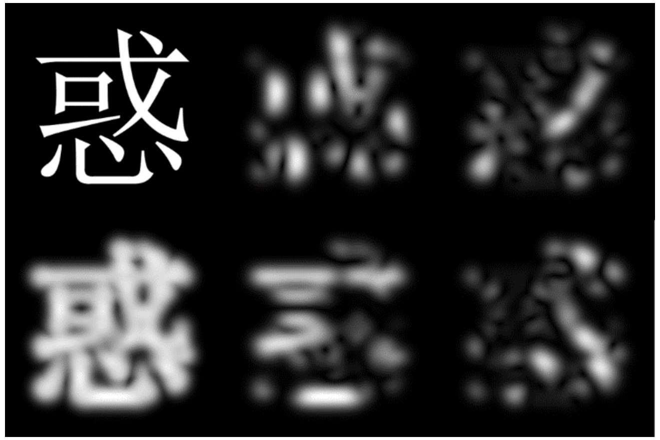
Orientation Descriptors Analysis - 1

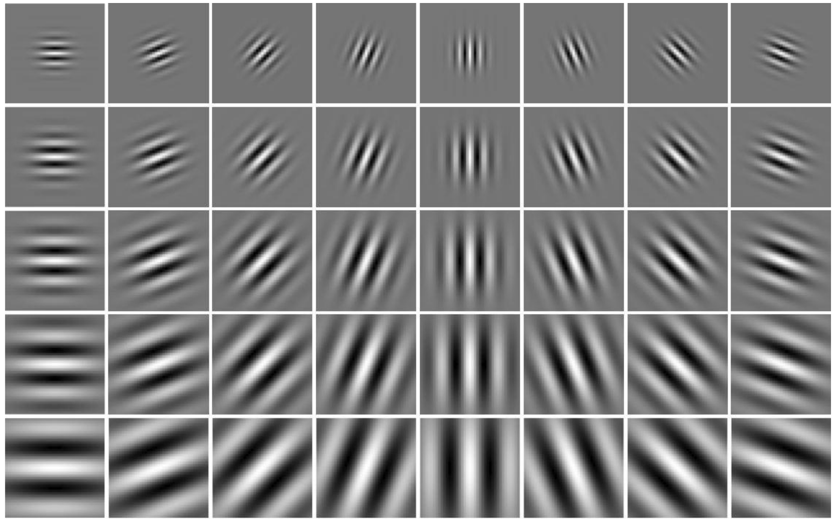
- HoG (Histogram of Oriented Gradients) data.
- Vector (size 64) fed to a multiclass SVM with a Gaussian Kernel. (60% Training + 40% Validation.)
- Best parameters led to a maximum mean accuracy of 25.90%.



Orientation Descriptors Analysis - 2

- Gabor Filter data.
  - Using Gabor Bank.
- Vector (size 70) fed to a multiclass SVM with a Gaussian Kernel. (60% Training + 40% Validation.)
- Best parameters led to a maximum mean accuracy of 49.88%.





Feature Descriptors Analysis - SIFT

Raw Images

SIFT

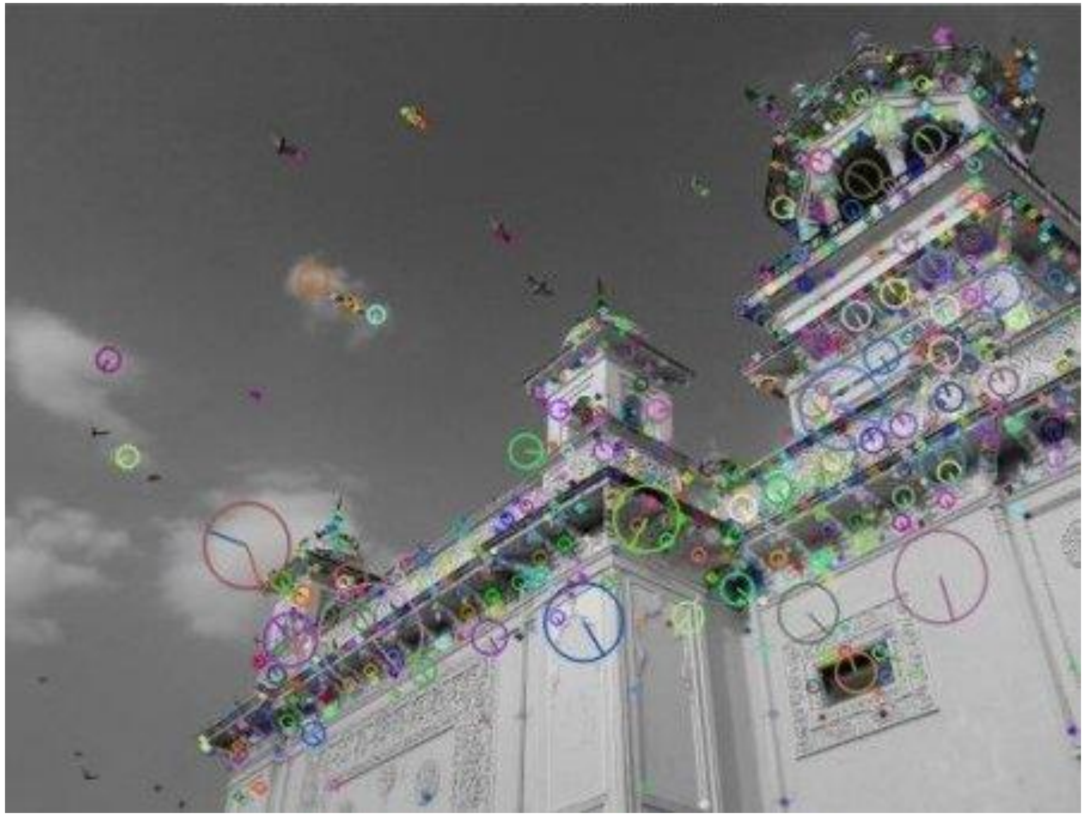
Bag of Key Point Vectors

K-means

Vector of size 16 for each image

Training

- SVM: 39.92%
- ANN: 19.23%



Training on Raw Pixel Data

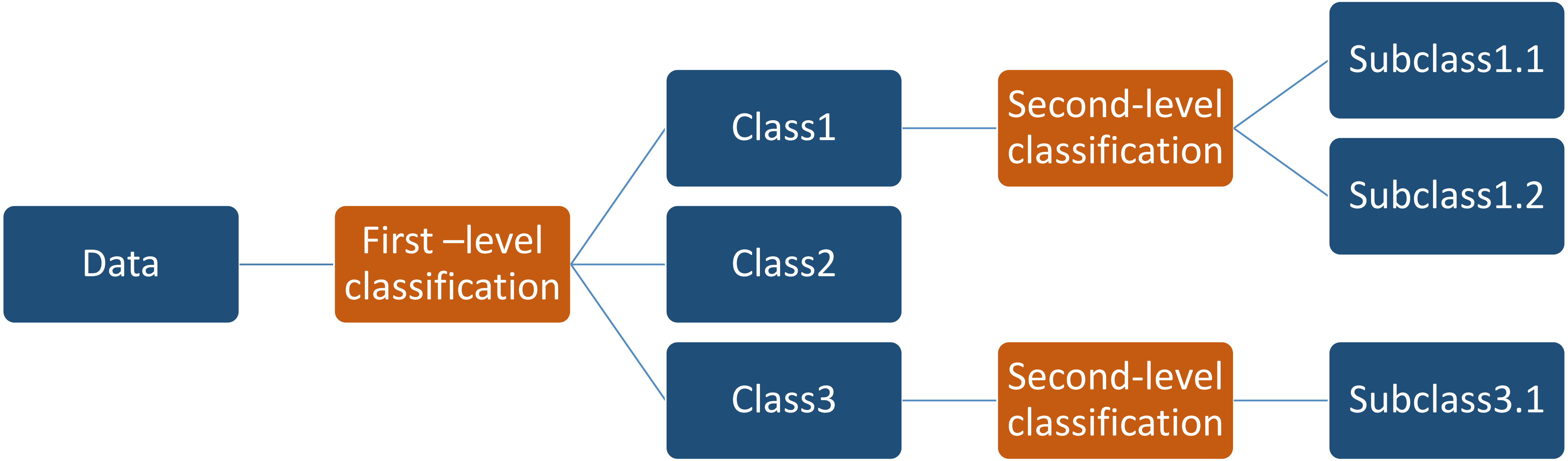
Model	**Validation Log Loss**	**Validation Accuracy (Top-1)**
Crude ANN	20.38	32%
Crude CNN	4.18	56%
ANN → SVM	2.38	58%

- Training Dataset 60% + Validation Dataset 40%

Future Work – 1

- *Revision Theory*
  - Reiterate over data with larger deviation from the true value.
- *Augmenting Multiple Feature Vectors*
  - For example: Matrix Product of Gabor and SIFT vectors.
- *Random Forest for Image Classification*
- *Large-Scale Object Classification using Label Relation Graphs*
  http://web.eecs.umich.edu/~jiadeng/paper/deng2014large.pdf
  - Hierarchy and Exclusion (HEX) graphs, a new formalism that captures semantic relations between any two labels applied to the same object.

Future Work – 2: *Hierarchical Paradigm*



Thank You. Questions?