Prediction of Genre from Movie Posters

Neha Yadav
College of Information and Computer Sciences
University of Massachusetts Amherst
140 Governors Drive, Amherst, MA
nyadav@cs.umass.edu

1 Introduction

For movie viewers, the movie posters are one of the first impressions which humans use to get cues about the movie content and its genre. Humans can grasp the cues like color, expressions on the faces of actors etc to quickly determine the genre (horror, comedy, animation etc). Valdez [1] Psychology research has shown that color characteristics of an image like hues, saturation, brightness, contour etc. affect human emotions. A given situation arouses these emotions in humans. If humans are able to predict genre of a movie by a single glance at its poster, then we can assume that the color characteristics, local texture based features and structural cues of posters possess some characteristics which could be utilized in machine learning algorithms to predict its genre. Hence, visual features based on colors and structural cues have been extracted from movie posters to detect its genre.

There has been huge success in the area of image classification and object recognition using deep learning methodologies. It would be interesting to see if a trained convolution network could learn high level features from raw image pixels and able to identify it's genre. Therefore, apart from extracting visual features from movie posters and training on traditional classification method, I have also trained a convolution neural network on raw pixels.

To the extent of my knowledge there is no standard dataset present for genre classification from movie posters. The dataset for this task is collected from IMDB. The task of genre classification is a complex process as a movie can be classified into multiple genre. In our dataset, at max three genre are assigned to a given movie. The purpose of my classification task is to classify movie into a single genre. Hence, for movie posters who belong to more than one genre, I have picked the most prominent label. There are 25 unique genre categories present in the dataset.

2 Related Work

The genre classification based on some additional information like synopsis etc have recently gained some attention. In paper [10], low level features were extracted from the movie trailers to classify the movie into four genre categories namely comedy, action, drama and horror. They collected 100 movie previews and extracted low level features from the video tracks.

In paper [5], low level features based on color and edges were extracted from poster images and experiments were conducted on a set of 1500 posters with 6 movie genres. They transformed the multi-label classification problem into single-label classification using two ways where multiple genres were grouped together to form a single genre category, while in other case class labels were transformed into a set of ordered pairs. The first element of each pair is the movie poster and the second one is the class label. Hence if a movie poster belong to 3 genres then 2 ordered pair would be created.

In paper [2], GIST, CENTRIST and W-CENTRIST features were extracted from series of shots from movie trailers and scene categorization was performed to automatically learn scene classes and build vocabulary of movie trailers. Each trailer was represented as segmented 2D histogram of scene categories using bag of words model. The experiments were conducted on 1238 annotated trailers and movie is classified into four most frequent genres: action, comedy, drama, and horror.

Paper[3] uses the same visual features as used in Paper[2], four computable video features (average shot length, color variance, motion content and lighting key) and visual effects are combined together to classify the movie genre. Three major genre categories: Action, Drama (including comedy, drama, romance) and Thriller (or Horror) were used for movie classification.

Paper[7] uses the movie posters and synopsis to classify movie genre. Visual textures like color, edge, texture and number of faces were extracted from movie posters. Two support vectors machine were trained on movie posters and synopsis independently and ”OR” of their prediction is used to predict the final result.

As most of the recent work is focused on extracting visual features from movie trailers or using some meta data like synopsis to predict movie genre. It would be really interesting to learn if traditional classification method like random forest would be able to predict movie genre if trained only on visual features extracted from movie posters or deep neural network trained on raw image pixels would be able to learn important feature to classify movie genre.

3 Methodology

Informative visual image features are computed to use machine learning algorithms which can model these features well and optimally separate each genre category. Below are the different types of features I have extracted for the task of Genre classification.

3.1 Features

Motivated by the idea that visual features like texture and color plays an important role in genre classification. Following are the image features used for the genre classification task.

3.1.1 Local Binary Patterns

Local Binary patterns (LBP) compute a local representation of texture. This local representation is constructed by comparing each pixel with its surrounding neighborhood of pixels. I have taken 40 points in a circularly symmetric neighborhood of radius 8 as parameters to compute LBP features. For each movie poster, a image descriptor of length 42 is obtained.

3.1.2 Statistics, Color, Moment and Hu Moments

Statistics include the mean, standard deviation, skewness of each image in HSV color space. This is a 9 dimensional feature vector. Hu Moments of an image, are used to characterize the outline or ”silhouette” of an object in an image. Hu Moments feature vector length is 7. Hence the total dimension of the image moments feature vector is 16.

3.1.3 3D color histogram in the HSV color space

HSV(Hue, Saturation, Value) color space maps pixel intensities into a cylinder. 3D Histogram estimate the probability P of a pixel color C occurring in image I. For the task of genre classification, I have used 8 bins for the Hue channel, 24 bins for the saturation channel, and 3 bins for the value channel, yielding a total feature vector of dimension 8 x 24 x 3 = 576. The size of feature vector is independent of image dimensions. 3D HSV color histogram for the entire image and for the different sub regions of the image is also computed. The region based histograms helps to simulate locality in color distribution. 3D histogram for 5 local regions is computed and each region is represented by a histogram with 8 x 24 x 3 = 576. Hence the image descriptor for 5 subregion of an image clubbed together has dimension of 2880 x 1.

3.1.4 SIFT + Bag of words

Scale Invariant Feature Transform features are used to detect and describe the local features in an image. Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters. Length of SIFT descriptor is Number of Key points x

128. SIFT descriptors of movie posters are clustered for the number of bags defined, then the bags are trained with clustered feature descriptors to obtain the vocab of visual features. Then, SIFT key points and descriptors are computed for each image and their feature descriptor are matched with the vocabulary to obtain their corresponding bag of words descriptor. This descriptor is used to used to train the classifiers.

3.1.5 Raw Image Pixel

Each poster of original dimension 182 x 268 is resized to 100 x 100. The raw image pixels are fed into a convolution neural net.

3.2 Modeling and Classification

I have used 5 classification methods namely Random Forest, Ada Boost, multi class SVM, Multi Layer Perceptron and convolution neural net. Each classifier is trained on different features described in Section 3.1. Random Forest classifier in general performed better than other classifiers, while the results of other classifiers were comparable. Support Vector Machine and Random forest have given best result with SIFT+BOW features.
Grid Search CV and pipeline is used to improve the performance of the classifier. Grid Search CV allows to construct a grid of all the combinations of parameters, tries each combination, and then reports back the best combination/model. Pipeline packages the transformation step of feature selection with the estimation step of classifier into a coherent work flow. *Note*: All hyper parameters values mentioned below will vary with the features and Dataset used.

- **Random Forests** are ensemble learning methods for classification or regression. They grow many classification trees(decision) and each tree gives a classification. The result is the mode of the classes predicted by individual decision trees.Hyperparameters(max_features: 'log2',n_estimators:100,random_state: 700)
- **AdaBoost**, a type of ensemble learning, is a stage-wise optimization of an exponential loss function. It combines classifier with performance, weak learners, into a strong classifier. It works by choosing the base algorithms and iteratively improving it by increasing the importance of samples that are still miss-classified.Hyperparameters(n_estimators: 50,random_state: 1)
- **Multi Layer Perceptron(MLP)** is a feed forward artificial neural network with one or more layers between input and output layer. Each node except the input node are neurons, simple computational units that have weighted input signals and produce an output signal using an activation function.The output of one layer is fed as an input to another layer in the network. The network is trained using back propagation algorithm.Hyper parameters(activation: 'logistic', alpha: 0.0001, hidden_layer_sizes: (90, 20, 10, learning_rate: 'constant',solver: 'adam')
- **Support Vector Machine**, is a supervised machine learning algorithm used for classification, regression tasks etc which produces a hyperplane separating the two class of data and maximizes the margin from this hyperplane. Hyperparameters(C: 0.001,kernel: 'linear')
- **Convolution Neural Network** is trained on raw pixels of movie posters. Each movie poster is represented as vector $x^{(i)} \epsilon \mathbb{R}^n$, where n is the number if pixels in the image. Each movie at most belongs to 25 Genre. Movie posters are labeled using one hot vector representation. It's a boolean vector $y^i \epsilon \mathbb{R}^{|G|}$, where G is the set of genre and $y_j^{(i)} = 1$ if i belongs to genre j, 0 otherwise. The algorithm will predict a single genre $\hat{y} \epsilon G$ for each movie poster. The prediction is defined as argmax of conditional probability:

$$\hat{y} = \underset{j}{\operatorname{argmax}} P(Y = j | x_{(i)})$$

4 Dataset

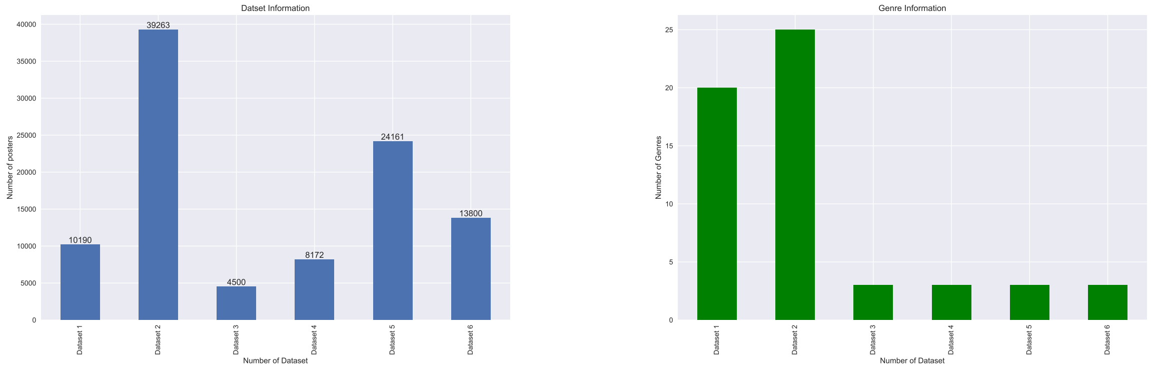
The movie posters are obtained from IMDB. In order to get web-link of movie posters, IMDB_id of 40,000 movie posters is collected from Movie Lens dataset[18]. The collected dataset contains IMDB_Id, IMDB Link, Title, IMDB Score, Genre and link to download movie posters.

4.1 Preprocessing

Entire dataset has 39,263 movie posters, where every poster belong to least one genre label. The size of each movie poster has dimension of 182268.
There are 25 unique genre and 1,293 different combinations of genre present in the dataset. 10,190 movie posters belong to only genre category while 29,103 movie posters have more than one genre label assigned. Each movie poster is assigned at max three genre labels.I have created following subsets of data, which would be used to train classifier:

- **Dataset 1:** Movie posters with single genre.
- **Dataset 2:** Movie posters with one or more genres.
- **Dataset 3** Top 3 genres, where every movie poster has one genre.
- **Dataset 4:** Top 3 genres, each 1500, where every poster has one genre.
- **Dataset 5:** Top 3 genres, where every poster has one or more genre.
- **Dataset 6:** Top 3 genres, each 4600, where every poster has one or more genre.

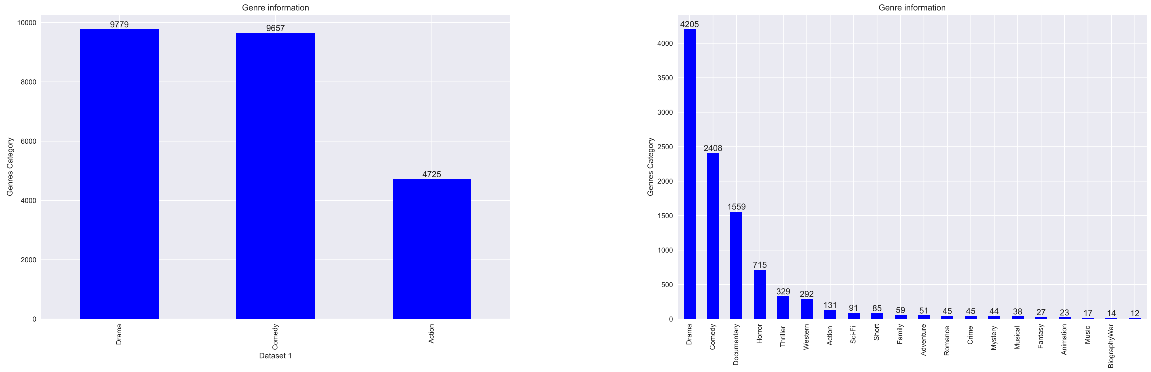
In dataset 2,5 and 6, movie posters belong to more than one genre label, and most prominent label is used for classification. Dataset 4 and 6 have balanced class label i.e each genre category has same number of movie posters. To create dataset with balanced class labels, movie posters for each genre categories were filtered, then relevant genres were merged and shuffled randomly.



(a) Number of posters

(b) Number of Genre

Figure 1: Graphs showing detailed information about the dataset, number of posters in each dataset, number of unique genres in entire dataset



(a) Dataset 5

(b) Dataset 1

Figure 2: Graphs showing detailed information about the dataset,top 3 genre in dataset 5 and number of genre present in dataset 1

5 Experiments and Results

In each dataset, 75% movie posters are used to training while rest are used for testing. The training data was further split into 80-20 ratio to get validation set. All the experiments were run using 3 fold cross validation and recursive feature selection to select the best features. The experiments were conducted on different datasets mentioned in section 4.

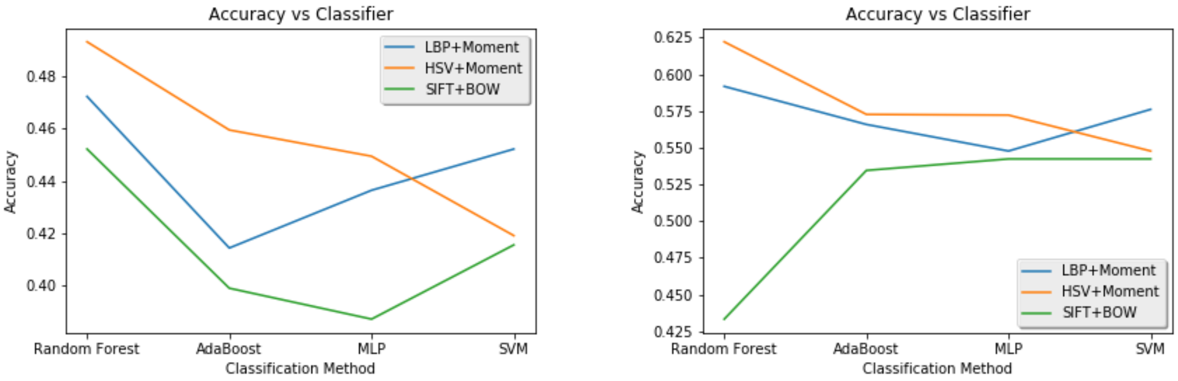
All traditional classifiers are trained using combination of features like LBP+Moment, Moments, HSV+Moments, LBP+HSV+Moments, selecting top 36 values from the histogram of each channel namely Hue, Saturation, Value, 3D color histogram for 5 regions of the image etc. The experiments are conducted on all except dataset 2 for traditional classifiers. It is observed combining moment features with local binary patterns or HSV increased the accuracy by 4-5%. The results for a subset of experiments are reported here.

Accuracy, precision, recall, F1 score was computed for each of the experiments. To understand about misclassification, confusion matrix is computed for all experiments. The diagonal elements show the number of correct classifications for each class,while off-diagonal elements represents the misclassification.

Convolution Neural Net consists of Input layer → Convolution Layer → Max pooling Layer → Convolution Layer → Max pooling Layer→ Dense layer → Dense Layer(Output Layer). The nonlinearity function used in convolution layer is tanh while rectify is used in second last dense layer and softmax function on last output dense layer. Max pooling layer has pool size(2,2) and filter size for Convolution Layer is (3,3). Raw image pixels are used to train convolution neural neural network on each of the 6 dataset. The best result are obtained on Dataset 3,4,5 and 6, where movie posters belong to top 3 genre categories in the dataset. There was a improvement of 10% - 15 % in the accuracies.

Table 1: Best results obtained on Dataset 3 for HSV+ Moment Features

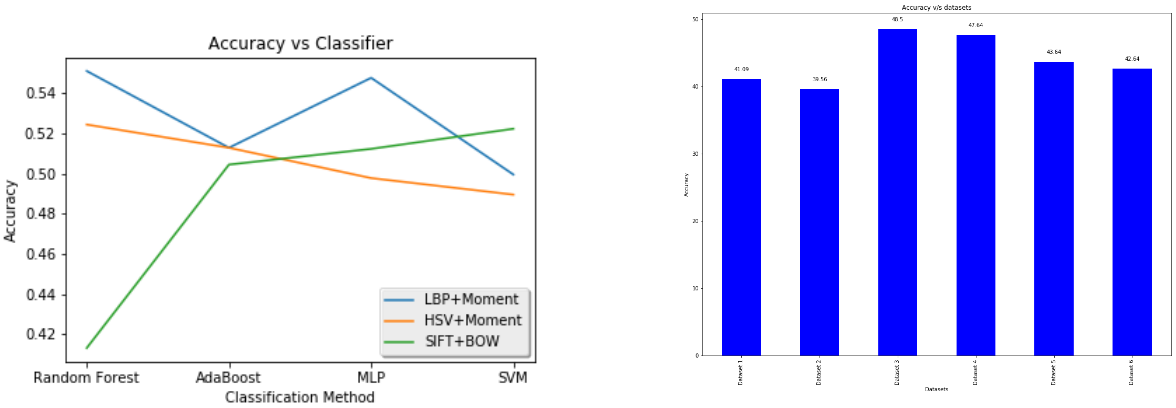
Evaluation Metrics	**Random Forest**	**AdaBoost**	**MLPClassifier**	**SVM Classifier**
Accuracy	0.622	0.583	0.572	0.573
Precision Score	0.595	0.557	0.530	0.532
Recall Score	0.622	0.583	0.572	0.573
F1 Score	0.571	0.542	0.533	0.536



(a) Dataset 1

(b) Dataset 3

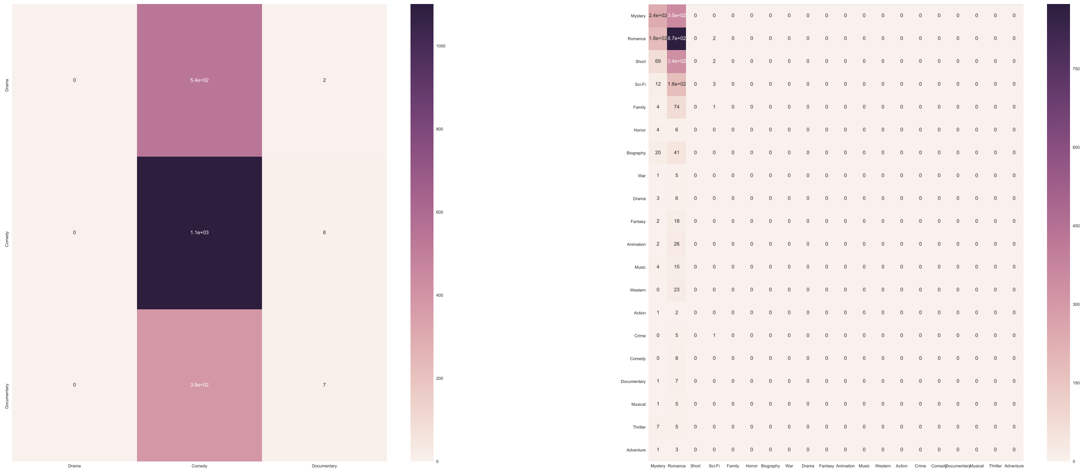
Figure 3: Accuracy Result of 4 classifiers namely Random forest, AdaBoost, MLP SVM on the HSV+Moments, LBP+Moments, SIFT+BOW feature vectors,on different dataset



(a) Dataset 5

(b) CNN on different dataset

Figure 4: Accuracy Result of 4 classifiers namely Random forest, AdaBoost, MLP SVM on the HSV+Moments, LBP+Moments, SIFT+BOW feature vectors,on different dataset. A bar graph displaying the accuracy obtained on convolution neural net trained on raw image pixels for different datasets



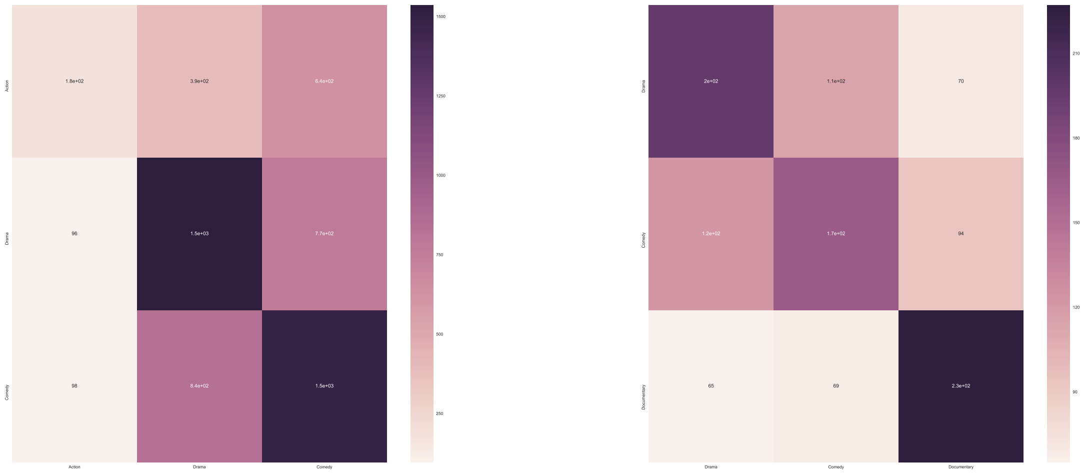
(a) Dataset 3

(b) DataSet 1

Figure 5: Confusion matrix for genre categories. The result shows confusion matrix for best accuracy result on a given dataset

6 Discussion and Conclusions

The HSV features combined with image moments(mean, standard deviation, skewness and Heu Moments) gave best results for top 3 genre (Drama, Comedy, Action). This shows color and image moments do play an important role in the genre classification. The texture of movie poster is equally important,as the results of classifiers on local binary patterns plus image moments were comparable. Initially the classifiers were trained for 20 genre categories on dataset 1, this dataset has class imbalance where top 3 genre comprise 80% of the dataset. Hence when classifiers were trained on top 3 genre, accuracy increased by approx 10%. The same observation was made for convolution neural



(a) Dataset 5

(b) Dataset 4

Figure 6: Confusion matrix for genre categories. The result shows confusion matrix for best accuracy result on a given dataset

network. The image moments, feature vectors are also able to classify genre correctly with accuracy of approx 50% for Dataset 3. Therefore, these features also represent important characteristics of image related to genre classification. As a movie could belong to more that one genre it would of great interest to extend this work to a multi label classification task where classifier should be able to predict the multiple genres associated with the given movie.

References

[1] Valdez P, Mehrabian A., *Effects of color on emotions. Journal of experimental psychology.* Genera, (1994)

[2] Zhou, H., Hermans, T., Karandikar, A.V., Rehg, J.M., *Movie genre classification via scene categorization*, pp. 747750 (2010)

[3] Huang, H.-Y., Shih, W.-S., Hsu, W.-H.: *A film classifier based on low-level visual features*, pp. 465468 (2007)

[4] John Arevalo, Thamar Solorio, Manuel Montes-y-Gomez, Fabio A. Gonzalez : *Gated Multimodal Units for Information Fusion*,arXiv:1702.01992, (2007)

[5] Ivasic-Kos, M., Pobar, M., Mikec, L.: *Movie Posters Classification into Genres Based on Low-level Features.*, pp. 11981203 (2014)

[6] Paris, G., Lambert, P., Beauchene, D., Deloule, F., Ionescu, B.: *Animated Movie genre detection using symbolic fusion of text and image descriptors*, pp. 3742 (2012)

[7] Fu Z., Li B., Li J., Wei S. *Fast Film Genres Classification Combining Poster and Synopsis*. In: He X. et al, (2015)

[8] Jitendra Malik, Serge Belongie, Thomas Leung, Jianbo Shi(2001), *Contour and Texture Analysis for Image Segmentation*, Kluwer Academic Publishers. Manufactured in The Netherlands.

[9] Egon l. VAN DEN broek, Peter M.F. Kisters, Louis G. Vuurpijl (2004), The utilization of human color categorization for content-based image retrieval, Proceedings of SPIE, 2004

[10] Z. Rasheed, Y. Sheikh, and M. Shah, *On the use of computable features for film classification*, pp. 5264, 2005. [10] Movielens Dataset: https://grouplens.org/datasets/movielens/latest/

[11] IMDB : http://www.imdb.com/

[12] Bag of words model: https://en.wikipedia.org/wiki/Bagofwords_model_in_computer_vision

[13] Color histogram: https://en.wikipedia.org/wiki/Color_histogram

[14] T. Gevers and A. Smeulders, *PicToSeek: Combining color and shape invariant features for image retrieval*, IEEE Transactions on Image Processing, vol. 9, pp. 102119, 2000

[15] Navneet Dalal , Bill Triggs, *Histograms of Oriented Gradients for Human Detection*, p.886-893, 2005

[16] T. Ojala, M. Pietikainen, and T. Maenpaa. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):971987, 2002.

[17] Sural, Shamik, Gang Qian and Sakti Pramanik *Segmentation And Histogram Generation Using The Hsv Color Space For Image Retrieval*, IEEE ICIP 2002

[18]Movie Lens dataset *https://grouplens.org/datasets/movielens/latest/*

[19]Multilayer Perceptron: *https://en.wikipedia.org/wiki/Multilayer_perceptron*

[20] Theano Library *http://deeplearning.net/software/theano/#*