Unsupervised Learning and Image Classification in High Performance Computing Cluster

Itauma Itauma, Xuewen Chen

2016 HPCC Systems Engineering Summit - Community Day
@hpccsystems

Data Sciences and Analytics Lab
Wayne State University
Department of Computer Science
College of Engineering
Detroit, MI

@amightyo

October 12, 2016

Overview

1. Introduction
   - Data Sciences and Analytics Lab
   - Objective
   - Background
   - Contributions of this Study
2. Methods
3. Experiments and Results
   - Evaluation on AR Face Database
   - Identity Recognition on the Wild and Multimedia Database
4. Discussion and Observation
   - ECL Codes
5. Conclusion and Next Steps

Data Sciences and Analytics Lab

discovery.cs.wayne.edu/lab_website/

Data Sciences and Analytics Lab @ Wayne State University

Home | Projects | Recent Publications | News and Updates | Software and Services | Forum | Contact

Home

At DSAL (Data Sciences and Analytics Lab) at Wayne State University, we focus on studying, researching and developing algorithms that are applicable to various big data analysis problems occur in areas of Machine Learning, Data Mining, Bioinformatics, and Healthcare Informatics.

Faculty

Recent Posts

New Post

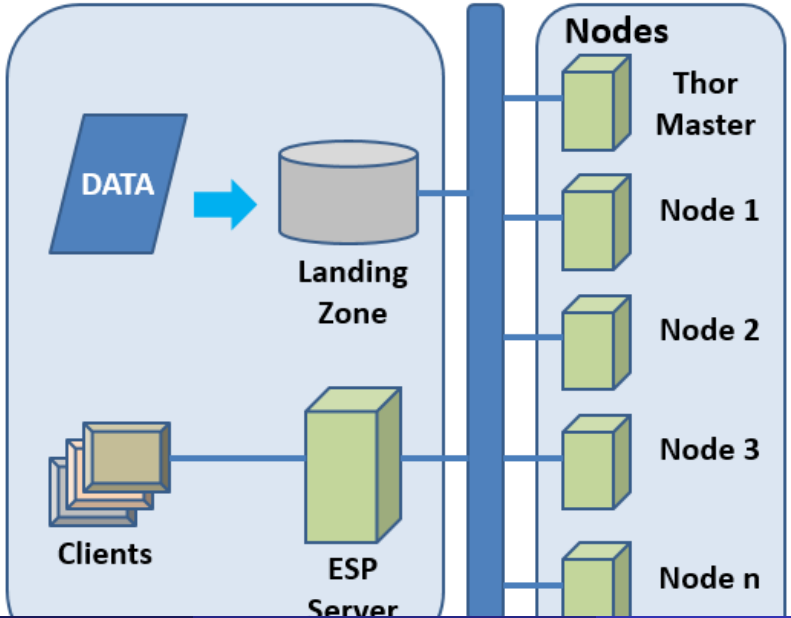
Tags

Objective

Unsupervised Learning and Image Classification in HPCC

- Feature representation from face images and sound data
- Classification based on the representations.

Background: Single and Multinode HPCC Architecture



Background: HPCC Systems Platform Architecture

- The experiments were run on a single node HPCC system platform.
- The single node HPCC system VM running Ubuntu 64-bit was setup as a guest machine on Windows.
- The host machine has a usable RAM size of 3.89GB of which 1.5GB is allocated to the guest machine.
- An important feature of the HPCC platform is its ability to ETL Big Data (tera bytes to peta bytes)
- At the core of the HPCC platform is the powerful and simple ECL programming language.

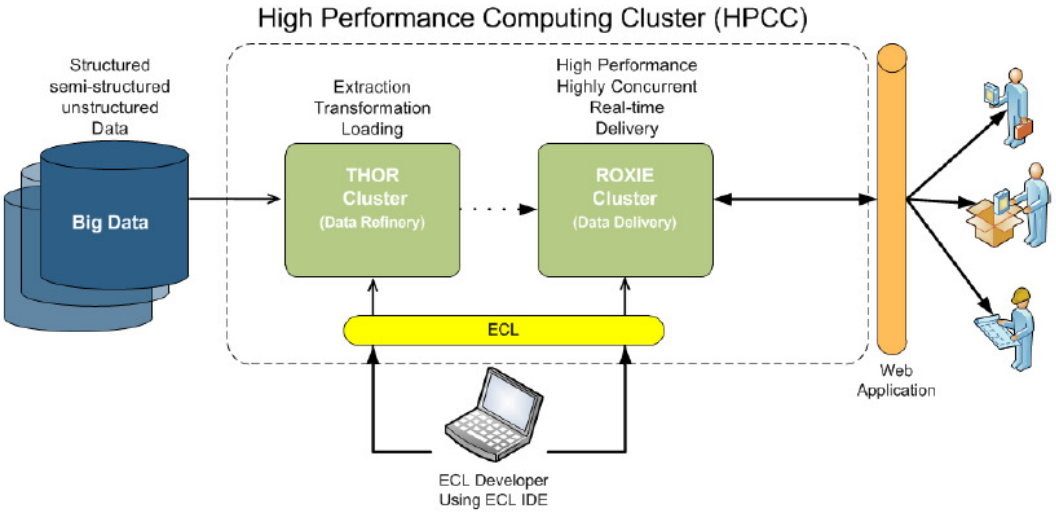
Big Data

- Any data matrix that can't fit into the RAM
- How much RAM do I need to store a matrix?
- A *rxc* matrix double precision matrix requires $r * c * 8/10^9$GB

Rows	Cols	GB	Fits into 0.75GB RAM
8300	250000	16.6	No
8300	62500	4.15	No
1400	10000	0.1	Yes
1000	10000	0.08	Yes

Table: Gigabytes of storage a matrix requires

HPCC Systems Platform Architecture



Contributions of this Study

- The use of HPCC platform in the implementation of the feature learning and object classification tasks.
- A feature learning and recognition framework using a multimodal strategy.
- This idea utilizes the HPCC platform to handle identity recognition using multimedia data in a multimodal framework which leads to high recognition accuracy.
- By dividing a face image into several subunits, we can extract intra-region information more precisely.

Methods: Project Pipeline

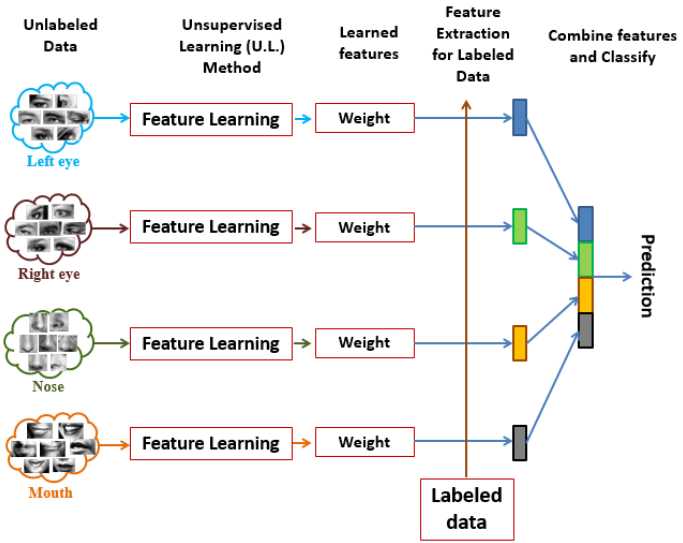


Figure: Framework for image classification.

Image Reading In HPCC

- Extract patches from images.
- Normalize the patches.
- Convert these patches to CSV.
- Spray the CSV to HPCC.

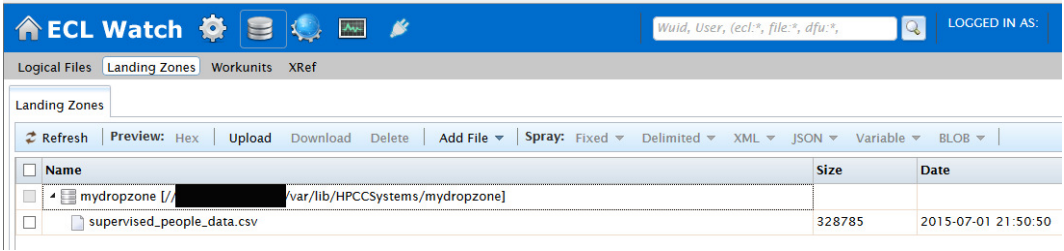


Image to CSV

What you see

RGB

Grayscale

Images converted to the same dimension

50 x 50

What the computer sees

88	61	60	67	93	134	84	53	58	83
118	71	45	42	77	128	128	75	49	60
89	67	56	48	57	81	79	71	78	84
74	69	77	79	75	79	72	77	85	83
95	79	72	70	72	77	86	85	80	74
83	71	62	62	67	70	76	85	89	89
79	76	78	91	100	100	102	98	90	91
109	108	104	105	103	95	95	100	102	105

Pixel values in an image region

1 x 2500

145	145	134	126

Given 3 images, a 3 x 2500 matrix is written to a CSV file.

Feature Learning

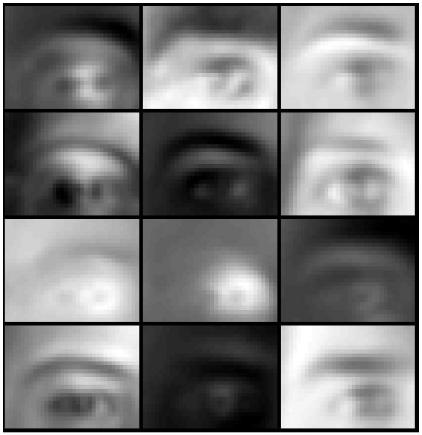
- The goal of unsupervised learning is to find a lower dimensional projection of the unlabeled data that preserves all the information in the data while reducing redundant dimension.
- The algorithm takes the dataset X and outputs a function $f : R^n \rightarrow R^k$ that maps an input vector $x^{(i)}$ to a new feature vector of k features.

- a good number of samples
- choice of parameters
- number of weights

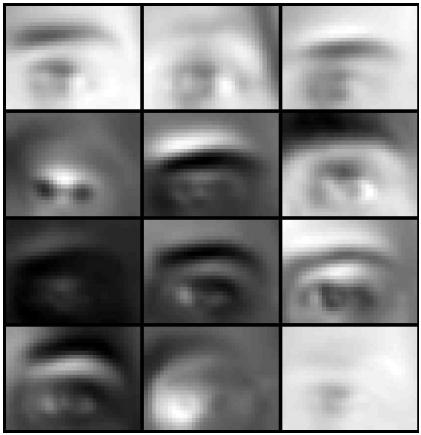
K-Means Clustering Algorithm

Select k points as initial centroids
repeat
  Form k clusters by assigning each point to its closest centroid
  Re-compute the centroids of each cluster
until convergence criterion is satisfied

Visualization of learned weight from AR Face Dataset



"(AR Left eye Weight)"



"(AR Right eye Weight)"

Figure: Selected bases (or centroids) trained on AR images using K-Means in HPCC.

Classification

Classifiers

1. C4.5 Decision Tree
2. Random Forest
3. Naive Bayes
4. Softmax

In machine learning, all classification algorithms follow three logical steps:

- Learning the model from a training set
- Testing with respect to obtaining measures of how well the classifier fits
- Classifying which involves testing the model on new data in order to compute a classification accuracy.

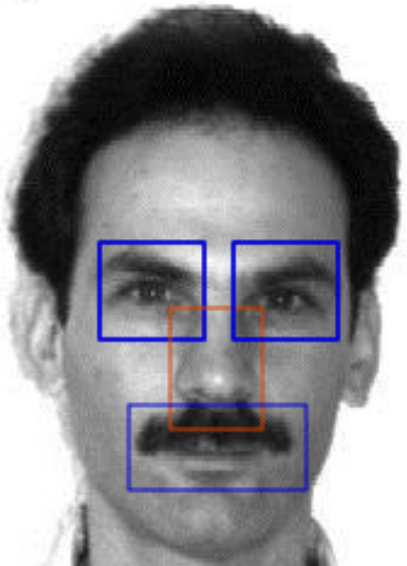
AR Data

- AR Face Database:
  - Aligned: 100 subjects
  - 10 different images per subject
- Occluded images removed

Figure: Example images from one subject in *AR* database with various facial expressions and illumination.

Region of Interest

Better representations obtained by running unsupervised learning for each region.



AR Face Results

Methods	Acc. (%) with Test *5 Train-Test*
K-means (HPCC) + Linear SVM (MATLAB)	69.3
K-means + Linear SVM [Coates et al., 2011]	74.3
K-means (H) + C4.5 Decision tree (H)	78.9
K-means (M) + C4.5 Decision tree (H)	**85.2**

Table: Comparison of face recognition rates on *AR* database.

AR Face Results

Region	Recognition rate (%)
Whole face	84.2
Left eye	76.4
Right eye	77.1
Nose	65.3
Mouth	57.9
Multimodalities	**91.5**

Table: Average recognition rates deviations on *AR* database with the use of features from each single facial region and the multimodal representation, respectively. C4.5 decision tree is used for the classification.

PubFig83 Data

- PubFig83 Aligned Face Images:
  - 10 subjects
  - 100 images for each subject

Figure: Example images of 10 celebrities with various real-world changes on facial expression, pose, illumination, occlusion, resolution, etc.

Identity recognition results using visual and/or speech contents on multimedia database.

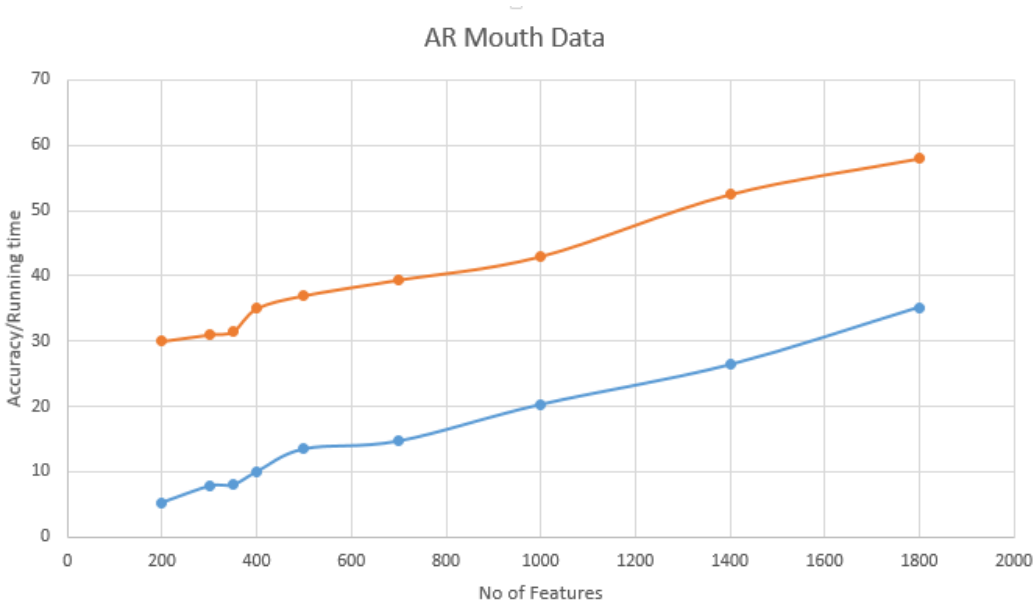
Methods	Acc. (%) *50 Train-Test*
Only Visual Content	
Naive Bayes (HPCC)	55.5
Linear SVM (MATLAB)[Coates et al., 2011]	71.6
Random Forest (HPCC) - maxLevel=10	74.2
Random Forest (HPCC) - maxLevel=15	75.5
Softmax Classification (HPCC)	83.0
C4.5 Decision tree (HPCC) - maxLevel=25	**91.5**
Only Speech Content	
C4.5 Decision tree (HPCC)	**91.6**
Multimedia (Visual + Speech)	
C4.5 Decision tree (HPCC)	**94.0**

Table: Identity recognition results using visual and/or speech contents on multimedia database.

Discussion and Observation

- Feature learning and object classification successfully performed in HPCC
- Single node HPCC platform was used.
  - A faster training time is expected with a multinode HPCC platform (future studies)
- C4.5 decision tree classification method achieved higher accuracy compared with other alternatives.
  - The deeper the tree, the more complex the decision rules and the fitter the model.

Discussion and Observation



Discussion and Observation

ECL Codes

c, [US] https://github.com/amightyo/ImageClassification_HPCC_ICMLA2015

SAE_People.ecl	ICMLA15 ECL codes	9 months ago
SoftMaxPeople.ecl	ICMLA15 ECL codes	9 months ago
SpeechDecisionTree.ecl	ICMLA15 ECL codes	9 months ago
computeAccuracy.ecl	ICMLA15 ECL codes	9 months ago
crossAssignmentsLayout.ecl	ICMLA15 ECL codes	9 months ago
kmeansSpeechMod.ecl	ICMLA15 ECL codes	9 months ago
multimodalDecisionTree.ecl	ICMLA15 ECL codes	9 months ago

Readme.md

Unsupervised Learning and Image Classification in High Performance Computing Cluster

Presented at IEEE ICMLA'15 in Miami, Florida.

All experiments were performed using a single node HPCC Systems.

Citation:
Itauma Itauma, Melih S. Aslan, Flavio Villanustre, and Xue-wen Chen. "Unsupervised Learning and Image Classification in High Performance Computing Cluster." In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), pp. 576-581. IEEE, 2015.

Contact Itauma Itauma @amightyo

Conclusion

- This study demonstrates feature learning and object classification problems in HPCC with faster optimization/calculation of algorithms and low cost of hardware.
- Several databases with image and multimedia content were explored in HPCC.
- Our learning framework leverages new representations that are learned over multimedia data automatically.
- C4.5 decision tree classification method boosts recognition rates.
- This identity recognition algorithm can lead to further exploration of face recognition problems.

Next Steps

- Setting up a Multinode HPCC Lab for Big Data processing
- Explore teaching ECL with a focus on Matrix Computation
- Explore more ML applications using HPCC

References

HPCC Systems: HPCC Client Tools. Boca Raton Documentation Team, 2014.

Coates, Adam; Lee, Honglak; Ng, Andrew Y. An analysis of single-layer networks in unsupervised feature learning, International Conference on Artificial Intelligence and Statistics, pp. 215–223, 2011..

Martınez, AM and Benavente, Robert, The AR Face Database, Computer Vision Center, Technical Report, vol. 24, 1998.

Becker, Brian C. and Ortiz, Enrique G., Evaluating Open-Universe Face Identification on the Web, Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops, pp. 904–911, 2013.

K-Means Clustering Algorithm, `http://en.wikipedia.org/wiki/K-means_clustering`.

Thank you!

Questions?