Social Networking Meta Data Based Image Classification

A Thesis submitted

In Partial Fulfilment of the Requirements

for the Degree of

BT-MT Dual Degree

by

Gaurav Krishna

under the guidance of

Dr. Harish Karnick

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur

May 2015

Certificate

It is certified that the work contained in this thesis entitled "**Social Networking Meta Data Based Image Classification**", by **Mr. Gaurav Krishna (Roll No. Y9227224)**, has been carried out under my supervision and this work has not been submitted elsewhere for a degree.

Dr. Harish Karnick

Professor,

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur

Kanpur, 208016.

Abstract

Every day, millions of people take photos and upload them to social media websites. Their goal is to share photos with people, but collectively they also create a vast repository of visual information.

When they engage with this visual information, in the form of social connection like *comments, likes, tags* etc. They also leave a semantic footprints about these visuals. This thesis finds methods to harness this semantic information to achieve better classification of this visual information. Improvement in classification accuracy leads to better organizing of data, better information retrieval, better recommendation.

Large Scale Image Retrieval Benchmarks are generally consists of photos from web. These photos also have rich social meta data attached with them. Therefore, these photos augmented with their social networking data provides a good data-set for evaluating our methods. We have introduced a novel method of constructing features from social networking data and fuse them with automated concept searching methods like Latent Semantic Indexing to give valuable classification features. We compared the classification on these social features and image content based features using structured classification techniques. The findings suggested that social network meta data are very useful in image classification tasks and many times outperforms image content based methods. We further proposed the ensembling of these two different modalities of features to boost the classification results.

Acknowledgements

I would like to express my sincerest gratitude to my thesis advisor, Dr. Harish Karnick, for his invaluable guidance and constant support during the past couple of years. His kind and encouraging nature has been very helpful in keeping me motivated for research. I am grateful for his patient guidance and advice in giving a proper direction to my efforts. I am very grateful for his help and guidance during all this time.

I would also like to thank the Department of Computer Science and Engineering at Indian Institute of Technology Kanpur for the excellent environment for research, all the facilities and the outstanding academic training they provided during my stay here.I am indebted to all my friends and wingies for making these past five years a memorable one for me. I especially thank Siddharth, Nirmal, Kanish and Urvesh for valuable inputs in my thesis.

Last, but not the least, I would like to thank my parents and siblings for their love, constant support and encouragement. Without their support and patience this work would not have been possible.

Gaurav Krishna

Contents

List of Figures

List of Tables

Chapter 1

Introduction

It has become a cliche to start a discussion on the field related to classification or tagging the online multimedia information by stating just how the meteoric growth of data has been in recent decade. While it may be a very trivial and obsolete way of introducing about the online digital corpora, but it is, nonetheless, true for work in this field.

The proliferation of personalized gadgets, cheap digital cameras and any such diversification away from single use devices like mobile phones and into everything from computer web-cams to tablets, it has now become a culture of saving the every moment of our lives in digitized format. The internet services like Facebook, Flickr, Instagram etc also made it very easy for common person to share their pictures with their (online) social connections.

In present scenario, where we have proliferating robust systems effectively handling the billions of photos, images and videos; the renaissance of the 'socialization' of web activity has produced a massive amount of social interaction information. The social interaction on multimedia supporting websites have also opened a new avenue of managing this ever growing corpora of multimedia data. The point is that this social interaction data like tags, comments, likes, groups, galleries, playlists etc. all leaves some clues about the content in the question.

This social meta-data, the data related to the content, can be a pathway to manage the whole multimedia burst in proper classifications, themes or may be ontological perspective. This meta-data can help us in many ways for example information retrieval, multimedia classification, heterogeneous learning etc. In our thesis, we are going to focus upon the images and their social meta data.

Photos are unlike the text documents have a far more immediate emotive impact - reportage from people dying in Somalia because of poverty to striking wild life photography can make an impression on viewers very quickly. In past, photographs were normally confined to an album and were rarely opened to relish the old memories. The photos were difficult to share and normally do not contain any contextual information. Where as in present scenarios, people can upload images to world wide web direct from the capture devices (like phone, tablets). They can also encode pertinent contextual information automatically and share it with all their social network contacts. This starts a cycle of interaction and annotation of that image. Over the time, the social aspect of image becomes more prevalent. This creates a need of expanding our understanding and power of utilizing this social data.

This thesis have focused on exploring the role; which can be played by social context information in enhancing the image classification. In image classification, an image is classified according to its visual content. For example does it contain clouds or not. We analyze some of the visual characteristics based features and then, we try to fuse these visual features with the social context information to follow a multimodal approach of classification.

1.1 Motivation

Image classification has many application ranging from multimedia information delivery to web search. In the past, image classification has faced two major difficulties. First, the labeled images for training are generally hard to extract out and short in supply plus labeling new images costs human labor. Secondary images can be ambiguous; e.g. an

image can have multiple concepts hidden. To overcome these problems, we can leverage the other information about the images. Even when labeled data is hard to create, we can take advantage of the text data related to images or the social interaction data associated with images [2].‘

Apart from these problems, we also experience the problem of ambiguity of key objects in image classification. The low level features of visual information can often be ambiguous. An object can belong to more than one concepts (polysemy). In such cases, we need some help of extra data to handle such ambiguous conditions gracefully in order to get high classification accuracy.

Text and images are two distinct type of information resources and belong to two different modalities, as they present an öbjectïn different ways. However, there are some implicit connections and invariant properties, which are shared between image and visual information (Smeulders et al. 2000). In fact, the texts associated with images includes some form of human generated descriptions of images; thus these can be regarded as supplement to the image content.

The combination of textual information with image features has been proposed to improve the image classification results(Swain et al. 1997; Zhao & Grosky 2002; Hu & Bagga 2004; Song et al. 2004; Smith & Chang 1997).

With all these cues about heterogeneous learning, we get ample motivation of taking the help of data in different modalities to have better classification of images. The unprecedented evolution of social networking gives us such data with very ease. People talk about images, puts likes, comments, tags and other different markers. These markers or rather, we say social meta-data can be instrumental in inferring the content of images.

Krystyna K. Matusiak has shown that how the social tagging can be a tool for enhancing the description of digital objects[3]. In some recent studies, it has been shown that the social-metadata can be utilize in case for better classification. (Yuan Lin and Yuqiang Chen et al [2], Lua, and Wei-Ying [4],Julian McAuley and Jure Leskovec [5])

Chapter 2

Background and Related Work

With the advancement of the machine learning algorithms, there have also been increase in the attempts of trying various innovative attempts of classifying images. There have been attempts to develop low level image features to have faster classification. Also people tried to mimic the human visual perception process in image classification. Most of the image classification solutions counter with an unanimous problem the deficiency of sufficient data. It is really difficult to have image data for all classes. Even if you get data, you have to deal with problems of noisy tags, ambiguous content etc. On the other side, we have lot of image data getting created on world wide web. This leads to a simple question that can not we use this data get some semantic structure, which would help us in image classification. The answer is yes, but the extent to which we can structure the world wide web and extract out the useful information still has lot of research potential.

The important concept, we want to coin here is heterogeneous learning. Kesorn, Kraisak [6], as shown that although, text and visual are distinct types of representation and modality. There are some strong implicit connections between images and any accompanying text information. Semantic analysis of image captions can be used by image retrieval systems to retrieve selected images more precisely. This is a form of heterogeneous learning. They had first extracted meta data automatically from text captions and restructured with respect to a semantic model. Second, they used Latent Semantic

Indexing to create a domain-specific ontology-based knowledge model. This enables the framework to tolerate ambiguities and variations (incompleteness) of meta data.

In [7], Kim, Kristen Grauman and Raymond Mooney solved the problem of learning robust models out of scenes and actions from partially labeled collections. They proposed to leverage the text accompanying visual data to cope with these constraints. They concluded that exploiting the multi-modal representation and unlabeled data provides more accurate image and video classifier compared to base-line algorithms. They have also asserted that this extra data and multi-modal representation can be formulated as a basis of a solution to the problem of managing the world's ever growing multi-media data of digital images and videos.

Julian McAuley and Jure Leskovec [5] has proposed that how we can use the interdependencies of image sharing common properties in multi-modal classification settings for image labeling in social networks. They used the same Large Scale Image Retrieval benchmarks (MIR, PASCAL, CLEF and NUS), which we are using in the thesis. They modeled their task as binary labeling problem on a network and used some machine learning technique to create a binary labeling model. They have analyzed about the relative importance of social features (eg. shared membership in a gallery, relational feature on the based on shared location, shared group etc.) in image labeling.

Stvilia, Besiki and Jörgensen, Corinne and Wu, Shuheng [8] came up with a framework for evaluating the added value of socially created meta-data to image indexing process. They were trying to define content models understandable and search-able for humans. Image indexing becomes a complex socio-cognitive task because it includes processing available data, then classifying it, abstracting and mapping that data into semantic concepts and entities often expressed through, which should also be justified on the basis of human linguistics. They have shown that how the meta-data or tags obtained by social network helps in understanding the higher level concepts and semantic relationships hidden in images, which would otherwise be not achievable.

It has been asserted that collaborative tagging, social classification, social tagging, which is also said "Folksonomy" [9], holds the key to develop a semantic web, in which

every web page contains machine readable meta-data that describes its content. Such meta-data can be harnessed to describe the content and classify the content. In [**?**], it has shown the impact of the meta-data for retrieving information from web-site.

Oded Nov and Mor Naaman [10] and Kern, R. and Granitzer, M. and Pammer, V. [11], have experimented on Folksonomy. They used some collaboratively created sets of meta-data, to organize multimedia information available on the Web. They addressed question how to extend a classical folksonomy with additional metadata. They also shown that it can be applied for tag recommendation.

Ed H. Chi and Todd Mytkowicz [12] has systematically analyzed the efficiency of social tagging systems using information theory. They try to find the answer about the efficacy of naturally evolved user generated vocubalary in identifying the objects (images, videos etc.).They have shown that information theory improvises an interesting way to understand the descriptive power of tags. Their results show that even information theory gives evidence that social tags can be used to identify the objects.

Liu, Dengsheng Zhang, Lua, and Wei-Ying [4] has demonstrated that In order to improve the retrieval accuracy of content-based image retrieval systems, we can fuse the data from HTML text and visual content of images for World-Wide-Web retrieval.

Yuan Lin and Yuqiang Chen et al [2] has shown that how labeled text form web help image classification. In this paper, they have investigated the interplay between multimedia data mining and text data mining. They addressed the problem of image classification with limited amount of labeled images and large amount of auxiliary labeled text data. They have considered the bag of Words model and Navie Bayesian Classification models. They have estimated the image feature mapping on the text-image co-occurrence data, acting like a bridge, which connects the text knowledge to image.

In [13], Nuo Zhang and Toshinori Watanabe has proposed a text-transformed image classification. They have proposed to first transform the images into texts. Then these text-transformed images are divided into segments and replaced by characters. Then, they use a novel method of using the similarity between compressibility vectors of texts is the basis of classification.

In 2010, Slaney and Mahajan [14] combined the information from the social graph with some semi-supervised techniques from all the unclassified images to create an enhanced image-classification model for multimedia data. They have shown that fusing image, text and social-graph features gives a large improvement over content features alone in an experiment, where they tried to classify the imahes of adults among all the images.

Leandro et al. [15]has shown a neural approach based on swarms to combine visual feature and text data for classifying the images.They has shown it as a two step process, one in which they query from text related to the image and in next they query for image characteristics to fulfill the objective of representing multimedia data and to extract knowledge from this kind of data.

In IEEE Conference on Computer Vision & Pattern Recognition 2010, Matthieu Guillaumin, Jakob Verbeek and Cordelia Schmid has considered a scenario where keywords are associated with the training images, eg as found on photo sharing websites. They demonstrated a semi-supervised multi-modal learning for image classification. They have shown that how the other source of information can aid the learning process when we have limited number of labeled images.They used PASCAL 2007 dataset and tried learning classed from Flickr Tags.

In [16] Honglak and Packer et al. has introduced a concept of self-taught learning. In this, they have presented a new machine learning framework. In this framework, they used unlabeled data used in supervised form to do classification tasks. They used a large number of unlabeled images, audio samples or text documents downloaded from the internet to improve performance on a given image classification task.

Adam Rae and Zwol van [17] addressed the task of recommending the additional tags to partially annotated images. For this they proposed to used the various collective contexts. They used a large-set of real world data from Flickr to show that collective social knowledge has capability of significantly improve the tag recommendation.

Zwol van [18] handled the problem of predicting of users' favorite photos in Flickr. They used a multi-modal machine learning approach, which fuses social, visual and textual signals into a single prediction system. They proposed that the visual, textual and

social modalities effectively infer the needs of most users. They used gradient-boosted decision trees (GBDT) for the classification of a user's favorite photos. For the evaluation of the performance of their classifier, they classified the data with respect to the individual modalities and various combinations. By using heterogeneous modalities, the GBDT becomes a highly effective classifier. The addition of textual features helps to significantly boost the recall, with a bit decrease in precision.

Jiayu Tang and Paul H. Lewis [19] proposed an visual image based feature space; mapping image regions and textual labels into one single space. They proposed that such representation makes the object recognition very straightforward. In this space, similar image segments linked with the same objects are clustered together, and also lie in the neighborhood of the labels linked to these objects.

In CVPR Paper[20] of 2004, Matthew Boutell and Jiebo Luo shown that how we can leverage the camera meta data to provide evidences independent of the captures scene content; to improve the classification performance. They used Bayesian network to fuse content based and metadata features. Their results demonstrates that this integration of camera meta data can only increases the efficacy of classification.

In 1994, Bartell and Cottrell [21] has shown that we can have combine multiple retrieval systems to get a superior retrieval model. They demonstrated that if we use multi-modal query to retrieve a single informational entity, we get more effective superior retrieval efficacy.

In [22], Teredesai and Bindra has shown the variety of methods for scalable topic modeling in social networks. They have talked about using Latent Dirichlet allocation (LDA), Latent Semantic Indexing (LSI) etc. unsupervised topic modeling techniques to harness social linkages to decipher the user interests for target recommendation.

Chapter 3

Data Set

The social network meta data related to images can be made the basis image classification tasks and many times can outperform image content based methods [23]. For the validation of this hypothesis, we needed a well labeled image data, which should also have a sufficient possibility of expanding it in social dimension. The data must have some ground truth provided by human annotators or by some standard benchmarks.

Most of the Large Scale Image Benchmarks are usually fabricated by the use of the vast sphere of images available on web. These images are majorly the part of social media holders like Pin-interest, Flicker, Facebook. After exploring the available image benchmarks, we narrowed our focus on the following four well established benchmarks, because these were created of flicker images and by implementing the flicker APIs on available information, we could extract enough social meta data about the images:

- The PASCAL Visual Object Challenge ('PASCAL') [24]
- The MIR Flickr Retrieval Evaluation ('MIR')[25]
- The ImageCLEF Annotation Task ('CLEF')[26]
- The NUS Web Image Database ('NUS')[27]

The creators of these data-set had obtained labels through crowd-sourcing, the ficker users or communities. The labels range from object based categories like *person* or *bicycle*, to subjective concepts like *Aesthetic_ Impression*. These labels satisfies the desired ground truth constraint for our classification process. We, therefore, use these labels as a classification base for our analysis.

3.1 Description of Data sets

The PASCAL Visual Object Challenge ('PASCAL') consists of over 12,000 images collected since 2007, with additional images added each year [24]. Flickr sources were available only for training images, and for the test images from 2007. There were total of 11,197 images, for which flicker sources were available.

The MIR Flickr Retrieval Evaluation ('MIR') [25] consists of one million images, 25,000 of which have been annotated. Flickr sources were available for 15,203 of the annotated images.

The ImageCLEF Annotation Task ('CLEF') [26] uses a subset of 18,000 images from the MIR data set, but it has more varied tagging annotation. There were total of 4,807 images, for which flicker sources were available.

The NUS Web Image Database ('NUS') [27] consists of approximately 270,000 Images. Flickr sources are available for all images.

3.2 Augmentation of Social Meta-data

Flicker Sources of above data sets were provided by the data set creators. We used the possible flicker APIs and tried to obtain the maximum meta data for each photo instance. The information, we could extract out were

- The photo itself

- Photo data,
  - Title
  - Description
  - Location
  - Time stamp
  - View count
  - Upload date
- User information, including the uploader's name, username, location, their network of contacts, etc.
- Photo tags, and the user who provided each tag
- Groups to which the image was submitted
- Collections (or sets) in which the photo was included
- Galleries in which the photo was included
- Comment threads for each image instance

We considered only the images which have all the above data available, which is roughly 90% of the images for which the URL Source was available.

3.3 Preliminary Observation of Data set

We also did some observation while extracting data. We would like to give those observations at this stage because those will help us describe the inferences and results.

FIGURE 3.1: Image MIR Examples

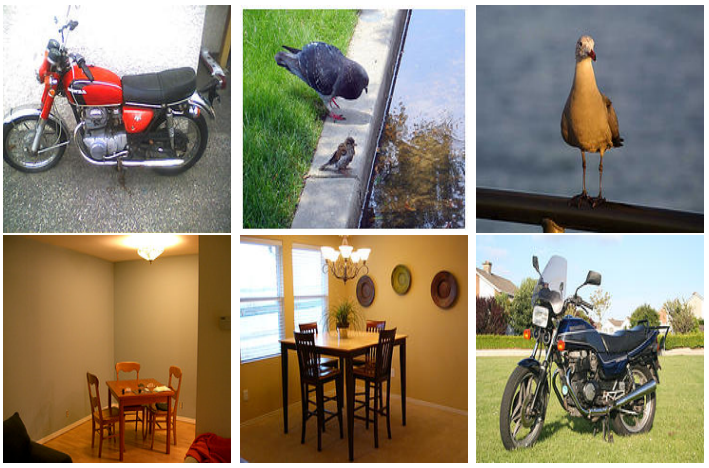

FIGURE 3.2: Image PASCAL Examples

The statistics obtained from enriching the images and elementary statistical analysis of image properties reveals that there is a large difference between the data-sets, for example PASCAL has least tags and comments compare to all other data-sets, because it is made up of less interesting images. The NUS Data Set favors the highly popular images as we can see that it has the highest tag vs image ratio of 19.4. Th images are highly tagged,

FIGURE 3.3: Image CLEF Examples

FIGURE 3.4: Image NUS Examples

have large number of comments and are submitted to many group. The MIR has 17+ tags comments showing that it contains 'interesting images' [25]

3.4 Label Selection for Classification Problem

Due to constraint of presence of less number of images for some particular labels and balanced learning, we have to selectively choose the labels for our Classification Problem. For eq. CLEF has 99 labels and some labels have image instances of less than 17 images, so doing a learning and testing on such small set is out of question. We, therefore, in case of CLEF select 20 Labels which have sufficient data. Similarly for NUS, because of some computational constraints and data availability, we reduce our computation for 12 labels. Following table shows the labels, which are considered in the whole classification problem.

On the line of [23] we decided to construct the problem in binary classification form.

TABLE 3.1: Labels

DataSet	Number of Labels Selected	Labels
NUS	10	animal, coral, dancing, harbor, military, mountain, snow, statue, tattoo, temple, waterfall, wedding
PASCAL	20	aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, diningtable, dog, horse, motorbike, person, pottedplant, sheep, sofa, train, tvmonitor
MIR	14	flower, car, bird, dog, night, tree, clouds, portrait, female, male, people, sea, river, baby
ImageCLEF	20	Adult, Aesthetic_ Impression, Animals, Autumn, Citylife, cute, Day, Flowers, Food, Graffiti, Landscape_ Nature, Painting, Portrait, Single_ Person, Sky, Street, Summer, Sunset_ Sunrise, Vehicle, Winter

This gave a leverage of easily use the bag of visual/ non-visual words, because when we try to indicate the features of an instances using some bagging, we can not cover the high number of labels in those limited bags and visual/non-visual dictionary. This gives us better learning of features for each label and also gives precise information retrieval.

Chapter 4

Feature Extraction

Feature extraction is a special form of dimensionality reduction, without loosing important information of data set. When we talk about images or text data generally actual data set is too huge to be processed. We need to use some transformation to extract out some interesting points from the data. Such points should help us in achieving our goal, which can vary from image classification to topic finding anything. Such reduced set of instrumental information is called as set of features. The process of computing these features is called as Feature Vector Extraction. These feature vectors are constructed in a way that they perform the task using reduced representation. For the feature extraction from our image data-set augmented with social meta-data, we break the process in two parts:

- Image Content Based Feature Extraction
- Social Content Based Feature Extraction

In the following part of chapter, we will explain about the features we have used and methodology used for extraction of those features.

4.1 Image Content Based Feature Extraction

Feature extraction is an important step in image processing. The performance of a classifier depends on the feature vector. Several kind of features are proposed in image processing field. Some primarily used common features for image classification are Color Histogram, HoG, LBP, SIFT, SURF, GLCM etc. Even after extracting important feature vectors, we face the problem of high dimensionality. So for reducing the dimesnionality, we use several other dimesionality reduction techniques. Some important techniques are PCA, Bag-of-words etc. In our work, we used following visual features:

- SIFT Features
- GIST Features
- COLOR Space Features
- Texture/GLCM Features
- HOG-LBP Features

In case of HoG-LBP and SIFT features we also used PCA and Bag-of-words model for further dimesnionality reduction. In the following part of this section, we will explain about these features and methodology used for extraction.

4.1.1 SIFT Features

SIFT(Scale-Invariant Feature Transform), as the name suggest, it is a feature descriptor, which is invariant to image scaling. But It is just not only consistent with the variation in scaling, it is also consistent with translation, rotation and till some extent also remains unaffected of (some) variations of the illuminations , 3D projections and other viewing conditions. The SIFT is normally bundled with a feature detector and a feature descriptor. The detector extracts attributed regions from in such a way, that the description of these

FIGURE 4.1: Example of SIFT Descriptors

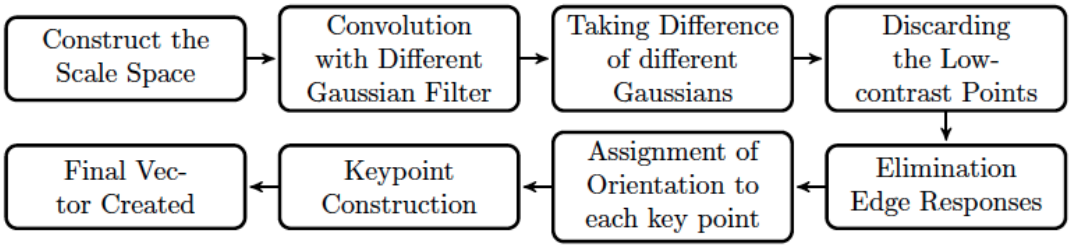

FIGURE 4.2: Process Flow of SIFT Descriptor

regions(descriptors) is consistent with all the aforementioned changes (illumination, view point etc). The descriptor associates with the regions a signature which recognize their appearances efficiently and accurately. For example, some sift descriptors are shown in Figure 4.1.

By above-mentioned description, we can easily judge that SIFT features are very instrumental for finding objects and recognizing scenes in an image. SIFT descriptors were first introduced by Lowe [28] in ICCV 1999. He took the idea from the primate vision process. The SIFT Features are actually similar to the neurons in inferior temporal cortex of a primate. Features are efficiently extracted through a staged filtering approach that focuses on some key invariable points in scale space. The steps of this filtering approach as mentioned below in Figure 4.2.

SIFT features have been proven very useful in objectives like: natural scene recognition by Li Fei-Fei and Pietro Perona in their paper [29]. It is also very instrumental in object recognition as shown in [28]. Motivated by the excellent performance of SIFT Features in Image Categorization, we decide SIFT as one of our features for image classification.

In [29], Fei fei et al. has shown that dense local scale-variant features performs better compared to sparse features. We. therefore, extracted dense SIFT features of 16×16 pixels frames. These frames were created over a grid with spacing of 8 pixels. A dense image descriptor has better chance to associate itself to overall scene recognition as it can capture uniformity of image such as landscapes, sea, sky etc.

We used VLFEAT Library for finding the SIFT descriptors. Once we have obtained the SIFT descriptors for the full set of images, we construct a visual vocabulary of 400 visual words as described in [30]. Lazebnik et al. [30] introduced the concept of visual bag of words, because this methodology overcomes the problem of high-dimensionality of SIFT features. We can constraint the definition of a visual concept in 400 visual words, which provide efficiency and robustness. We used k-means from VLFEAT library [31] for creating 400 clusters out of full training data. In figure 4.3, we have given an example of using SIFT Bag of Words features to match two natural scenes.

Lazebnik et al. [30] gave an extension to normal BoW (Bag of Words model) by introducing Spatial Pyramid technique. In this technique, we create a spatial pyramid of an image by dividing the image in hierarchical spatial layers. Each division provides a finer sub-region and more spatial localized information about the image. We keep on dividing the image in layers and then calculate histogram over those 400 Visual words on these divisions.This approach works gives improved results on challenging scene recognition tasks as shown by Lazebnik et al. [30]. So considering this work, we also added Spatial Pyramid technique as defined in [30] . We did 2-level Spatial partitioning. In level 0, we have the full image,

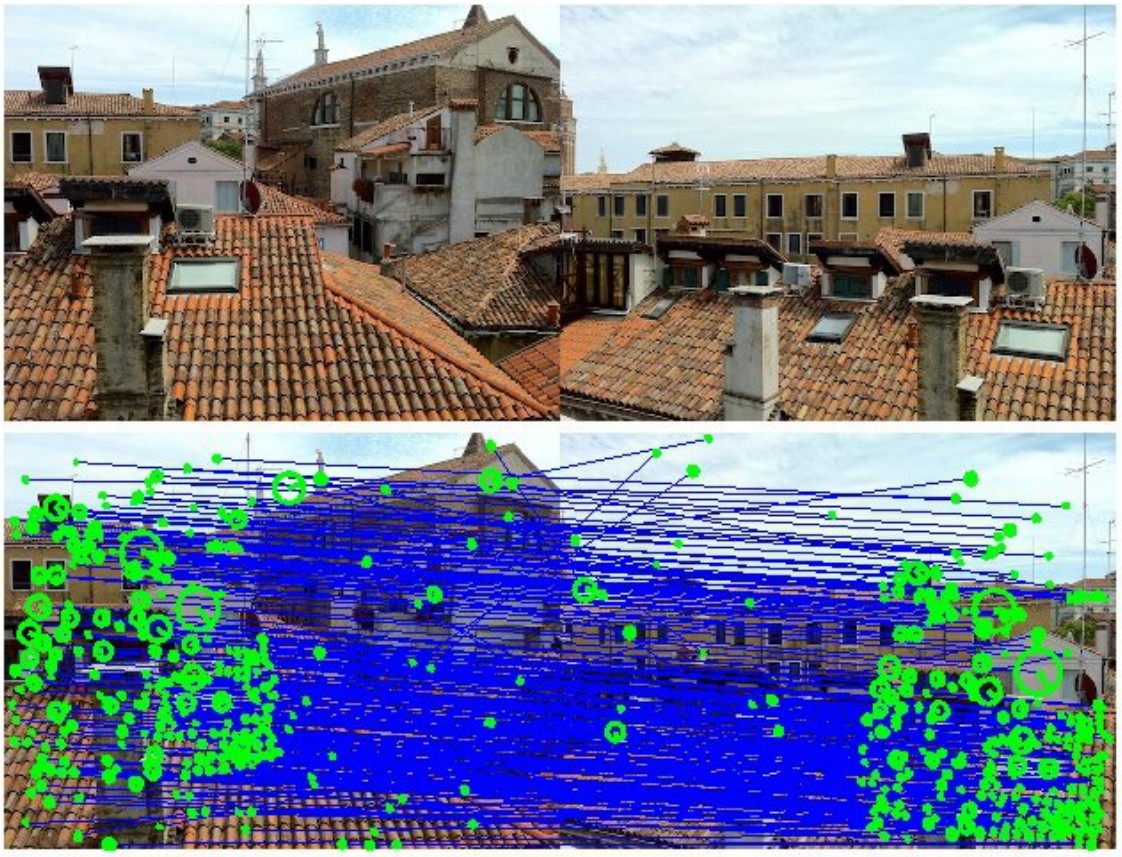

FIGURE 4.3: Use of SIFT Descriptor in matching

this gives 400 dimensional vector (the size of bag of visual words). In level 1, we partition the image in 4 sub regions, this gives 4×400 dimensional vector. In level 2, we have again partion each sub-region into 4 sub-regions, so we have $4 \times 4 = 16$ sub-regions. This gives us a 16×400 dimensional descriptor. In this way, we have an image descriptor of size 8400. We used the MATLAB implementation developed by [30] for computing the pyramid on the SIFT descriptors. A schematic diagram of an image is shown in Figure 4.4

In figure 4.4, we have represented the construction of the spatial pyramid. In this case, we have dictionary of three visual words represented by plus, circles and rhombus. The image is partitioned in 3 layers and for each layer in each sub-division, the count of each visual words is used to create the bin and then the spatial pyramid bins are weighted as given in [30].

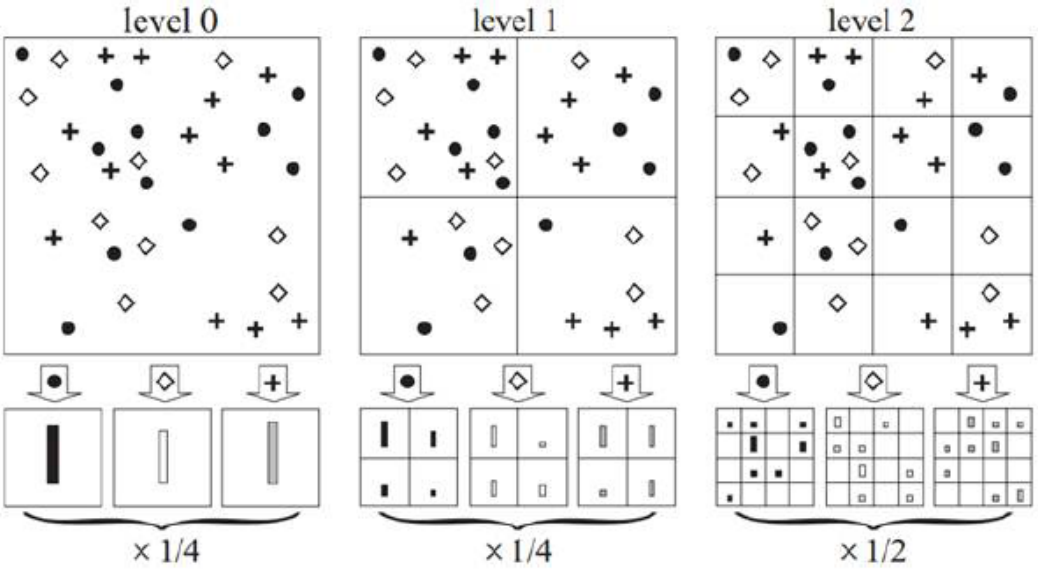

FIGURE 4.4: Computation of Spatial Pyramid over an image

4.1.2 GIST Features

Just like SIFT features, GIST features also have their roots in the concept of primate vision. GIST features were first introduced by Aude Oliva and Antonio Torralba [32] in 2001. They took the reference of Barrow, H.G. and Tannenbaum, J.M. 1978. paper [33] which describes seminal vision in humans. Olivia and Torralba [32] portrayed this seminal conception in computational vision. In case of scene recognition, human mind actually do progressive reconstruction of the input of local descriptors (edges, surfaces) integrated into complex decision layers. Therefore, the recognition of read word pictures may be initiated from some basic global descriptors, ignoring most of the details and object data.

In [32], Olivia and Torralba suggested that recognition of real world pictures can be attempted with some small set of perceptual dimensions : ruggedness, naturalness,roughness, expansion, openness. This small set of perceptual dimension can be used as a way of recognizing a picture without going into tiresome process of segmentation and processing individual regions/objects. This low-dimesnional representation is termed as "Spatial Envelope" in [32]. These Spatial Envelope Perceptual Dimension Descriptors can reliably computed using spectral and coarsely localized information.

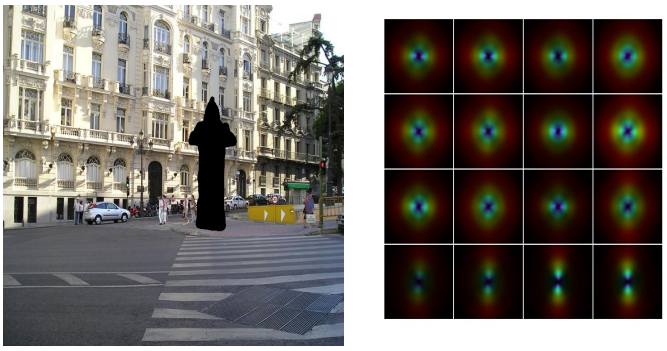

FIGURE 4.5: Example of GIST descriptors

The model based on this spatial envelope generates a multidimensional space. In this projected space, scenes with semantically closed categories (e.g. sea, water, river, lake) are projected closely. The performance of this model emphasizes that for scene categorization and modeling a holistic representation of a scene, we do not need specific information about object shape or identity. This holistic representation is define as GIST of the scene. Douze et al. [34] has shown that the GIST descriptor is very efficient and instrumental for web-scale search system for images. This indicates that GIST can also be an efficient feature for image classification. We, therefore, include GIST in our feature list.

We use the GIST implementation available on [**?**] as shown given by Olivia and Torralba [32]. We first decompose the image using filters of 8 orientations for each of the 4 scales mentioned in [32]. This way we 32 oriented filter. Then the image is represented as 4×4 matrix. Output values of all filters are also normalized to 4×4 matrix. THen the image is represented by the weighted combination of all these values representing into 8(orientation)$\times$ 4(scales) $\times$ 4 $\times$ 4 (size of matrix) = 512 dimensional vector.

GIST can be easily extended for larger database because it is memory efficient and also computationally efficient. Even after being computationally efficient, It turns out to categorize natural images very well.

In Figure 4.5, we have shown GIST descriptors of an image. The left part is an image from our database and right is GIST descriptor for the image.

4.1.3 COLOR Space Features

Colors for each pixel in an image can be represented using tuples of numbers, numbers can be three as in RGB model or four as in CMYK model). A color space is the way of such representation of colors, it is also called as color model or color system. Every color can be represented by a point in this color space. There are multiple color spaces, which are used to represent a color according to the application. Some of them are RGB , CMYK , HSV, CIELAB. In this section, we will give a brief overview of RGB and CIELAB, because we are using these for our thesis.

RGB Color Space

RGB color space is consisted of three components Red, Green and Blue. Red, Green and Blue are considered as additive primary colors because these colors can be used to create a broad range of colors.
Colors can be created on computer monitors with color spaces based on the RGB color model, using the additive primary colors (red, green, and blue). In every pixel , we define the color using intensity values for each of these three colors. Yhe range of intensity value is 0-255. This leads to 16,777,216 different colors, when use different combinations of each of these colors. It is a reproduction medium dependent color space because it depicts different RGB values for the same image when computed on different devices, such as the phosphor (CRT monitor) or backlight (LCD monitor). RGB color space is been used in most of the modern display devices like Television, Computers, Mobile Phone displays etc.

CIELAB Color Space

CIELAB Color Space describers all colors which are perceptibly visible for human beings. It was first introduced by the International Commission on Illumination in 2003. [35]. It is also a three dimensional color space with three components of CIELAB are $L*$,$a*$ and

$b*$. $L*$ represents the lightness of color ranging from 0 to 100 in which 0 is black and 100 is white. $a*$ and $b*$ are color spaces. The range of both of these are -128 to +127. In this $a*$(+127) represent red color where as $a*$(-128) represents green color. $b*$(+127) represent yellow color where as $b*$(-128) represents blue color.

To convert an image from CIELAB to RGB, we first convert the RGB image in CIEXYZ color space. Then we convert CIEXYZ to CIELAB. CIEXYZ is a color space introduced by CIE in 1931. For doing this we used MATLAB functions (makecform, applycform). Following are formula's for converting RGB color space to CIELAB. First conversion is from RGB to XYZ anf then we convert this into CIELAB. Conversion formula is mentioned below: Conversion from CIEXYZ space to CIELAB space:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} 0.412453 & 0.357580 & 0.180423 \\ 0.212671 & 0.715160 & 0.072169 \\ 0.019334 & 0.119193 & 0.950227 \end{pmatrix} \begin{pmatrix} R \\ B \\ G \end{pmatrix}$$

$$L^* = \begin{cases} 116 \times (\frac{Y}{Y_n})^{(\frac{1}{3})} & \text{if } (\frac{Y}{Y_n}) > 0.008856 \\ 903.3 \times (\frac{Y}{Y_n}) & \text{otherwiise} \end{cases}$$

$$a^* = 500 \times (f(\frac{X}{X_n}) - f(\frac{Y}{Y_n}))$$

$$b^* = 200 \times (f(\frac{Y}{Y_n}) - f(\frac{Z}{Z_n}))$$

where,

$$f(x) = \begin{cases} x^{\frac{1}{3}} & \text{if } (x > 0.008856) \\ 7.787 \times t + \frac{16}{116} & \text{,otherwise.} \end{cases}$$

Here X_n, Y_n, and Z_n are the tristimulus values of the reference white.

Color Histogram

Color Histogram is a very prominent feature in image classification problems. Color Histogram is a way to represent the cumulative distribution of colors in an image. It calculates the number of pixels lie in particular color range. The color ranges are histogram bins for the color histogram model. Color Histogram are independent to the rotational state of image. Therefore even if the image is tilted, it won't effect the classification. Apart from that computation efficiency is another strong reason behind using color histogram in classification problem. The disadvantage of color histogram is that it fails to capture the spatial distribution of the color in images and only captures the color information.
As we described earlier in RGB color space subsection, RGB color space is dependent on the device., because of this device dependency, RGB does not feel to be a right choice for our color histogram. Adding to that, CIELAB color space appears to be a better choice because of its perceptual uniformity and device in-dependency. Perceptual Uniformity means same quantity of perceptual effect is generated with same quantity of change in color values or we can say that visual effect is proportional to color values.
We first used inbuilt MATLAB functions to convert the given images from RGB space to CIELAB color space. After that we calculated the color histogram after dividing the image into 16 parts called blocks. Now we construct bins of 4 for each L, a and b components. Thus we have totally 64 color bins. Thus for each block we get a 64 dimensional color histogram. When we combined the vectors we get a vector of $(16 \times 64) = 1024$ dimensionality.

4.1.4 Texture/ GLCM Feature Extraction

In order to extract some meaningful information from an image, it is important to get into human interpretable features. There are three types of such features which lead to perceptive interpretation of color images: spectral, textual and contextual features.
In such human interpretable features, texture comes as an important feature. In normal terms we can define texture of an image as estimate of smoothness of that image. In

everyday terms, texture can be defined with words as rough, bumpy, silky.

A texture, which is rough, when touched has large difference between high and low points, with the distance between those points be very low. A smooth texture will usually has small difference between high and low points, with these points be distant.

Image texture also works in the same. Except the high and low values, we have brightness values (also referred as grey levels) instead of elevation. Instead of using hand or finger to judge the surface, a window or box is used to define the size of probe.

GLCM or Gray Level Co-occurrence Matrix acts like a texture indicator for an image. This co-occurrence matrix represents the inter-pixel distance and spatial relationship between gray values over an image. This spatial interrelations of the grey tones actually determines the textural pattern.

It was first introduced by Haralick et al [36], that's why it is also called as Haralick features .

For calculating GLCM feature for an image, we first convert that image into a gray-scale one, because GLCM is actually an estimate of the occurrence of different combinations of pixels in a gray-scale image. The gray level co-occurrence matrix $G(i, j, \theta, d)$ can be as follows. The value of $G(i, j, \theta, d)$ is count of occurrences of the pair of pixels having gray value i and j, where the distance between these pixels as d and direction specified by angle θ. In standard GLCM matrix, the angle are considered as 0°, 45°, 90°and 135°with $d = 1$ *pixel*. This directional component of θ makes it more powerful in a sense that it represent features from every angle of an image.

Figure 4.6 illustrate the process of finding co-occurrence matrices using $N = 5$ levels. IT is showing gray-scale co-occurrence G(0°,$d = 1$). We can observe that pixels (0,2) of the input is shown in G(0°,$d = 1$) as 2 because we only have two occurrence of the pixel intensity value 0 with horizontally adjacent pixels with intensity = 2 in the input. We computed matrix G as symmetric, as we considered pair (0, 2) as (2, 0) as well. Matrix G

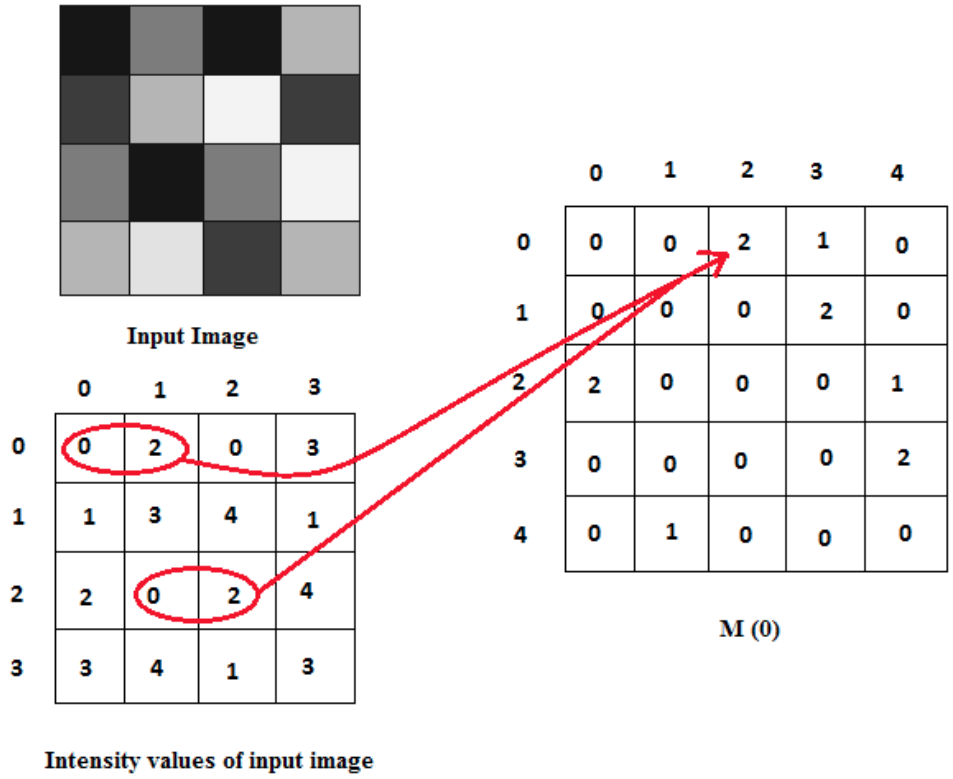

FIGURE 4.6: Co-occurence Matrix G(0°) generation for N=5 levels

can also be computed with non-symmetric measure.

In our approach, we computed G for all θ angles using $N = 8$ with symmetry because increasing the gray levels further was decreasing the accuracy and lesser number of gray levels may not be sufficient enough to capture the texture adequately. We used MATLAB function to get this matrix. This step gives us four 8×8 matrices. As input is not re-sized to some predefined dimensions, we normalized each matrix for better comparison. After normalization, we got 1×64 dimensional vector for each matrix. We merge these to get 1×256 size vector for an image.

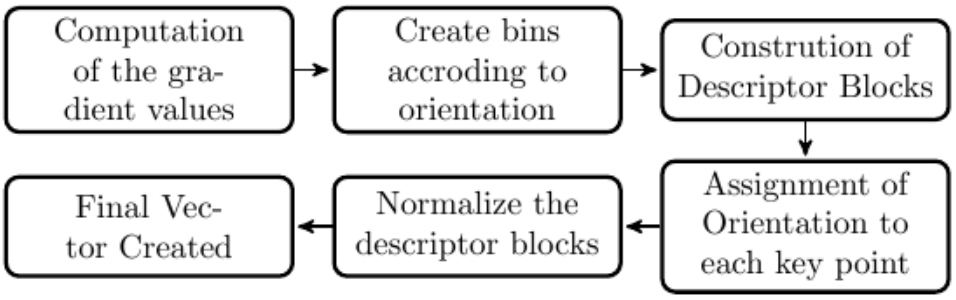

FIGURE 4.7: Construction of HoG descriptors

4.1.5 HOG-LBP Features

HOG or Histogram of Oriented Gradients in one of the prominent used features for object recognition in Computer Vision. The idea behind this descriptor is that the object in an image can be shown by the some intensity gradients or distributed edge directions. HOG descriptor works in a localized region, therefore it does not get affected with illumination changes or geometric transformations like rotation, scale, or change in viewpoint. These descriptors were first used by Dalal and Triggs [37] for human in 2005. After that these descriptors clubbed with LBP features are usually practiced for object recognition [38], [39],[40] etc. The steps of constructing HoG descriptors are shown in figure 4.7.

Local Binary Pattern (LBP) is a texture classification feature first introduced by Ojala et al. [41] in 1994. LBP captures the appearance of an image in a neighborhood of the pixel. A LBP is a string of bits, which contains one bit for each of the pixels in the neighborhood. LBP does not get affected by monotonic gray level changes and acts as a good discriminator. Wang et al. [42] tried combining HoG with LBP. The results indicated high improvement in performance in case of object detection. HoG looses its discriminating capability, if the image is cluttered with blurred edges. LBP acts as complementary to HoG in such cases. LBP uses uniform pattern to remove out the noisy edges from such

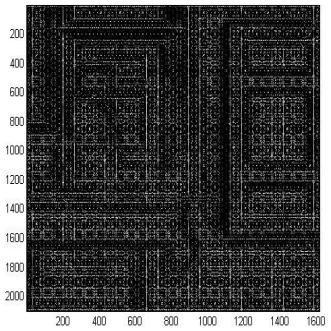

FIGURE 4.8: Example of HOG Descriptors

cluttered image. In [43], Santana et al. has also shown that the combination of HoG and LBP acts much better than the individual ones. We, therefore, consider combination of HoG and LBP for our classification.

We used VLFEAT [31] library for computing HoG features. VLFEAT has two variants of HoG. One is UoCTTI variant, other is Dalal and Triggs's variant [37]. We computed the UoCTTI variant HoG on each painting. This variant computes directed and also undirected gradients. Apart from this, it also has 4-dimensional texture-energy feature on a window size of 16 . We therefore obtain 31 dimensional HoG vector for each cell.

For computing LBP features, we again used VLFEAT library [31]. VLFEAT considers $3 * 3$ neighborhood, this leads to LBP feature of 8 bit long string vector. This 8 bit long vector can assume $2^8 = 256$ possible values. These 256 possible values are further quantized into a smaller number of patterns. This uniform quantization makes LBP features computationally efficient.

In this uniform quantization, we use following observations. There is one quantized pattern, for every bit, which has exactly one transition from 0 to 1 and one from 1 to 0 when scanned in anti-clockwise order. Plus one pattern comprising of two uniform LBPs and one pattern comprising all other LBPs. These observations yields total 58 patterns. When we concatenate both HoG and LBP vector descriptor we get combined vector of 89 dimension, Now we use bag of words approach on this combined HoG-LBP vector. We form a bag of 4000 visual words using the K-means and after that we combine the histogram on this visual dictionary for each vector, which gives a vector of 4000 dimension for each image.

In figure 4.8, we have shown the HOG descriptor on an image. The left part is the original image and right part show the HoG descriptor for the image.

4.2 Social Content Based Feature Extraction

Social meta-data obtained from images has similar properties like text data-set, because we have tags, comments, groups all in normal language text. So, It makes sense to just only use the text processing methods here. But, this text data also has inherit structure of social network. We, therefore, to utilize this extra aspect, first construct node features over the social-metadata for each image as shown by [23]. These node/social feature vectors have high dimensionality. We, therefore, use the topic modeling/text processing methods over these social features to construct a better and reduced representation. In [19], Tang and Jie et al. have shown that such topic-level modeling of social-networking data leads to good result in finding patterns and inferences. These final low-dimensional features project the semantically close node features (like mountain, hill etc.) near to each other. We tried Latent Semantic Indexing(LSI), Latent Dirichlet allocation (LDA) and Random Projection (RP) methods for the purpose of dimesionality reduction and topic modeling. The process of constructing useful feature vectors out of the social meta-data obtained can be divided into following steps:

- Pre-analysis of Social Data
- Constructing Node Features
- Applying Topic Modeling/Text Processing Methods on Binary Social Features

In the following part of this section, we will discuss each of these steps.

4.2.1 Pre-analysis of Social Data

We first do preliminary observation about the tags and featured groups for image instances. The elementary observation suggests that tags are less structured are provided by any number of annotators and can include the information that is not easily detectable from content alone, such as location like sea-side or mountain ranges. Groups are similar to rags, with difference that the groups to which an image is featured, are chosen entirely by image's author.

4.2.2 Constructing Node Features

There are some properties, which can be defined for a single image instances. eg. tags, groups etc. We call such features as node features, because these properties can be separately defined for each image/node.

We first construct indicator vector encoding those words, groups and tags that appear in an image. For this , we first consider the 1000 most popular words, groups and tags across the entire data-set . As described in [23], this data set of only 1000 most popular words does not sufficiently represent the whole data. We, therefore, also consider any words, group and tags that our at least twice as frequently in images having the questioned label compared to the overall rate. This way we will get the similar node features as described in [23]. For developing this word feature, we utilize text from the image's title, it's comment thread, description after eliminating stop-words. This will give us more than 40000+ points. The node-feature vector is in a binary form and have high dimension. We have 0 and 1 as the value for each field in this vector corresponding the presence of word in the image data or not. We further convert this raw binary 0 and 1 form, in usable social features with the use of advance text processing methods like Latent Semantic Indexing and Random Projections.

4.2.3 Applying Text Processing/Topic Modeling Methods on Binary Social Features

We, now, use dimensionality reduction cum text processing methods over these feature vectors. For text processing, we can consider each image as a document and the current node feature vector as a representation of dictionary. This vector represents the presence of a word in the document. Now, we have corpora of word features and in next step, we just need to use some topic modeling/text processing methods to get a feature vector with reduced dimension. We experimented with Latent Semantic Indexing, Random Projection and Latent Dirich-let Allocation as some possible methods to do such dimensionality reduction for sparse binary data.

4.2.3.1 Latent Semantic Indexing

Latent Semantic Indexing is actually a singular value decomposition method to identify patterns in the relationship among the semantic concepts in an unstructured collection of text. It is normally used for extraction of conceptual content of a body of text by establishing associations between those words which are show a similar contextual presence.[44] LSI is used in a variety of information retrieval and text processing applications which are increasingly used for electronic document discovery, publishing, government/intelligence community. [45] In typical information retrieval methods, an information retrieved by literally matching terms in search space with those of search query. However, this method, which purely depends on lexical matching, can be inaccurate. Since, there are many ways to express a given concept in real world, the literal matching may not provide us the relevant information. A better approach will be create a basis of conceptual topic for the search space. Latent Semantic Indexing tries to overcome the problems of lexical matching by using statistically derived conceptual indices instead of raw data. Latent Semantic Indexing assumes that there is a hidden latent conceptual structure in raw features and which is not visible because of variability of word choices. A truncated SVD (Singular

Value Decomposition) is used to estimate this latent semantic structure. These statistically derived vectors proves to be more robust indicators of information than individual terms.

4.2.3.2 Basic Concept

Latent Semantic Indexing is a technique that projects the feature vectors into a space with "latent" semantic dimensions. In latent semantic space, two feature vectors can have high cosine similarity even if they do not share any terms - as long as their terms are semantically similar in a sense to be described later. We can look at LSI as a similarity metric that is an alternative to word overlap measures like tf.idf. In terms of topic modeling and text processing, latent semantic indexing is the application of Singular Value Decomposition or SVD, to a "word-by-document" matrix. The projection into the latent semantic space is chosen such that the representations in the original space are changed as little as possible when measured by the sum of the squares of the differences. SVD represents a matrix A as $\hat{A}$ in a lower dimensional space such that the "distance" between the two matrices (Which is measured by the 2-norm is minimized): [1]

$$\delta = \|A - \hat{A}\|_2$$

SVD project an n-dimensional space onto a k-dimensional space where $n \ll k$. Thus, the projection transforms a feature vector in n-dimensional word space into a vector in the k-dimensional reduced space. We used GENSIM library [46] in python for Latent Semantic Indexing of our data. It was developed by Radim Rehurek and Petr Sojka [47] for topic modeling with large corpora.

[1] The 2-norm for matrices is the equivalent of Euclidean distance for vectors.

4.2.3.3 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is an another frequently used process in natural language processing. It is a generative model that allows set of observations to be depicted by unobserved groups explaining why some parts of the data are quite similar. For example in general natural language processing scenario, when the observations are words associated with a document, it tries computation to provide observations that each document is a mixture of a small number of topics and the each word's presence is dedicated to one of the document's concept. LDA was actually a graphical model presented in [1] for topic discovery. LDA has connection with image classification because in [48] a variation on LDA was used to automatically pit the natural images into categories, such as forest and mountain, by assuming the images as words. Latent Dirichlet allocation (LDA)uses generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. There are many variants of LDA computation. In [1], LDA assumes the following generative process for each document w in a corpus D:

- Choose $N \sim Poisson(\xi)$.
- Choose $\theta \sim Dir(\alpha)$.
- For each of the N words wn:
  - Choose a topic $z_n \sim Multinomial(\theta)$.
  - Choose a word w_n from $p(w_n|z_n, \beta)$, a multi nomial probability conditioned on the topic z_n.

In figure 4.9, we have shown an example of Linear Discriminant Analysis on a two-dimensional data set. The line in between discriminate the data set in two classes, where as the colors show the actual class labels. In our case the LDA does not provide a good

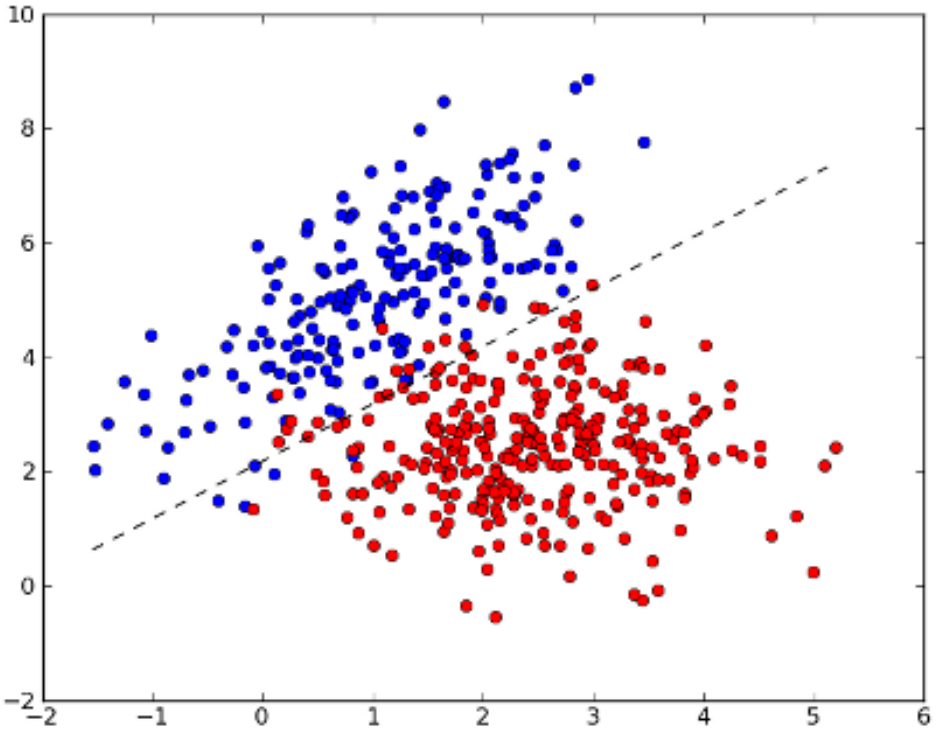

FIGURE 4.9: Example of LDA in two-dimensional feature vector [1]

results because we already have a textual data, which is oriented to solve a binary classification problem. This binary nature of topic leads to a too sparse feature generation from Latent Dirichlet allocation (LDA). This sparseness of features around images lead to low quality classification. LDA also fail if discriminatory information is not in the mean but in the variance of the data [1]. We, therefore, discard this method for our computations.

4.2.3.4 Random Projections

Random Projections are a powerful methods for dimensionality reductions in application to image and text data. [49]. In [50], Bingham introduced the random projections as a simpler and less erroneous dimensionality reduction tool for information retrieval from text and processing of images. It is very instrumental in such case where reduction of the high dimensional data in a low dimension is essential, which if not reduced leads

to heavy computation penalty without any significant gain. Using random projection is significantly less expensive compared to techniques like principal component analysis. In random projection, the original high dimensional data is projected onto a lower dimensional subspace using a random matrix R. In random projection, the original d-dimensional data is projected to a k-dimensional ($k << d$) subspace through the origin, using a random $k \times d$ matrix R whose columns have unit lengths. Using matrix notation where $X_{d\times N}$ is the original set of N d-dimensional observations,

$$X^{RP}_{k\times N} = R_{k\times d} X_{d\times N}$$

is the projection of the data onto a lower k-dimensional subspace. The key idea of random mapping arises from the Johnson-Lindenstrauss lemma [51]: if points in a vector space are projected onto a randomly selected subspace of suitably high dimension, then the distances between the points are approximately preserved. We write the Euclidean distance between two data vectors x_1 and x_2 in the original large-dimensional space as $\|x_1 - x_2\|$. After the random projection, this distance is approximated by the scaled Euclidean distance of these vectors in the reduced space:

$$\sqrt{d/k}\,\|R_{x_1} - R_{x_2}\|$$

where d is the original and k the reduced dimensionality of the data set. The scaling term $\sqrt{d/k}$ takes into account the decrease in the dimensionality of the data: according to the Johnson-Lindenstrauss lemma [51].the expected norm of a projection of a unit vector onto a random subspace through the origin is $\sqrt{k/d}$ The choice of the random matrix R is one of the key points of interest. The elements r_{ij} of R are often Gaussian distributed, but the Gaussian distribution can be replaced by a much simpler distribution such as

$$r_{ij} = \sqrt{3} * \begin{cases} +1 & \text{with proabibility } \frac{1}{6} \\ 0 & \text{with proabibility } \frac{2}{3} \\ -1 & \text{with proabibility } \frac{1}{6} \end{cases}$$

In fact, practically all zero mean, unit variance distributions of r_ij would give a mapping that still satisfies the Johnson Lindenstrauss lemma. This means further computational savings in feature computation, as the computations can be performed using integer arithmetic. Again for computing the random projections, we used the GENSIM library [46] in python. It has been found that even after being computationally light, Random projections is sufficiently accurate method for dimensionality reduction of high dimensional data [52].

4.2.4 Implementation of Dimensionality reduction

Considering that we are using a large database and we need to do dimensionality reduction for such a data. We use an online version of aforementioned techniques. So that we don't have to bother about loading a whole data into memory space.
Both LSI and RP rely on TF-IDF (term frequency - inverse document frequency) as a fast pre processing step .
[47] gives a framework [46] for doing all these text processing on large corpora in memory independent fashion. We use this as a tool to doing LSI and RP Compuation. On varying the number of dimension, in dimensionality reduction, we found that using 300 features in LSI and 400 features in RP gives us the best results.
In selecting the dimension the whole point is to reduce dimensionality in such a way that we can go non-linear, which would be too costly and too susceptible to over

tting with thousands of binary features. We directly convert the node features of dimension 40000+ in social features of dimension 300 (LSI) and dimension 400 (RP). This conversion has been done in python using GENSIM and the converted file are in the LIBSVM format.

Chapter 5

Experimental Results

In this chapter, we first give an overview of the visual and social feature vectors constructed and after that we discuss the classifiers tested. In remaining part of chapter, we describe the results on four benchmark data-sets.

5.1 Feature Vectors

We extracted the five image features SIFT, GIST, HOG-LBP, CIELAB color space vector and GLCM as described in the previous chapters. Apart from these visual features, we construct two feature vectors on the basis of social meta data. These feature vectors are constructed after doing analysis of social data and obtaining a binary feature vector, indicating the presence or absence of a social/textual element for the image. This high dimensional binary vector leads to two different feature vectors. One which is made of dimensionality reduction by Latent Semantic Indexing and second by which is the resultant of Random Projections.

5.2 Classifiers Used

After experimenting with various classifiers like Random Forest, MLP (Multi Layer Perceptron), libSVM (with various kernels linear, RBF, χ^2, histogram intersection), we found out that in case of image features libSVM with χ^2 kernel works best and in case of social features libSVM with the linear kernel gave us the best results. We,therefore, use libSVM with χ^2 kernel as classifier for visual feature vectors and libSVM with Linear Kernel as classifier for social feature vectors.

5.3 Classification Results

In the following part of this chapter, we have shown classification results on various labels of four data sets as mentioned in 3 .

The results are divided in four subsections according to four data-sets. We have shown classification results from all the features extracted and then provided the result of ensemble of all those feature.

We have given a qualitative and quantitative conclusion/observation of our results. We have also shown a comparison with the results directly published material on each of these four benchmark, or associated competition.
The goal of all these comparison is to assess the improvement that can be obtained by using social meta-data of images. We report the mean average precision (MAP) for the sake of comparison with published materials and competition results. We also give the accuracy for this binary prediction/classification of labels.
All these results are for tenfold cross validation. For ensemble method, we have divided the data in three parts: training, testing and validation set. We randomly chose the 10% feature instances for validation set and learn weights for the linear combination of the

classifiers, which provide best results. After learning those weights we use these optimized combination of these classifiers to do classification testing on test set.

5.3.1 MIR Flickr collection

MIR has images has high quality photographic images. It has a rich meta data attached with it. This provides a wide variety of image retrieval bench-marking scenarios.

In [53], a combination of social data and low-level content-based descriptors to improve the accuracy of visual concept classifiers. We use the results of this paper as a comparison metric for our results. In [53], they have used the following four sets of image features:

- HMMD Color Histogram descriptor.
- Spatial Color Mode descriptor.
- MPEG-7 Edge Histogram
- MPEG-7 Homogeneous Texture descriptor

Apart for these low-level content based descriptors they also use flicker tags of visual concepts. A set consisting of 293 binary features is developed using these tags. The tags corresponds with at least 50 images in the MIR Flickr collection.
They have further compared the classification accuracy between classifiers based on low-level features only and classifiers that additionally use the Flicker tags (Set 5 above) as features.
They have used two classifiers one is Linear Discriminant Analysis and other is support vector machines.

We have shown the map comparison of our computation and result of [53]in table5.1.
The accuracy obtained with our computation is shown in table 5.2.

Observations

- The classifications based on visual only features give the average precision of 76.15%, which outperforms the low level image descriptor based classification with 40.43% in case of LDA and 44.38% in case of SVM. The result is also better than classification based on the combination of low level image descriptors and flicker tags. Here we see a precision increment of 28.28% as compared to SVM results in paper and 26.40% as compared to LDA results in paper.

- When we do classification on the basis of social features computed using LSI, we get an average precision of 87.59%. This is 39.71% more compared to SVM classification of combined features of flicker tags and low level image descriptors and 37.83% more compared to LDA classification of this combined feature set.

- LSI based social feature classification outperforms our visual only classification with 11.43% precision increment. It shows a precision gain of 51.86% and 55.81% on the low level image descriptor based SVM classification and LDA classification respectively.

- Ensemble of the social features and visual features provides average precision of 90.58% which is 40.83% more compared published result and 2.99% more compared to LSI method.

- This 42.70% precision gain compared to published result is coherent with the results shown in [Jure paper]. They obtained a precision gain of 42% using the social features.

- Our LSI based method works better for all labels except *Clouds*. The images in this label shows better result with HOG_ LBP features. This is due to low volume of social information attached with the data and the high visual concepts entailed in these images, which is more emphasized by HOG-LBP features.

- The results of social features and image features are quite close for *night, tree, sea* and *river*. This is the result of a great degree of visual information present with

these images contrast to social information in comments (or other social entities), as they are natural outdoor photographs.

- The classifications based on visual only features give the average accuracy of 76.03%.
- The classifications based on social only features with LSI pre-computation give average accuracy of 86.25 outperforming the visual features only method with 10.22%.
- The fusion of all the features provide an average accuracy of 88.80% outperforming the social feature only accuracy with 2.55%.
- GLCM plays an important role in the case of *Bird , male , baby* labels.
- GIST plays a pivotal role in case of labels *female, male* and *tree*.

5.3.2 ImageCLEF

ImageCLEF has 99 labels. For some labels we have less than 20 instances. So learning visual bag words for such a low number of instance is not favorable. We, therefore, discard such labels. After doing analysis on the available instances of labels and their meta data, we decided to do classification for the following Labels.
'Adult','Aesthetic_ Impression','Animals','Autumn','Citylife','cute','Day','Flowers','Food', 'Graffiti' ,'Landscape_ Nature','Painting','Portrait','Single_ Person','Sky','Street', 'Summer','Sunset_ Sunrise','Vehicle','Winter'

The ImageCLEF competition [26] has the best comparative published results for judging our hypothesis as it already has results based on Flicker User tags and multi-modal approaches that consider visual information and/or Flicker user tags and/or Exif Information.

Table 5.3 shows the mean average precision obtained for ImageCLEF Data. Table 5.4 shows the average accuracy obtained for ImageCLEF Data.

Observations

- While comparing MAP for the labels we find that when we use social features with LSI, even then we can achieve an average improvement of 4.09% compared to best results (visual, multimodal, textual) used in CLEF competitions.
- Ensembling of all the features outperforms the published results in precision with average of 6.93%.
- When we use the social features with LSI method, even then we can obtain comparable accuracy with the CLEF best results.
- Ensembling of all the features outperforms the published results in accuracy with average of 1.26%.

TABLE 5.1: MIR: Precision Comparison

Labels	Ensemble Method	MIR Published Results				Individual Feature Based Classificatiom						
		SVM		LDA		Social Features			Visual Features			
		Flicker tags + Image Descriptors	Image Desciptors Only	Flicker tags + Image Descriptors	Image Descriptors Only	LSI	RP	HOG-LBP	SIFT	GIST	COLOR	GLCM
flower	92.96	48.00	46.90	56.00	30.10	91.37	74.34	78.94	41.60	73.19	70.37	64.40
car	96.91	33.90	17.90	29.70	14.20	92.21	72.68	83.27	56.54	71.70	64.35	76.65
bird	95.77	44.30	12.80	42.60	9.70	93.37	79.26	70.11	47.96	61.77	60.63	71.17
dog	98.15	60.70	15.50	62.10	10.80	95.94	73.38	73.68	47.39	68.80	65.31	67.67
night	90.53	58.80	55.40	61.50	51.50	87.81	71.85	82.40	58.00	70.40	79.16	81.37
tree	84.05	55.90	51.40	51.50	43.40	81.43	69.18	76.92	54.59	74.97	61.64	76.11
clouds	92.00	69.50	65.10	65.10	57.70	82.75	74.12	90.91	50.88	79.78	67.31	76.55
portrait	84.08	48.00	49.30	54.30	43.20	83.54	71.20	69.31	50.61	65.30	59.99	65.85
female	83.55	46.40	46.10	49.40	40.40	81.52	69.90	63.72	52.56	62.27	54.11	58.77
male	76.34	41.30	40.70	43.40	35.60	74.51	67.04	60.16	50.62	60.96	52.88	61.79
people	93.91	74.80	36.10	73.10	62.80	90.77	73.32	68.36	60.37	58.19	55.56	59.18
sea	98.01	52.90	36.60	47.70	25.50	93.34	73.85	88.48	53.04	77.20	73.42	81.05
river	83.35	15.80	17.90	31.70	13.00	82.39	73.58	77.14	46.68	69.73	68.55	71.10
baby	98.55	20.00	8.40	28.50	6.90	95.24	73.24	77.58	47.71	72.29	73.86	80.06
Average	90.58	47.88	35.72	49.76	31.77	87.59	72.64	75.79	51.33	69.04	64.80	70.84

Table 5.2: MIR: Accuracy Comparison

Labels	Ensemble Method	Social Features	Visual Features						
		LSI	RP	HOG-LBP	SIFT	GIST	COLOR	GLCM	
animal	94.09	83.46	90.01	70.44	73.46	51.62	66.21	57.18	64.11
coral	95.45	45.73	94.82	83.10	79.45	53.11	65.22	69.23	67.64
dancing	95.98	53.66	92.82	74.07	75.59	59.13	67.39	63.30	60.74
harbor	93.63	36.47	92.24	77.58	81.40	53.22	79.32	72.40	73.07
military	97.86	55.72	93.07	79.28	70.67	56.42	61.35	65.02	71.94
mountain	90.05	34.23	89.26	72.42	84.30	54.58	78.03	78.62	72.60
snow	92.32	42.43	87.54	66.07	69.11	56.66	60.82	64.92	67.05
statue	88.56	54.63	84.56	73.67	57.66	54.79	59.57	52.99	56.49
tattoo	88.92	56.24	86.81	71.95	83.43	58.64	70.74	71.13	75.49
temple	93.68	43.78	89.72	69.59	64.42	54.15	57.92	58.67	58.14
waterfall	98.58	57.38	95.88	82.59	86.45	54.65	76.68	76.56	83.82
wedding	94.22	72.34	89.98	76.31	66.43	54.44	56.91	61.50	60.27
AVERAGE	93.61	53.01	90.56	74.76	74.36	55.12	66.68	65.96	67.61

- For the three labels *Landscape_ Nature, Sky* and *Sunset_ sunrise* our method provides lesser accuracy and precision because these labels were more connected to visual data and social data on them were less in amount.

- For the labels like Autumn, Landscape, Autumn, CityLife, Paintings, Sky, Street and Sunset_ Sunrise we see that visual feature based computation is works quite well because these images are more visual concept centric and social data has lesser role to play there.

- HOG-LBP feature works best among all the features for all the labels except painting and single_ person. In these two cases GLCM works better because GLCM reads the texture of the image and texture of a 'Painting' or 'Single_ Person' based image plays pivotal role because it is quite different than the normal natural photo clicks.

5.3.3 PASCAL

PASCAL data-set is actually a competition which is based on the challenge of recognizing visual object classes in realistic scenes. Therefore, the data-set is actually very much visual content specific and has image instances having information specifically related to these objects. We can actually subclass the 19 labels, we used in our computations, in following 3 sub classes.

- Animal: bird, cat, cow, dog, horse, sheep
- Vehicle: aeroplane, bicycle, boat, bus, car, motorbike, train
- Indoor: bottle, chair, dining table, potted plant, sofa, tv/monitor

Table 5.5 shows the MAP comparison of various labels and also comparison with the VOC Competition results. Table 5.6 shows the accuracy of classifying various labels in binary prediction environment.

Observations

- Observations:
- The classification based on only visual features gives an average precision of 67.50% with maximum for 'aeroplane' of 78.79% and minimum for 'potted plant' of 59.72%.
- The classification based on only LSI computed social features outperforms the competition's results with an average of 16.74% better MAP and only visual features result with 6.17%. The average precision obtained is 73.67%.
- Fusion of social and visual features provides a better precision of 76.86% which is 3.19% more than only social features and 9.36% more than usual visual classification.
- Our ensemble method and social feature based method gives better precision for 17 labels out of 19 labels considered compared to published results.

TABLE 5.3: ImageCLEF: Precision Comparison

Labels	Ensemble Method	CLEF Best Results	Social Features		Visual Features				
			Latent Semantic Indexing	Random Projection	HOG-LBP	SIFT	GIST	COLOR	GLCM
Adult	91.76	77.21	89.56	75.24	60.94	56.88	54.70	56.45	60.24
Aesthetic_ Impression	67.74	60.66	63.91	59.30	59.58	53.33	54.94	55.60	50.82
Animals	93.76	84.34	92.83	74.05	65.67	54.52	58.87	58.65	60.09
Autumn	88.12	83.51	85.89	65.67	76.01	69.43	67.71	58.74	65.18
Citylife	82.02	78.37	78.47	63.79	69.87	44.81	62.13	52.60	62.63
cute	64.49	59.71	63.11	61.01	55.11	51.49	52.67	52.41	51.13
Day	87.43	80.75	84.15	68.67	67.39	54.79	59.98	63.30	61.41
Flowers	94.13	82.72	93.54	69.79	73.75	52.97	65.75	65.94	57.58
Food	92.30	85.20	87.51	71.04	73.30	68.41	72.84	67.89	59.68
Graffiti	84.09	66.21	81.17	67.22	66.02	61.50	61.32	56.33	56.63
Landscape_ Nature	88.21	88.68	84.45	75.04	77.95	52.49	66.03	62.66	68.30
Painting	73.04	72.43	70.51	59.10	60.26	51.96	55.70	53.26	62.49
Portrait	90.27	81.34	85.82	75.59	64.98	57.47	60.80	61.90	63.16
Single_ Person	93.99	76.41	91.25	76.48	58.97	56.14	56.37	54.80	60.28
Sky	88.05	89.26	87.30	76.27	80.42	54.78	73.07	69.15	67.28
Street	83.41	76.76	79.20	65.91	67.86	60.37	62.71	53.38	63.46
Summer	75.87	74.08	71.80	65.56	62.96	52.30	56.92	59.51	59.18
Sunset_ Sunrise	91.74	91.83	87.09	78.51	85.49	57.40	71.96	71.15	82.35
Vehicle	88.97	78.64	87.98	69.92	74.42	51.68	62.11	58.25	69.10
Winter	88.10	80.71	85.02	75.90	69.89	60.94	58.30	58.47	64.10
Average	85.37	78.44	82.53	69.70	68.54	56.18	61.74	59.52	62.25

TABLE 5.4: ImageCLEF: Accuracy Comparison

Labels	Ensemble Method	ImageCLEF Competetion Results	Social Features		Visual Features				
			Latent Semantic Indexing	Random Projection	HOG-LBP	SIFT	GIST	COLOR	GLCM
Adult	90.03	80.73	88.50	75.17	60.33	55.50	53.83	56.33	60.67
Aesthetic_ Impression	68.01	60.89	64.83	58.83	59.33	45.33	54.50	54.83	48.83
Animals	92.13	88.43	90.17	73.17	65.67	53.17	58.67	58.00	59.33
Autumn	86.16	88.31	83.57	62.14	74.29	66.43	63.57	57.86	63.57
Citylife	81.02	82.67	78.00	63.50	69.83	47.67	62.83	52.33	62.83
cute	65.72	59.24	63.00	60.17	55.17	49.50	52.17	52.00	47.00
Day	82.82	85.17	82.17	67.17	66.83	49.33	58.67	62.67	62.00
Flowers	91.74	87.03	89.75	69.75	72.25	52.25	65.75	65.25	56.75
Food	87.36	88.70	86.00	71.67	73.33	51.67	72.33	67.00	60.33
Graffiti	75.90	69.85	75.00	63.57	56.43	50.71	57.14	53.57	56.43
Landscape_ Nature	84.69	91.54	83.67	73.67	77.00	51.50	66.00	60.83	67.50
Painting	71.96	76.28	69.17	57.78	60.00	51.67	50.28	53.06	61.39
Portrait	90.50	85.59	86.67	75.17	66.00	52.00	60.33	62.00	63.33
Single_ Person	91.89	80.04	91.33	75.00	58.33	47.17	55.67	54.33	59.67
Sky	87.36	91.75	86.33	76.50	79.50	52.67	71.83	67.33	65.50
Street	79.52	81.20	78.50	65.67	67.83	50.83	63.33	53.00	64.17
Summer	71.97	78.50	71.00	65.00	62.67	52.00	56.83	58.33	59.67
Sunset_ Sunrise	88.24	93.24	86.84	77.89	85.00	55.00	72.37	70.53	81.05
Vehicle	88.50	82.38	87.50	70.00	73.50	48.33	61.83	57.83	69.33
Winter	86.48	85.36	84.58	76.25	70.00	57.50	58.75	56.25	64.17
Average	83.10	81.84	81.33	68.90	67.66	52.01	60.83	58.67	61.68

- The accuracy obtained with using only visual features is average of 66.68%. With minimum 60.67% for 'cat' and maximum of 73.50% for 'bottle'.

- While comparing the accuracy of our classification, we find that accuracy obtained using LSI based social features is 70.22%, which is greater than 3.54% for the visual only features.

- The fusion of social and visual features provides accuracy of 72.49 % which is 2.7 1% more than only social features based classification and 5.81% more than only visual features based classification.

- The accuracy for visual only features of the most labels are comparable with social features and not have difference more than 3%. This difference is much higher for all other three data-set s. This low difference shows that the social data available for the PASCAL Data-set is not as enriched as other data-sets. The images are not contextually interesting as compared to other data-sets, so do not encountered with much human interaction leading to lesser social data.

5.3.4 NUS

In [27], authors have used six types of low-level features extracted. These six type of image features include 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT descriptions. Further they have used traditional k-NN algorithm on these features to provide the baseline results for web image annotation. We use the results of this paper as a comparison base.

Table 5.7 shows the MAP comparison of various labels and also comparison with the VOC Competition results. Table 5.8 shows the accuracy of classifying various labels in binary prediction environment.

Observations

- The classifications based on visual only features give the average precision of 74.63% which is better with the large margin of 10.13% from the published results in [27].
- The classifications based on social only features (with LSI pre-computation) give the average precision of 90.56% outperforming the results obtained from the visual only features with the margin of 15.93%.
- Fusion of all the features and ensemble classification leads to the average precision of 93.61%, which is better than precision of social only features with 3.05% margin. This result is way better than baseline results mentioned in [27] and shows a margin of 29.11% in the precision.
- The classifications based on visual only features give the average accuracy of 74.84%.
- The classifications based on social only features (with LSI pre-computation) give the average accuracy of 88.61% outperforming the results obtained from the visual only features with the margin of 13.77%.

Table 5.5: PASCAL: Precision Comparison

Labels	Ensemble Method	PASCAL VOC Results	Social Features		Visual Features				
			Latent Semantic Indexing	Random Projection	HOG-LBP	SIFT	GIST	COLOR	GLCM
aeroplane	85.83	77.50	81.76	70.60	75.42	56.53	72.82	65.20	78.79
bicycle	68.29	63.60	67.65	60.58	67.51	55.38	60.60	53.57	62.41
bird	82.38	56.10	79.22	69.80	62.49	55.64	65.89	50.86	65.30
boat	68.54	71.90	63.67	54.88	67.14	57.14	63.19	57.79	62.99
bottle	71.09	33.10	66.30	61.95	57.54	48.67	55.70	57.80	62.44
bus	77.58	60.60	76.79	63.74	74.26	56.43	66.24	57.97	66.13
car	73.50	78.00	68.71	56.02	61.25	54.20	58.41	54.69	60.63
cat	81.73	58.80	77.73	68.49	69.36	56.27	64.01	59.97	68.32
chair	73.87	53.50	71.76	61.50	61.19	54.38	51.94	57.86	61.14
cow	75.59	42.60	71.63	63.67	68.83	56.62	61.20	60.39	64.33
diningtable	85.35	54.90	82.07	70.89	67.67	62.70	65.11	57.66	62.59
dog	83.18	45.80	78.93	64.97	65.07	49.69	61.33	55.64	64.52
horse	82.19	77.50	78.80	65.44	67.49	46.98	58.99	56.93	62.26
motorbike	80.51	64.00	76.80	59.73	66.52	58.15	68.11	55.44	62.67
pottedplant	67.87	36.30	64.34	57.02	59.60	53.57	55.97	59.72	57.10
sheep	80.13	44.70	78.75	65.73	73.34	49.68	67.70	62.14	68.27
sofa	70.64	50.90	70.16	63.35	64.89	53.21	56.27	54.22	62.71
train	86.96	79.20	83.48	69.57	71.61	56.76	65.79	57.61	63.24
tvmonitor	74.00	53.20	69.25	61.77	66.90	57.38	67.69	59.13	67.84
Average	76.86	56.93	73.67	63.28	66.26	54.60	61.90	57.19	63.61

Table 5.6: PASCAL: Accuracy Comparison

Labels	Ensemble Method	Social Features		Visual Features				
		Latent Semantic Indexing	Random Projection	HOG-LBP	SIFT	GIST	COLOR	GLCM
aeroplane	78.46	74.5	69	73.67	47.33	70.67	64.17	74.83
bicycle	69.82	65.17	59.17	66.17	54.67	60.17	53.17	62.5
bird	75.22	74.83	68.33	62	55.67	64.83	51	63.33
boat	67.85	62.5	54.83	65.67	52	62.67	57	61
bottle	67.69	63.83	60.33	56.33	48.67	55.33	52.17	61.83
bus	77.38	71.5	62.83	73.5	55.5	66	57.5	65.83
car	69.77	67.83	56	61.5	49.5	57.67	54.17	60.5
cat	74.07	73.5	68	68.33	55.83	63.83	59.5	67
chair	72	68.33	60	60.67	54	51.33	56.83	60.5
cow	73.17	69.33	62.83	68.83	51.5	61.33	60	63.67
diningtable	75.81	75.67	68	67	47.83	63.33	57.33	63
dog	77.9	74.17	63.17	64.5	47	60.83	54.67	64.83
horse	76.7	73.67	64	67.83	48	59.17	56.33	61.17
motorbike	73.74	72.17	58.67	66.33	57.17	66.67	55.67	62.67
person	59.85	56.33	52.33	59.17	52	52.33	51.33	55.33
pottedplant	61.79	61.67	56.17	59.33	53	55.17	52.5	56.5
sheep	74.85	74.67	64.5	72.83	50.17	67.67	61.5	67.17
sofa	68.96	65.67	62.17	63.83	52.17	56	52.5	63.5
train	80.1	78.83	68.5	70.5	49.5	65.17	57.33	63.33
tvmonitor	67.97	66.33	59.83	65.83	51.33	67	57.83	67.83
Average	72.16	69.53	61.93	65.69	51.64	61.36	56.13	63.32

- Fusion of all the features and ensemble classification leads to the average accuracy of 91.23%, which is better than accuracy of social only features with 2.62% margin.
- The results of social features and image features are quite close for “mountain”, “tattoo”. This is the result of a great degree of visual information present with these images, compared to social data.

TABLE 5.7: NUS: Precision Comparison

Labels	Ensemble Method	NUS Baseline Results	Social Features		Visual Features				
			Latent Semantic Indexing	Random Projection	HOG-LBP	SIFT	GIST	COLOR	GLCM
animal	94.09	83.46	90.01	70.44	73.46	51.62	66.21	57.18	64.11
coral	95.45	45.73	94.82	83.10	79.45	53.11	65.22	69.23	67.64
dancing	95.98	53.66	92.82	74.07	75.59	59.13	67.39	63.30	60.74
harbor	93.63	36.47	92.24	77.58	81.40	53.22	79.32	72.40	73.07
military	97.86	55.72	93.07	79.28	70.67	56.42	61.35	65.02	71.94
mountain	90.05	34.23	89.26	72.42	84.30	54.58	78.03	78.62	72.60
snow	92.32	42.43	87.54	66.07	69.11	56.66	60.82	64.92	67.05
statue	88.56	54.63	84.56	73.67	57.66	54.79	59.57	52.99	56.49
tattoo	88.92	56.24	86.81	71.95	83.43	58.64	70.74	71.13	75.49
temple	93.68	43.78	89.72	69.59	64.42	54.15	57.92	58.67	58.14
waterfall	98.58	57.38	95.88	82.59	86.45	54.65	76.68	76.56	83.82
wedding	94.22	72.34	89.98	76.31	66.43	54.44	56.91	61.50	60.27
AVERAGE	93.61	53.01	90.56	74.76	74.36	55.12	66.68	65.96	67.61

Table 5.8: NUS: Accuracy Comparison

Labels	Ensemble Method	Social Features		Visual Features				
		Latent Semantic Indexing	Random Projection	HOG-LBP	SIFT	GIST	COLOR	GLCM
animal	91.62	88.00	70.17	72.86	50.54	66.07	56.96	64.29
coral	95.99	92.33	80.83	79.63	52.59	65.56	68.33	67.41
dancing	92.39	92.00	74.33	75.58	58.85	67.88	62.88	60.96
harbor	93.02	90.83	75.83	82.96	53.15	80.37	72.59	73.15
military	94.69	90.83	77.00	69.81	55.74	61.11	64.26	70.93
mountain	90.55	86.67	72.67	85.00	53.33	78.33	76.85	71.85
snow	88.44	86.50	66.00	70.19	47.59	60.93	64.63	65.00
statue	85.23	84.67	72.50	58.15	54.07	59.26	52.59	55.93
tattoo	88.83	85.17	71.33	84.07	57.04	70.93	71.67	77.41
temple	91.50	87.67	68.83	65.00	53.52	58.15	58.52	57.96
waterfall	93.98	93.83	81.17	86.67	53.70	75.93	76.48	84.07
wedding	88.57	84.83	74.67	65.93	53.33	56.67	60.74	60.93
Average	91.23	88.61	73.78	74.65	53.62	66.76	65.54	67.49

Chapter 6

Conclusion and Future Work

In this work, we have demonstrated that if we use the social networking meta data related to images present on social networks like flicker, instagram, facebook, google plus etc. We achieve much better image classification results. We started with the five different visual features like SIFT, GIST, HoG, Color Histogram and GLCM. First, we classified our data set based on these individual features and then we took an ensemble model developed by weighted combination of these feature models. This analysis shows for image classification problem individual features are less effective. We need to combine these features for better classification.

We have further presented a simple and efficient approach of converting the social meta data of image in feature vectors. We used the text processing methods to contain the semantic structure of these social feature vectors in lower dimension. These social feature vectors gave better classification in most of the labels as compared to visual features. This phenomena can be explained by the fact that the social feature vector is actually constructed from social meta-data, which is very much similar to textual information about the image. The comments, the tags, the groups and the galleries of an image has the direct textual information of that image class. Therefore, correlating this textual information with the class of image is not so complex. Where as in case of visual features, we have do a lot of image processing to get features, which indicates some visual property like texture,

objects, color pattern and spatial envelope. But none of these property directly tell the exact label of the image, but in social meta data we might have the label as a tag or as a word in a comment.

Several machine learning technique were explored for analysis and classification including Spatial Pyramid, Bag of Words Model, Latent Semantic Indexing and LibSVM. The results of these techniques came out to be much better then the best results of the competitions held on these four datasets. We, therefore, can conclude that the social-meta data gives that extra information, which proves to be very instrumental in image classification. The fusion of social features and visual features enhances the accuracy of image classification.

6.1 Future Work

Some of the labels in our data-set does not had enough images like fox, lake, mountain, protest, earthquake etc. In such cases, when number of images available were less then 50, we discarded that label for classification because we did not have the enough positive instances. In such cases we can try the web-information retrieval methods to get more images for such labels.

We used the strategy of extracting the features, using dimensionality reduction techniques like Spatial Pyramid, Latent Semantic Indexing, Bag of Words etc and then using the classifying techniques like LibSVM on these low dimensional features. This problem can also be explored using Deep Belief Networks [**?**] with different number of hidden states, stacks of Restricted Boltzmann Machines and layers to get prominent reduced features; which can further be classified using supervised classification technique.

We have currently used MATLAB and python for implementation purpose. This implementation was done on signle thread with local system. We can change our implementation to faster executing language like C++ , with parallel processing for computationally difficult portion of feature extraction. This can increase the time efficiency of creating the model. We used the approach of weighted features, we could used some better fusion techniques to ensemble all the features.

Our social feature vector will be efficient only, if the social meta data of image already has enriched textual information in it. To overcome this constraint, we can explore the dimension of using the network structure of social meta data and co-relating images on the basis of the network. This technique can also be supervised to retrieve images from social networking sites on the basis of linkages and also we can do automated tag recommendations.

Bibliography

[1] Andrew Y.; Jordan Michael I Blei, David M.; Ng. "latent dirichlet allocation". in lafferty, john. journal of machine learning research. *Journal of Machine Learning Research*, 2003.

[2] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. In *AAAI*, 2011.

[3] Krystyna K. Matusiak. Towards user-centered indexing in digital image collections. *OCLC Systems & Services: International digital library perspectives*, 22(4): 283–298, 2006. doi: 10.1108/10650750610706998. URL `http://dx.doi.org/10.1108/10650750610706998`.

[4] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262 – 282, 2007. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2006.04.045. URL `http://www.sciencedirect.com/science/article/pii/S0031320306002184`.

[5] Julian J. McAuley and Jure Leskovec. Image labeling on a network: Using social-network metadata for image classification. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV (4)*, volume 7575 of *Lecture Notes in Computer Science*, pages 828–841. Springer, 2012. ISBN 978-3-642-33764-2. URL `http://dblp.uni-trier.de/db/conf/eccv/eccv2012-4.html#McAuleyL12`.

[6] Kraisak Kesorn. Multi-modal multi-semantic image retrieval, 2010. URL `https://qmro.qmul.ac.uk/xmlui/bitstream/handle/123456789/438/KESORNMulti-modal2010.pdf?sequence=1`.

[7] Sonal Gupta, Joohyun Kim, Kristen Grauman, and Raymond J. Mooney. Watch, listen & learn: Co-training on captioned images and videos. In Walter Daelemans, Bart Goethals, and Katharina Morik, editors, *ECML/PKDD (1)*, volume 5211 of *Lecture Notes in Computer Science*, pages 457–472. Springer, 2008. ISBN 978-3-540-87478-2. URL `http://dblp.uni-trier.de/db/conf/pkdd/pkdd2008-1.html#GuptaKGM08`.

[8] Besiki Stvilia, Corinne Jörgensen, and Shuheng Wu. Establishing the value of socially-created metadata to image indexing. *Library & Information Science Research*, 34(2): 99–109, April 2012. ISSN 07408188. doi: 10.1016/j.lisr.2011.07.011. URL `http://dx.doi.org/10.1016/j.lisr.2011.07.011`.

[9] Wikipedia. Folksonomy — wikipedia, the free encyclopedia, 2015. [Online; accessed 18-May-2015].

[10] Oded Nov, Mor Naaman, and Chen Ye. What drives content tagging: The case of photos on flickr. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1097–1100, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357225. URL `http://doi.acm.org/10.1145/1357054.1357225`.

[11] R. Kern, M. Granitzer, and V. Pammer. Extending folksonomies for image tagging. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on*, pages 126–129, May 2008. doi: 10.1109/WIAMIS.2008.43.

[12] Ed H. Chi and Todd Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08: Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88, New York,

NY, USA, 2008. ACM. ISBN 978-1-59593-985-2. doi: 10.1145/1379092.1379110. URL http://portal.acm.org/citation.cfm?id=1379110&coll=GUIDE&dl=GUIDE&CFID=37458772&CFTOKEN=13998061&ret=1.

[13] Nuo Zhang and Toshinori Watanabe. Text-transformed image classification based on data compression, 2013. URL http://dx.doi.org/10.1007/978-3-642-28807-4_51.

[14] Dhruv Mahajan and Malcolm Slaney. Image classification using the web graph. ACM Multimedia, 2010. URL http://research.microsoft.com/apps/pubs/default.aspx?id=192157.

[15] Leandro Augusto da Silva. Image classification combining visual features and text data: neural approach and based on swarms. 2013. URL http://www.bv.fapesp.br/25294.

[16] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 759–766, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273592. URL http://doi.acm.org/10.1145/1273496.1273592.

[17] Adam Rae, Börkur Sigurbjörnsson, and Roelof van Zwol. Improving tag recommendation using social networks. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, RIAO '10, pages 92–99, Paris, France, France, 2010. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. URL http://dl.acm.org/citation.cfm?id=1937055.1937077.

[18] Roelof van Zwol, Adam Rae, and Lluis Garcia Pueyo. Prediction of favourite photos using social, visual, and textual signals. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1015–1018, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874138. URL http://doi.acm.org/10.1145/1873951.1874138.

[19] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 807–816, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557108. URL `http://doi.acm.org/10.1145/1557019.1557108`.

[20] Matthew Boutell and Jiebo Luo. Beyond pixels: Exploiting camera metadata for photo classification. *Pattern Recognition*, 38(6):935 – 946, 2005. ISSN 0031-3203. doi: http://dx.doi.org/10.1016/j.patcog.2004.11.013. URL `http://www.sciencedirect.com/science/article/pii/S0031320304003978`. Image Understanding for Photographs.

[21] Brian T. Bartell, Garrison W. Cottrell, and Richard K. Belew. Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '94, pages 173–181, New York, NY, USA, 1994. Springer-Verlag New York, Inc. ISBN 0-387-19889-X. URL `http://dl.acm.org/citation.cfm?id=188490.188554`.

[22] Ashish Bindra. Sociallda:scalable topic modeling in social networks. Master's thesis, University of Washington, 2012.

[23] Jure Leskovec Julian McAuley. Image labeling on a network: Using social-network metadata for image classification. *Computer Vision – ECCV*, 2012.

[24] Van Gool L. Williams C. Winn J. Zisserman A. Everingham, M. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[25] Lew Huiskes, M. The mir flickr retrieval evaluation. *CIVR*, 2008.

[26] Huiskes M.: Nowak, S. New strategies for image annotation: Overview of the photo annotation task at imageclef 2010. *CLEF (Notebook Papers/LABs/Workshops)*, 2010.

[27] Tang J. Hong R. Li H. Luo Z. Zheng Y.T. Chua, T.S. Nus-wide: A realworld web image database from the national university of singapore. *CIVR*, 2009.

[28] David G Lowe. Object recognition from local scale-invariant features. In *The proceedings of the seventh IEEE international conference on Computer Vision.*

[29] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005*, 2005.

[30] Cordelia Schmid Svetlana Lazebnik and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.

[31] A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008.

[32] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of Computer Vision*, 2001.

[33] H.G. Barrow and J.M. Tannenbaum. Recovering intrinsec scene characteristics from images. *A. Hanson and E. Riseman (Eds.)*, 1978.

[34] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of gist descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 19:1–19:8, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-480-5. doi: 10.1145/1646396.1646421. URL `http://doi.acm.org/10.1145/1646396.1646421`.

[35] Jurij F Tasic Marko Tkalcic. Colour spaces perceptual, historical and applicational background. *Eurocon*, 2003.

[36] Robert M Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 1979.

[37] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision*

and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2372-2. doi: 10.1109/CVPR.2005.177. URL `http://dx.doi.org/10.1109/CVPR.2005.177`.

[38] P.A. Torrione, K.D. Morton, R. Sakaguchi, and L.M. Collins. Histograms of oriented gradients for landmine detection in ground-penetrating radar data. *Geoscience and Remote Sensing, IEEE Transactions on*, 52(3):1539–1550, March 2014. ISSN 0196-2892. doi: 10.1109/TGRS.2013.2252016.

[39] Yinan Yu, Junge Zhang, Yongzhen Huang, Shuai Zheng, Weiqiang Ren, and Chong Wang. Object Detection by Context and Boosted HOG-LBP. 2010. URL `http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2010/workshop/nlpr.pdf`.

[40] Junge Zhang, Kaiqi Huang, Yinan Yu, and Tieniu Tan. Boosted local structured hog-lbp for object localization. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1393–1400, June 2011. doi: 10.1109/CVPR.2011.5995678.

[41] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision amp; Image Processing., Proceedings of the 12th IAPR International Conference on*, volume 1, pages 582–585 vol.1, Oct 1994. doi: 10.1109/ICPR.1994.576366.

[42] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32–39. IEEE, 2009.

[43] Modesto Castrillón-Santana, Javier Lorenzo-Navarro, and Enrique Ramón-Balmaseda. Improving gender classification accuracy in the wild. In José Ruiz-Shulcloper and Gabriella Sanniti di Baja, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, volume 8259 of *Lecture Notes*

in Computer Science, pages 270–277. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-41826-6. doi: 10.1007/978-3-642-41827-3_34. URL `http://dx.doi.org/10.1007/978-3-642-41827-3_34`.

[44] et al Deerwester, S. Improving information retrieval with latent semantic indexing. *Proceedings of the 51st Annual Meeting of the American Society for Information Science 25,*, 1988.

[45] et al Deerwester, S. Best practices commentary on the use of search and information retrieval methods in e-discovery. In *the Sedona Conference*, 2007.

[46] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[47] Radim Rehurek and Petr Sojka. Gensim:software framework for topic modelling with large corpora. *Natural Language Processing Laboratory Masaryk University, Faculty of Informatics*, 2010.

[48] Pietro. Li, Fei-Fei; Perona. A bayesian hierarchical model for learning natural scene categories. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[49] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data.

[50] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, pages 245–250, New York, NY, USA, 2001. ACM. ISBN 1-58113-391-X. doi: 10.1145/502512.502546. URL `http://doi.acm.org/10.1145/502512.502546`.

[51] WilliamB. Johnson, Joram Lindenstrauss, and Gideon Schechtman. Extensions of lipschitz maps into banach spaces. *Israel Journal of Mathematics*, 54(2):129–138,

1986. ISSN 0021-2172. doi: 10.1007/BF02764938. URL http://dx.doi.org/10.1007/BF02764938.

[52] S. Dasgupta. Experiments with random projection. *In Proceeding on Uncertainty in Artificial Intelligence*, 2000.

[53] Thomee B. Lew M. Huiskes, M. New trends and ideas in visual concept detection:the mir flickr retrieval evaluationinitiative. *CIVR*, 2010.

[54] Chang, Chih-Chung and Lin, Chih-Jen. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[55] Siddharth Agarwal. Genre and style tagging of paintings. Master's thesis, IIT Kanpur, 2014. Unpublished Manuscript.

[56] D. Achlioptas. Database-friendly random projections. *In Proceeding of ACM Symposium on the Principles of Database Systems*, 2001.

[57] Shu-Yuan Chen Ya-Chun Cheng. Image classification using color, texture and regions. *Image and Vision Computing*, 2003.

[58] D.G. Lowe. Method and apparatus for identifying scale invariant features in an image and use of same for locating an object in an image, March 23 2004. URL http://www.google.com/patents/US6711293. US Patent 6,711,293.

[59] Tony Lindeberg. A computational theory of visual receptive fields. *Biological Cybernetics*, 107(6):589–635, 2013. ISSN 0340-1200. doi: 10.1007/s00422-013-0569-z. URL http://dx.doi.org/10.1007/s00422-013-0569-z.

[60] Todd A Letsche and Michael W Berry. Large-scale information retrieval with latent semantic indexing. *Information sciences*, 100(1):105–137, 1997.

[61] Pedro R Kalva, Fabricio Enembreck, and Alessandro L Koerich. Web image classification based on the fusion of image and text classifiers. In *Document Analysis and*

Recognition, 2007. ICDAR 2007. Ninth International Conference on, volume 1, pages 561–568. IEEE, 2007.

[62] Gang Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1367–1374, June 2009. doi: 10.1109/CVPR.2009.5206816.

[63] Khaled A.F. Mohamed. The impact of metadata in web resources discovering. *Online Information Review*, 30(2):155–167, 2006. doi: 10.1108/14684520610659184. URL `http://dx.doi.org/10.1108/14684520610659184`.

[64] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA, 1998. ACM. ISBN 1-58113-057-0. doi: 10.1145/279943.279962. URL `http://doi.acm.org/10.1145/279943.279962`.

[65] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 902 – 909, jun 2010. URL `http://lear.inrialpes.fr/pubs/2010/GVS10`.