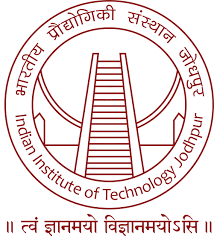
**Learning Document Structure**

**For**

**Retrieval And Classification**

Chinmay Garg (B16CS041) Anurag Shah (B16CS034)

Mentor: Dr Gaurav Harit

*Department of Computer Science and Engineering*

*Indian Institute of Technology, Jodhpur*

**Motivation**

The sole biggest motivation behind choosing this project was our fascination towards machine learning.

Machine Learning is the rave of the moment, as it is gradually becoming the brain behind business intelligence.

We wished to implement machine learning for document classification as it is more challenging due to handling of voluminous and highly non-linear data, generated exponentially in the era of digitization. Proper representation of documents increases efficiency and performance of classification, ultimate goal of retrieving information from large corpus.

And hence, in this project we will be implementing a already proposed algorithm for document classification.

**Objective**

To retrieve “similar” documents from a large heterogeneous collection of document images.

**Sub-Objective**

To classify the Arabic dataset provided on the basis that whether they contain a table or not.

**Challenges/Research Issues:**

1. Integration of OpenCV libraries with a C++ IDE.
2. Poor documentation of functions and classes implemented in OpenCV libraries on the official website of OpenCV.
3. Size of a matrix that can be constructed is limited (around 4700000 bytes). Thus, limited number of descriptors can be extracted from a limited number of images for codebook creation.
4. Very small dataset available for training and hence during testing on externally provided images, accuracy obtained was not up to the expectation.
5. The algorithm proposed in the provided paper – Horizontal Vertical Partitioning has almost non-existent documentation.
6. No proper reasoning given behind choosing a specific value behind any constant in this paper.

**Methodology/Algorithm:**

1. SURF (Speed Up Robust Features) algorithm for extracting descriptors from each image.
2. K-means algorithm on extracted descriptors for codebook creation.
3. SURF algorithm and HVP (Horizontal Vertical Partitioning) algorithm for writing features.
4. RFC (Random Forest Classifier) algorithm for classification.

**Results:**

1. For Hessian Threshold = 8000, maximum descriptors = 10000,

On training data of 150 relevant and irrelevant images each,

When cross-validated on 66 relevant and 252 irrelevant data,

Accuracy = 96.8558%.

1. For Hessian Threshold = 60000, maximum descriptors = 50000,

On training data of 150 relevant and irrelevant images each,

When cross-validated on 56 relevant and 233 irrelevant data,

Accuracy = 93.5374%.

1. On manual testing of external images:

Relevant: 8/11 images classified correctly.

Irrelevant: 19/19 images classified correctly.

**Conclusion:**

The algorithm proposed in the paper was successfully implemented.

The accuracies obtained were nearly same as that expected in the paper (97.2%).

Efficiency of HVP and RFC were successfully tested.

More importantly we got our first brush with machine learning, document page segmentation, document classification and OpenCV libraries.

**References:**

https://docs.opencv.org/3.0-beta/doc/tutorials/tutorials.html

https://docs.opencv.org/3.1.0/d2/d75/namespacecv.html

https://docs.opencv.org/3.1.0/modules.html

https://docs.opencv.org/3.1.0/annotated.html

https://kevinhughes.ca/tutorials/opencv-install-on-windows-with-codeblocks-and-mingw

https://stackoverflow.com/questions

https://lampsrv02.umiacs.umd.edu/projdb/project.php?id=72

https://pdfs.semanticscholar.org/31da/8cace27a08e6339145f95f190a77197f01b2.pdf

https://www.quora.com/

https://github.com/opencv/

https://github.com/

Date: April 16, 2018 Mentor’s signature