# 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**

July 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 of online multimedia information by stating just how the meteoric growth of data has been in recent decades. While this may be a very mundane way to introduce the need for annotated online digital corpora and the quantum of effort required, it is nonetheless, true for work in this field.

The proliferation of personalized gadgets, cheap digital cameras and diversification away from single use devices like old generation mobile phones to new generation smart phones, tablets and wearable devices has spawned a culture of documenting and saving every moment of our lives in digital form. Internet services like Facebook, Flickr, Instagram etc also make it very easy for anyone to share their pictures with their (online) social connections.

In the present scenario, where we have several robust systems effectively handling 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 leave some clues about the content in question.

This social meta-data, the data related to the content, can be a pathway to organize the whole multimedia corpus through a labelling exercise using classification, theming and ontology creation and use. This meta-data can help us in many ways for example in information retrieval, multimedia classification, heterogeneous learning etc. In this thesis, we focus on images and their social meta data.

Photos unlike 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 the past, photographs were normally confined to an album and were rarely opened to relive old memories. The photos were difficult to share and normally did not contain any contextual information. Whereas today people can upload images to the world wide web directly from capture devices (like phones, 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 time, the social aspect of the image becomes more prevalent. This creates a need for expanding our understanding and power of utilizing this social data.

This thesis focuses on exploring the role which can be played by social context information in enhancing 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 social context information to follow a multimodal approach to 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 and in short supply also

labeling new images is costly and requires significant human input. Second, 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.

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 concept (polysemy). In such cases, we need some help from extra data to gracefully handle such ambiguous conditions in order to get high classification accuracy.

Text and images are two distinct types of information resource and belong to two different modalities as they present an object in different ways. However, there are some implicit connections and invariant properties shared between images and textual information. In fact, the texts associated with images include some form of human generated description of images; thus these can be regarded as a supplement to the image content.

The combination of textual information with image features can be a way of improving image classification results. With all these cues about heterogeneous learning, we wish to take the help of data in different modalities to have better classification of images. The unprecedented evolution of social networking gives us such data with reasonable ease. People talk about images, put likes, comments, tags and other markers. These markers or rather social meta-data can be instrumental in inferring the content of images.

In the light of all these motivating facts, we decided to pursue social meta data as an additional soure of information for enhancing image classification.

## 1.2 Problem Statement

In this thesis, we will focus on classification of images, which have some social meta-data available with them. The social meta data for an instance can be tags on images, groups in which image is featured or comments on the image etc.

If we try to break an image classification problem for multi-labeled images, we can talk in terms of labeling an image with different labels. We will be using supervised learning techniques to do this, which means we will first develop a classification system by creating classification models using already labeled images and then use these models to automatically label the unknown images with the label.

Annotating an image with a labels can be considered as a multi- class classification problem where the classes are represented by labels. For example an image with a ship floating in the sea, can be labeled as "sea", "ship" or "water". So, the problem of annotating an image with its labels is equivalent to classifying that image into one of the labels. We can define a problem as follows:

Suppose we have an image data set consisting of N images $X = \{x_1...x_N\}$ and a label space consisting of L categories $L = \{-1, 1\}^L$.

We can denote the ground truth labeling for the image $x_n$ as $y^n \in L$. Then the ground truth for a particular category $c$ can be denoted as $y^n : y^n \in \{-1, 1\}$. When we combine the ground truth for the entire data set for category c, we get $Y_c \in \{-1, 1\}^N$.

We try to learn a prediction made for an image $x_n$ and category $c$, which will be $\bar{y_c}(x_n, \theta_c) \in \{-1, 1\}$. Predictions across the entire data set for category $c$ will be $\bar{Y_c}(X, \theta_c) \in \{-1, 1\}^N$.

We propose methods to enhance image classification using social meta-data. If we only have images and no auxiliary data , we will get only the visual features $\theta_c^{Visual} \in R^{Visual\ features}$ for an image. We, therefore, first consider some widely used visual features (SIFT Features, GIST Features, COLOR Space Features, Texture/GLCM Features and HOG-LBP Features) to learn a classification model on these images.

We, then, test our visual-only classification model, to see how accurate it is in classifying the images. Now, we consider the features generated by the social features $\theta_c^{Social} \in R^{Social\ features}$ for an image to learn a classification model on these images. We, then, compare our results of social-only classification model to visual-only classification model. In the third step,

In third step, we ensemble these visual only and social only classifiers to obtain the final result. We compare all the results to find the improvement obtained by using some auxiliary social- meta data.

## 1.3 Outline of the Thesis

The remaining thesis is structured as follows:

- **Chapter 2** reviews the related work done in utilizing the meta-data and other information related to images in order to enhance image classification.
- **Chapter 3** gives an overview of the data-sets, we considered in our thesis.
- **Chapter 4** describes the feature extraction, which is quite important in our experiments.
- **Chapter 5** presents the experiments we did and the results we got.
- **Chapter 6** contains conclusions and pointers to extensions and future work.

# Chapter 2

# Background and Related Work

The world wide web has huge amounts of image data, mostly unlabelled, and separately a lot of meta data on images in terms of comments, tags, data on individuals etc. from social media. The obvious question is whether and how the meta data can be used to give meaningful labels to the images. Note that this is not a simple classification problem in the usual sense. It is clearly a multi-class and a multi-label problem at the same time but more importantly there is no apriori fixed set of labels. The label set can grow and shrink based on the corpus of images that we are considering. Also, images in the corpus may have no labels or can be partially labelled and there is the added problem of deciding whether the labelling process for a particular image with respect to the current set of labels is complete or not.

The important concept, we want to coin here is heterogeneous learning. Kesorn [2010] has shown that although, text and visual are distinct types of representations and modalities there are some strong implicit connections between images and any accompanying text information. They used multimodal cues (visual features and text captions) for retrieving images, which depict semantically similar concepts. For example, given a large corpora of images and associated texts, finding the images which have been taken in a sports event. Sincei, in such cases discovering the semantics of an image is an extremely challenging problem, they have used any associated textual information that accompanied the image

as a cue to predict the meaning of an image. By transforming this textual information into a structured annotation to enhance visual content it can be used by image retrieval systems to retrieve selected images more precisely. This is one form of heterogeneous learning.

Second, they used Latent Semantic Indexing to create a domain- specific ontology-based knowledge model. The ontologies describe visual content using well-structured concepts and relationships that are also human readable and meaningful. The ontology of a certain domain is about its terminology (domain vocabulary), all essential concepts in the domain, their classification, their taxonomy, their relations (including all important hierarchies and constraints) and domain axioms. The use of the ontology-based knowledge model allows the system to find indirectly relevant concepts in image captions and thus leverage these to represent the semantics of images at a higher level. This also enabled the framework to tolerate ambiguities and variations (incompleteness) of meta data. They designed ontologies for some specific domains like e.g. sports. They collected data from different sports organization websites and then used a designed sport taxonomy to convert the data to an ontology model.

**?** has solved the problem of learning robust models out of scenes and actions from partially labeled collections. They proposed that visual cues are generally too ambiguous to recognize the visual scenes or activities. Obtaining manually labeled examples from which a robust model can be learned is also impractical. They proposed leveraging the text accompanying visual data to cope with these constraints. To classify images, their method learns from captioned images of natural scenes. For actions, they used videos of athletic events with commentary. They concluded that exploiting the multi-modal representation and unlabeled data provides more accurate image and video classification compared to base-line algorithms. They also asserted that this extra data and multi-modal representation can be the basis of a solution to the problem of managing the world's ever growing multi-media data of digital images and videos.

McAuley and Leskovec [2012] proposed that we can use the inter-dependencies of images 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 use in the thesis. They modeled their task as a binary labeling problem on a network and used max-margin SVM training for simultaneous binary predictions over the entire data-set. They studied the use of social meta data for three binary classification problems: predicting image labels, tags, and groups. They analyzed the relative importance of social features (e.g. shared membership in a gallery, relational features based on shared location, shared group etc.) in image labeling. They first learned flat models using single indicators like tags only, groups only or gallery only to learn the labels. In the second step, they created a graphical model of shared properties, in this graphical model two images contained in the same gallery have a high probability of similar label. In terms of graphical models, this means that they formed a clique from photos sharing common meta data.

It has been asserted that collaborative tagging, social classification, social tagging, which is also called "Folksonomy" Wikipedia [2015], holds the key to developing a semantic web, in which every web page contains machine readable meta-data that describes its content. A folksonomy is a system of classification derived from the practice and method of collaboratively creating and translating tags to annotate and categorize content. IMohamed [2006] showed the impact of meta-data for retrieving information from websites.This study focuses on indexing web pages using metadata and its impact on search engine's rankings.

Kern, Granitzer, and Pammer [2008] has experimented with Folksonomy. They used some collaboratively created sets of meta-data, to organize multimedia information available on the Web. They addressed the question of how to extend a classical folksonomy with additional metadata. They have also shown that it can be applied for tag recommendation. A Folksonomy is called plain, if it contains only one kind of information, for example it can only be collaboratively given photo tags or it can only be favorite photos selected by users. In extended folksonomy, you find connection between two type of folksonomies. In

their paper, Kern, Granitzer, and Pammer [2008] has used a similarity graph built from the graph created from selected data. For this they used 12 different type of folksonomical data about photos including groups, photo tags, photo favorites, testimonials, comments etc. Their study was on Flickr images related to Group "Fruit & Veg".

Chi and Mytkowicz [2008] have systematically analyzed the efficiency of social tagging systems using information theory. They tried to find an answer to the efficacy of a naturally evolved user generated vocabulary in identifying objects (images, videos, documents etc.). They have shown that information theory provides an interesting way to understand the descriptive power of tags. Their results show that information theory gives evidence that social tags can be used to identify objects. Their experiment was to find the efficacy of a user generated vocabulary in identifying documents. They collected bookmarking data from Delicious (formerly del.icio.us). Delicious is a social bookmarking web service for storing, sharing, and discovering web bookmarks. They started at the del.icio.us homepage and harvested a set of users. For each user, they collected their bookmarks, as well as links to other users that bookmarked the same document. In their data set, the ratio of unique documents to unique tags was almost 84. Given this multiplicity of tags to documents, they tried to answer the question: How effective are tags at isolating a single document?

Liu, Zhang, Lu, and Ma [2007] have given a survey of the recent technical achievements in order to improve the retrieval accuracy of content-based image retrieval systems. Their survey shows that in recent times research focus has shifted from designing sophisticated low-level feature extraction algorithms to reducing the 'semantic gap' between the visual features and the richness of human semantics. They have discussed fusing the data from HTML text and visual content of images for World-Wide-Web retrieval. They suggest that this technique can be used to narrow down the 'semantic gap'. They observe that the Web page containing an image generally has some additional information available, which can facilitate semantic-based image retrieval. For example, the URL of the image file often has a clear hierarchical structure including some information about the image such as image category. In addition, the HTML document also contains some useful information

in image title, tags, the descriptive text surrounding the image, hyperlinks, etc. They have used the evidence from both the HTML text and visual features of images and developed two independent classifiers based on text and visual image features, respectively. The experimental results using a pre-defined set of 15 concepts demonstrates a substantial performance improvement.

**?** have shown how labeled text from the web helps image classification. In this paper, they have investigated the interplay between multimedia data mining and text data mining. They address 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 Naive Bayes Classification models. They have proposed that for a targeted domain classification problem, some extra annotated images can be found on many social Web sites, which can serve as a bridge to transfer knowledge from the abundant text documents available on the Web. By using latent semantic features generated by the auxiliary data, they were able to build a better integrated image classifier.

Mahajan and Slaney [2010] combined the information from social graphs 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 give a large improvement over content features alone in an experiment where they tried to classify the images of adults among all the images. They have exploited the link structure of the web graph. A web page related to a given category is normally linked to other pages describing related objects. They combine information from the webgraph structure with semi-supervised learning from all the unlabeled images to create a superior image-classification model for multimedia data. They show that fusing image, text and web- graph features gives a 12% improvement (in the area under the ROC curve) over content features alone in an adult image-classification experiment.

mmodel have considered a scenario where keywords are associated with the training images, e.g. as found on photo sharing websites. They demonstrated a semi-supervised multi-modal learning algorithm for image classification. They have shown 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 MIR Flickr data-set for their experiments. They utilized Flickr tags as the aiding information in the process.

They used images, which were annotated for 24 concepts, including object categories but also more general scene elements such as sky, water or indoor. They choose tags for these images which were appearing in at least some percentage of images, resulting in a vocabulary of tags. They used a binary vector to encode the absence or presence of each of the different tags in this fixed vocabulary in a linear kernel, which counts the number of tags shared between two images. They also extracted several different visual descriptors. After that they averaged the distances between images based on these different descriptors, and used it to compute an RBF kernel. They first used visual features and then used the textual one for image classification. They observed that for many classes in both data sets the visual classifier is stronger than the textual one, yielding a 10 % higher MAP(mean average precision) score, where as the combined classifier significantly improved the classification results, the MAP score increased by more than 13 %.

van Zwol, Rae, and Garcia Pueyo [2010] handled the problem of predicting 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. For the social signals, they used attributes of the users like his current favorite list, his friends, the galleries he/she followed etc. For textual signals, they used tags and comments on the image. They used gradient-boosted decision trees (GBDT) for the classification of a user's favorite photos. For the evaluation of performance 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 and social features helps to significantly boost the recall, with a small decrease in precision.

Boutell and Luo [2005] has shown, how we can leverage the camera meta data to provide

evidence independent of the captured scene content. They used this meta data to improve classification performance. They proposed that the EXIF specification for camera metadata (used for JPEG images) includes hundreds of tags. Among these, some relate to picture taking conditions (e.g., FlashUsed, FocalLength, ExposureTime, Aperture, FNumber, ShutterSpeed, and Subject Distance). They said that some of these cues can help distinguish various classes of scenes. For example, flash tends to be used more frequently on indoor images than on outdoor images. They broke this meta-data in families of meta data tags: Scene Brightness, Flash, Subject Distance. and Focal Length. They introduced Bayesian networks based probabilistic scheme to fuse low-level image content cues and camera meta data cues for improved scene classification. Their results demonstrate that this integration of camera meta data increases the efficacy of classification. They used this technique for some very specific problems Indoor- Outdoor Classification, Sunset Scene Detection and man-made-Natural Scene Classification.

Bindra [2012] have shown a 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 user interests for target recommendation. They called it SocialLDA. They propose a LDA model by taking into account the social connections among users in the network. This model was used to categorize a user's incoming document stream as well as finding user interest based on the user's authored document. This is primarily text classification but it has novel use of LDA and social meta data for classification, which is quite similar to our method of using social meta-data.

# Chapter 3

# Data Set

*The social network meta data related to images can be used for image classification tasks and some times can outperform image content based methods.* For the validation of this hypothesis, we needed a well labeled image data set, which should also be amenable to expansion along social dimensions. The data must have some ground truth provided by human annotators or by some other standard way of labeling.

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

- The PASCAL Visual Object Challenge ('PASCAL') **?**
- The MIR Flickr Retrieval Evaluation ('MIR') Huiskes [2008]
- The ImageCLEF Annotation Task ('CLEF') Nowak [2010]
- The NUS Web Image Database ('NUS') Chua [2009]

The creators of these data sets had obtained labels through crowd-sourcing from the Flickr users or communities. The labels ranged from object based categories like *person* or *bicycle*, to subjective concepts like *Aesthetic_ Impression*. These labels satisfied the desired ground truth constraint for our classification process. We, therefore, used 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. Flickr sources were available only for training images, and for the test images from 2007. There were a total of 11,197 images, for which Flickr sources were available.

The MIR Flickr Retrieval Evaluation ('MIR') 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') uses a subset of 18,000 images from the MIR data set, but it has more varied tagging as annotation. There were a total of 4,807 images, for which Flickr sources were available.

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

## 3.2 Augmentation of Social Meta-data

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

## 3.3 Preliminary Observation of Data set

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

FIGURE 3.1: Image MIR Examples

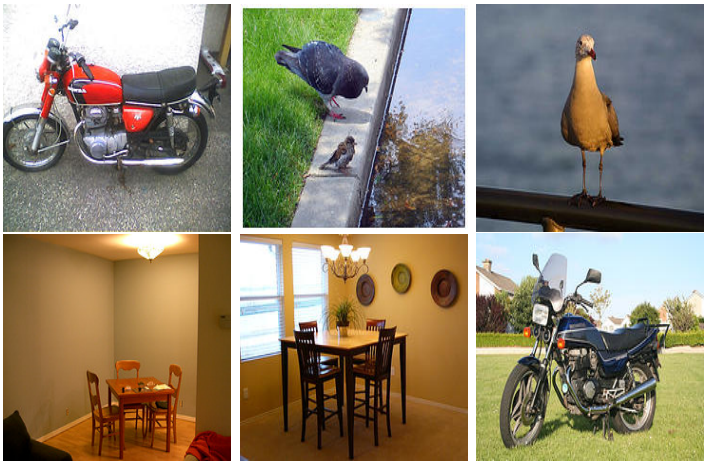

FIGURE 3.2: Image PASCAL Examples

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 had the least tags and comments compared to all other data-sets, because it was made up of less interesting images. The NUS data set favored the highly popular images and we saw that it has the highest tag vs image ratio of 19.4. The images were highly tagged, had large number of comments and were submitted to many groups. The MIR had 17+ tags and comments showing that it also had some what interesting images.

FIGURE 3.3: Image CLEF Examples

FIGURE 3.4: Image NUS Examples

## 3.4 Label Selection for Classification Problem

Due to constraints of less number of images for some particular labels and balanced learning, we have to selectively choose the labels for our Classification Problem.

For example, CLEF has 99 labels and some labels have number of images less than even 17. Doing learning and testing on such a small set was not useful and would not have given good results. We, therefore, in case of CLEF selected 20 Labels which had sufficient data.

Similarly for NUS, because of some computational constraints and data availability, we reduce our computation to 12 labels. The following table shows the labels, which were considered for the whole classification problem.

TABLE 3.1: Labels

| Data set | Count 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 |

Following the ideas give in Julian McAuley [2012], we decided to construct the whole problem of labeling an image with multiple label as taking a single label and then decide, if that image should or should not be given that label.

We, therefore, supposed to learn a prediction for an image $x_n$ and label $l$, which will be $\bar{y}_l(x_n, \theta_l) \in \{-1, 1\}$. $\theta^l$ are the descriptors/features extracted for label $l$. The above strategy transformed the whole problem in multiple binary classification problem. We followed the following strategy:

Given an image data set consisting of N images $X = \{x_1 ... x_N\}$ and a label space consisting of L categories $L = \{-1, 1\}^L$. We take single label $l$, when we combine the ground truth for the entire data set for label $l$, we get $y_l^n \in \{-1, 1\}^N$. Now we learn a classifier, which predicts for an image $x_n$ as $\bar{y}_l(x_n, \theta^l) \in \{-1, 1\}$. $\theta^l$ are descriptors/features extracted for label $l$.

This gave us good learning of features for each label and also precise information retrieval against a label, because if we use such classifier for a label $l$, we can easily retrieve images for that label by predicting $\bar{y}_l(x, \theta^l) \in \{-1, 1\}$ for set of images.

It also gave a leverage of easily using the bag of visual/non-visual words. Bagging of visual/non-visual words means we tried to learn some small set of descriptors, called as words, which represent a category.

## 3.5 Comparison of Results

In next two chapter, we would be first using the visual descriptors to classify images for each label. After that, we would be using social meta-data for image classification. We would be comparing these two results to figure out, which works out best among these two. These all tasks will be done for each label individually. In subsequent part, we will be doing ensemble of these classifiers to leverage all the descriptors. We will also show comparisons with published results on each of these four benchmarks or with results in associated competitions. All these benchmarks has their own competition and also various authors used these benchmarks in their paper, which we will be using.. In Huiskes [2010], Nowak [2010] ,**?** and Chua [2009], authors have either surveyed the various results in competition or have depicted a way of using visual + social data as classification basis for

MIR, ImageCLEF, PASCAL and NUS respectively. We will consider results from these papers. We will also be giving qualitative and quantitative conclusions/observations of our results.

The goal of all these comparisons would be to assess the improvement that was obtained by using social meta-data for images. We reported the mean average precision (MAP) for the sake of comparison with published materials and competition results. We also gave the accuracy for the binary prediction/classification of labels.

# Chapter 4

# Feature Extraction

Feature extraction is a special form of dimensionality reduction. The specialty of such dimensionality reduction is that, we do not lose important information of the data. Actual image or text data is generally too huge to be processed. We, therefore, need to extract useful information from this huge data. The constraint of this process is that we should not loose the information, which would help us in achieving our goal. The goal can vary from image classification to topic finding. The reduced set of this goal specific instrumental information is called as the set of features. The process of computing these features is called *feature vector extraction.*

Our data set was multi-modal, we had images, which contained visual form of data and meta-data(social and textual) of image, which was mostly textual. We, therefore, broke the process of feature extraction in two following parts:

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

In the following sections, we have described the features, we extracted and respective methods used for extracting those features.

## 4.1 Image Content Based Feature Extraction

Feature extraction is an important step in image processing. The performance of a classifier heavily depends on the feature vector used. Several kinds of features have been proposed in image processing. Some commonly used features for image classification are Color Histogram, HoG, LBP, SIFT, SURF, GLCM etc. With these commonly used features also, the problem of high dimensionality is faced. We, therefore, took the help of several dimesionality reduction techniques. Some important techniques, which can be named here are, Principal Component Analysis(PCA), Bag-of-words etc. In our work, we used the following image features:

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

In case of HoG-LBP and SIFT features, we used PCA and bag-of-words model for dimesnionality reduction. In the remaining part of this section, we describe these features and the methods 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 not just invariant to scaling, it is also consistent with translation, rotation and to some extent also remains unaffected by (some) variations of illumination, 3D projection and other viewing conditions. The SIFT is normally bundled with a feature detector and a feature descriptor. The detector extracts

FIGURE 4.1: Example of SIFT Descriptors

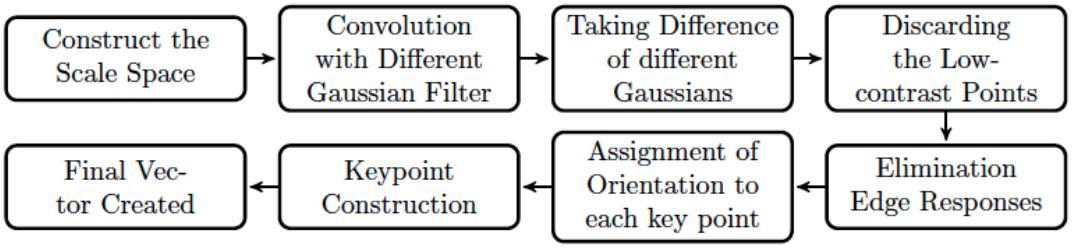


FIGURE 4.2: Process Flow of SIFT Descriptor

attributed regions in such a way, that the description of these regions (descriptors) is consistent with all the aforementioned changes (illumination, view point etc). The descriptor associated with the region is can be assumed as a signature, which recognizes the appearances efficiently and accurately. For example, some SIFT descriptors are shown in Figure 4.1.

SIFT features are very useful for finding objects or recognizing scenes in an image. SIFT descriptors were first introduced by Lowe in ICCV1999. They took the idea from the vision process in primates. 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 are mentioned in Figure 4.2.

SIFT features have been proven very useful in objectives like natural scene recognition Fei-Fei and Perona [2005]. These features are also very useful in object recognition, this was shown in Lowe. Motivated by the excellent performance of SIFT Features in Image Categorization, we chose SIFT as one of visual features, we were going to use for image classification purpose.

Fei-Fei and Perona [2005] has been shown that dense local scale-variant features performs better compared to sparse features. We. therefore, extracted dense SIFT features of $16 \times 16$ 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 the VLFEAT Library for finding the SIFT descriptors. Once, we had obtained the SIFT descriptors for the full set of images, we constructed a visual vocabulary of 400 visual words as described in Svetlana Lazebnik and Ponce. Svetlana Lazebnik and Ponce introduced the concept of visual bag of words, because this methodology overcomes the problem of high-dimensionality of SIFT features, quite intelligently. Using these bag of words, we could constraint the definition of a visual concept in these 400 visual words, which also provided efficiency and robustness. We used k-means from VLFEAT library Vedaldi and Fulkerson. [2008] for creating 400 clusters from the full training data. In figure 4.3, we have given an example of using SIFT Bag of Words features to match two natural scenes. In these images, we can see the SIFT descriptors with green circles. We have also shown the connection (an edge with blue line) matching these SIFT descriptors on both the images. We can see that SIFT descriptors calculated on image 1 also exist in image 2, hence these two images matches.

Svetlana Lazebnik and Ponce extended the normal BoW (Bag of Words model) by introducing the Spatial Pyramid technique. In this technique, we create a spatial pyramid of

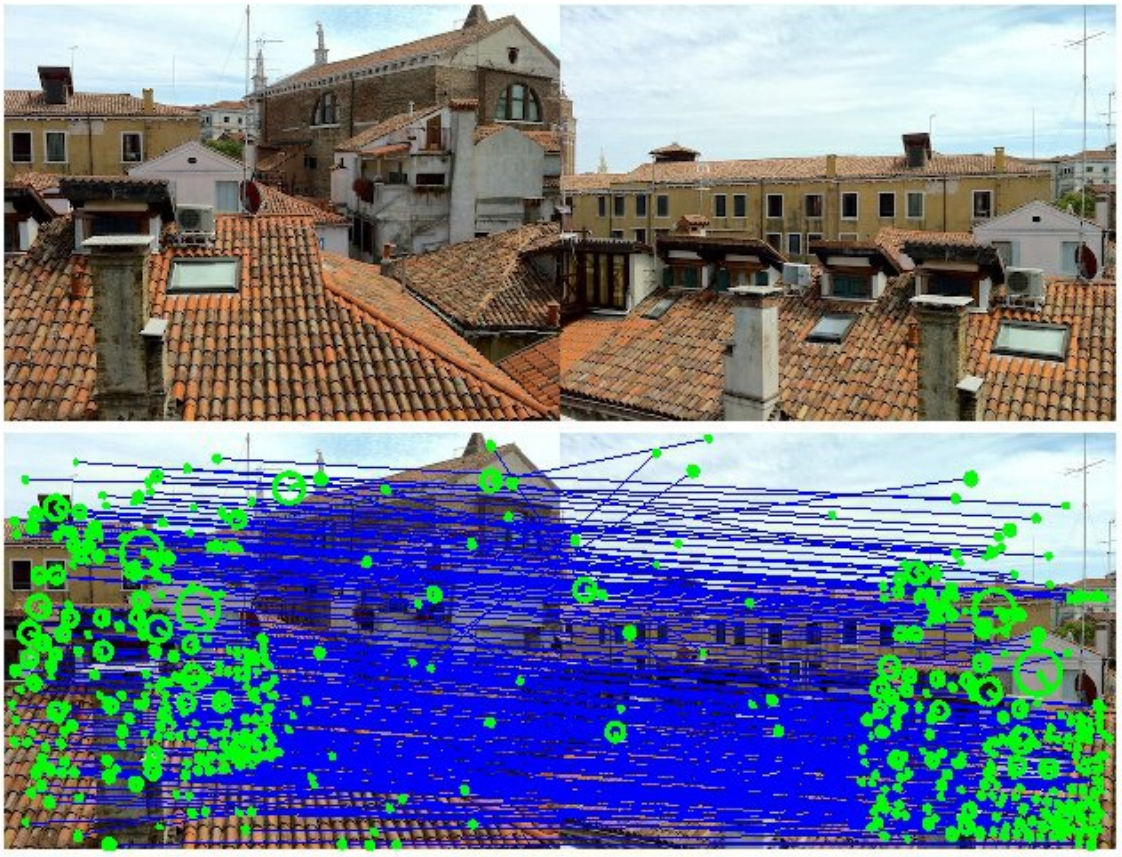

FIGURE 4.3: Use of SIFT Descriptor in matching

an image by dividing the image in hierarchical spatial layers. Each division provides a finer sub-region and more spatially 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 gives improved results on challenging scene recognition tasks as shown by Svetlana Lazebnik and Ponce.

We also used the Spatial Pyramid technique as defined in Svetlana Lazebnik and Ponce. We did 2-level Spatial partitioning. In level 0, we had the full image, this gave a 400 dimensional vector (the size of bag of visual words). In level 1, we partitioned the image in 4 sub regions, This gave $4 \times 400$ dimensional vector. In level 2, we again partitioned each sub-region into 4 sub-regions, so we obtained $4 \times 4 = 16$ sub-regions.This gave us a $16 \times 400$ dimensional descriptor. In this way, we have an image descriptor of size 8400.

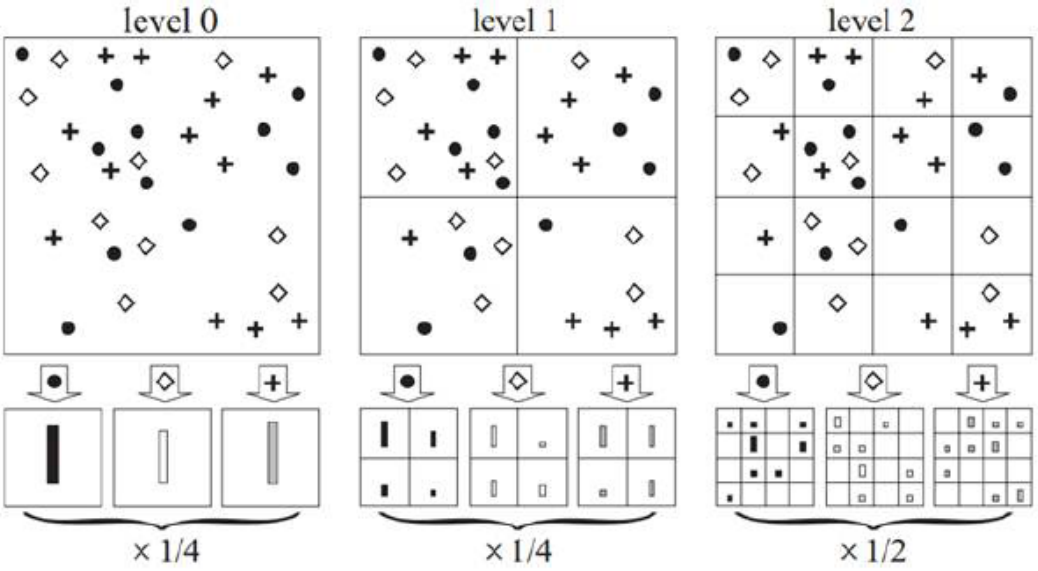


FIGURE 4.4: Computation of Spatial Pyramid over an image

We used the MATLAB implementation developed by Svetlana Lazebnik and Ponce for computing the pyramid on the SIFT descriptors.

A schematic diagram of calculating Spatial Pyramid 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 a dictionary of three visual words represented by plus, circle 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 Svetlana Lazebnik and Ponce.

### 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 Oliva and Torralba. [2001] in 2001. They took the reference of paper Barrow and Tannenbaum [1978] which describes vision in humans.

In case of scene recognition, human system actually does 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. Oliva and Torralba. [2001] 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 dimensions can be used as a way of recognizing a picture without going into the tiresome process of segmentation and processing individual regions/objects. This low -dimensional representation is termed as "Spatial Envelope" in Oliva and Torralba. [2001]. These Spatial Envelope Perceptual Dimension Descriptors can be reliably computed using spectral and coarsely localized information.

The model based on this spatial envelope generates a multidimensional space. In this projected space, scenes with semantically close 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 defined as the GIST of the scene.

Douze, Jégou, Sandhawalia, Amsaleg, and Schmid [2009] have shown that the GIST descriptor is very efficient and useful for web-scale search systems for images. This indicates that GIST can also be an efficient feature for image classification. We, therefore, include GIST in our visual feature list.

We used the GIST implementation available at **?** given by Oliva and Torralba. [2001]. We first decomposed the image using filters of 8 orientations for each of the 4 scales mentioned in Oliva and Torralba. [2001]. We got 32 oriented filters by this way.Then, the image was represented as a $4 \times 4$ matrix. Output values of all filters were normalized to $4 \times 4$ matrix. Then, the image was represented by the weighted combination of all these values giving $8(\text{orientation}) \times 4(\text{scales}) \times 4 \times 4(\text{size of matrix}) = 512$ dimensional vector.

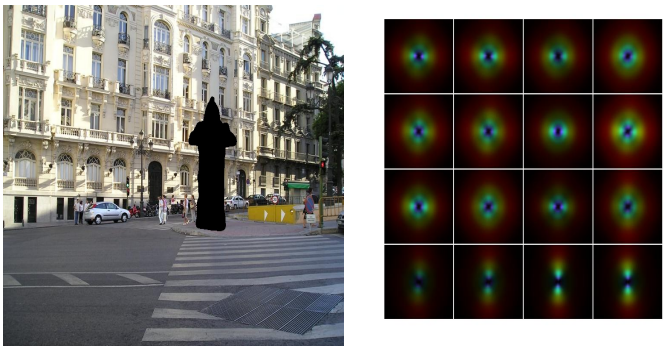

FIGURE 4.5: Example of GIST descriptors

GIST can be easily handle large databases because it is memory efficient and also computationally efficient. It categorizes 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 on the right is the 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 a way to represent colors, it is also called a 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 used these in our visual features.

#### RGB Color Space

RGB color space consists 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. The range of intensity values is 0-255. This leads to 16,777,216 different colors, when used in 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 backlit (LCD monitor). RGB color space has been used in most 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 Marko Tkalcic [2003], International Commission on Illumination in 2003. It is also a three dimensional color space with three components. The three components in 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) represents red color where as $a*$(-128) represents green color. $b*$(+127) represents yellow color where as $b*$(-128) represents blue color.

To convert an image from CIELAB to RGB, we first converted the RGB image to CIEXYZ color space. Then we converted CIEXYZ to CIELAB. For doing this, we used MATLAB functions (makecform, applycform). The formulas for converting RGB color space to CIELAB are as follows: First conversion is from RGB to XYZ and 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 tri-stimulus values of the reference white.

**Color Histogram**

The Color Histogram is a very prominent talked 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 that lie in a particular color range. The color ranges are histogram bins for the color histogram model. Color Histograms are independent to rotation of the image. Therefore, even if the image is tilted, it won't affect the classification. Apart from this, computational efficiency is another strong reason behind using color histograms in classification problems. 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. We, therefore, do not choose it for our color histogram because of it being device dependent

CIELAB color space appears to be a better choice because of its perceptual uniformity and device independence. 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 constructed bins of 4 for each L, a and b component. Thus we had a total of 64 color bins. Thus for each block we get a 64 dimensional color histogram. When we combine 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 some human interpretable features from the image. 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 is an important feature. In normal terms, we can define texture of an image in estimate of smoothness of that image. In everyday terms, texture can be defined with words as rough, bumpy or silky.

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

Image texture also works in the same way. 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 interrelation of the grey tones actually determines

the textural pattern. It was first introduced by Haralick. [1979], so these are also called Haralick features .

For calculating GLCM features for an image, we first converted it to a gray-scale image, 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 calculated 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 is $d$ and direction specified by angle $\theta$. In standard GLCM matrix, the angles used are 0°, 45°, 90°and 135°with $d = 1$ *pixel*. This directional component of $\theta$ makes it more powerful in the sense that it represents 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 occurrences of the pixel intensity value 0 with horizontally adjacent pixels with intensity = 2 in the input. We computed the matrix G as symmetric, as we considered pair (0, 2) as (2, 0) as well. Matrix G can also be computed with non-symmetric measure.

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

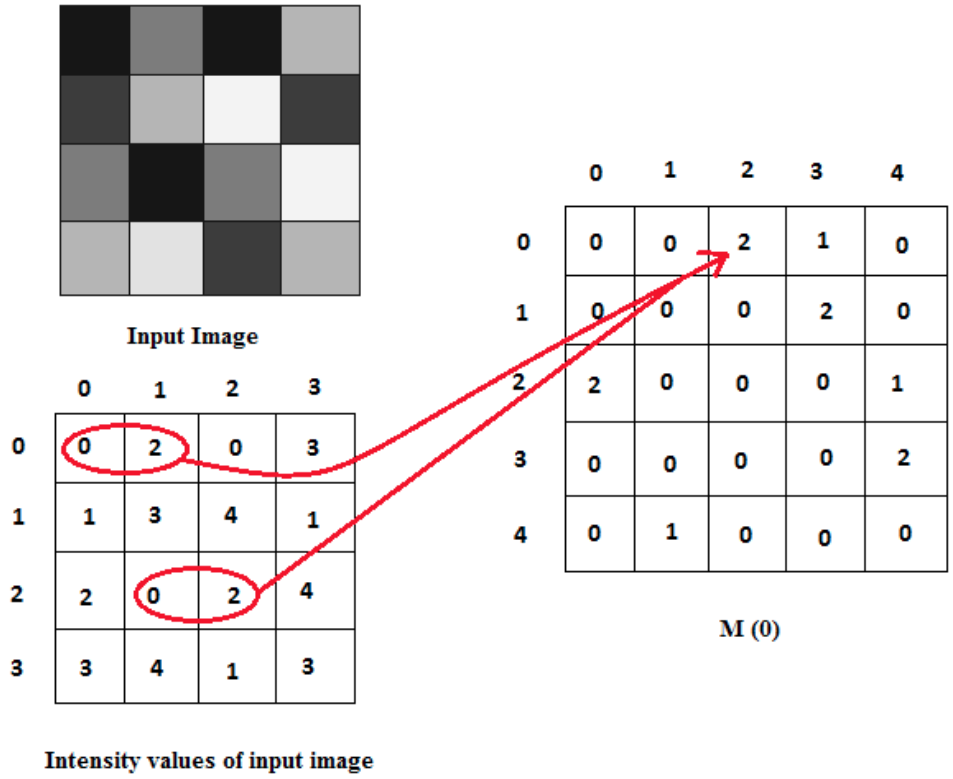


FIGURE 4.6: Co-occurence Matrix G(0°) generation for N=5 levels

### 4.1.5 HOG-LBP Features

HOG or Histogram of Oriented Gradients is a widely used feature for object recognition in Computer Vision. The idea behind this descriptor is that the object in an image can be characterized by the 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 [2005] for pedestrian detection in 2005. After that these descriptors clubbed with LBP features are usually used for object recognition Torrione, Morton, Sakaguchi, and Collins [2014], Yu, Zhang, Huang, Zheng, Ren, and Wang [2010],Zhang, Huang, Yu, and Tan [2011] etc.

The steps for constructing HoG descriptors are shown in figure 4.7.

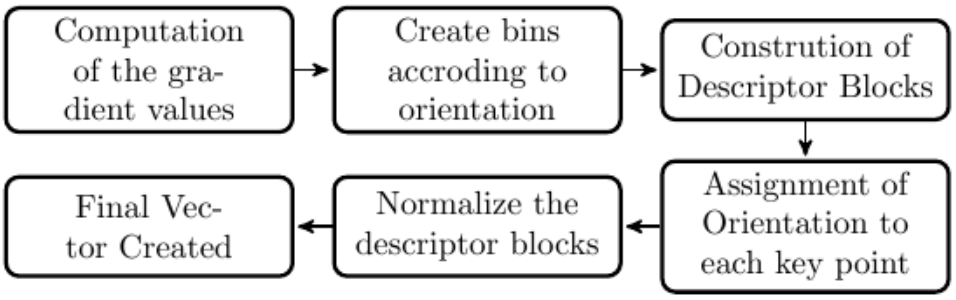


FIGURE 4.7: Construction of HoG descriptors

Local Binary Pattern (LBP) is a texture classification feature. It was first introduced by Ojala, Pietikainen, and Harwood [1994] 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.

**?** tried combining HoG with LBP. The results indicated high improvement in performance in case of object detection. If the image is cluttered with blurred edges, HoG loses its discriminating capability, LBP acts as a complementary feature to HoG in such cases. LBP uses uniform patterns to remove the noisy edges from a cluttered image. Castrillón-Santana, Lorenzo-Navarro, and Ramón-Balmaseda [2013] have shown that the combination of HoG and LBP acts much better than each individually. We, therefore, considered a combination of HoG and LBP for our classification.

We used VLFEAT Vedaldi and Fulkerson. [2008] library for computing HoG features. VLFEAT has two variants of HoG. One is UoCTTI variant, other is Dalal and Triggs's variant (Dalal and Triggs [2005]). 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 obtained 31 dimensional HoG vector for each cell.

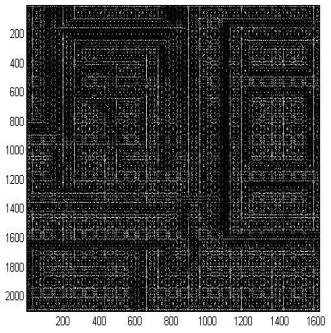


FIGURE 4.8: Example of HOG Descriptors

For computing LBP features, we again used VLFEAT library (Vedaldi and Fulkerson. [2008]). VLFEAT considered a $3 * 3$ neighborhood, this lead to LBP features that is a 8 bit long string vector. This 8 bit long vector can assume $2^8 = 256$ possible values. These 256 possible values were further quantized into a smaller number of patterns. This uniform quantization makes LBP features computationally efficient.

In this uniform quantization, we made the following observations. There was 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 comprised of two uniform LBPs and one pattern comprised all other LBPs. These observations yield a total 58 patterns. When we concatenated both HoG and LBP vector descriptors we get combined vector of 89 dimensions. We further used the bag of words approach on this combined HoG-LBP vector. We formed a bag(visual dictionary) of 4000 visual words using K-means implementation of VLFEAT and after that we combined the histogram on this visual dictionary for each vector, which gave 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 the right part shows the HoG descriptor for the image.

## 4.2 Social Content Based Feature Extraction

Social meta-data obtained from images had similar properties like a text data-set, because we had tags, comments, groups all in normal language text. So, It makes sense to just use text processing methods here. But, this text data also had inherited structure of a social network. We, therefore, tried to utilize this extra aspect, first constructed node features over the social-metadata for each image as shown by Julian McAuley [2012]. These node/social feature vectors had high dimensionality. We, therefore, used the topic modeling/text processing methods over these social features to construct a better and reduced representation. Tang, Sun, Wang, and Yang [2009] have shown that such topic-level modeling of social- networking data leads to good results in finding patterns and making inferences. These final low-dimensional features projected 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 from the social meta-data obtained can be divided into the following steps:

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

In what folows we will discuss each of these steps.

### 4.2.1 Pre-analysis of Social Data

Preliminary observations of the data suggest 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 for example sea-side or mountain ranges. Groups

are similar to tags, with the difference that the groups in which an image is featured, are chosen entirely by the image's author.

### 4.2.2 Constructing Node Features

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

We first constructed an indicator vector via 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 Julian McAuley [2012], this data set of only 1000 most popular words did not sufficiently represent the whole data. We, therefore, also considered any words, groups and tags that are at least twice as frequent in images having the label in question compared to the overall rate. This way we got similar node features as described in Julian McAuley [2012].

For developing this word feature, we utilized text from the image's title, it's comment thread and description after eliminating stop-words. This will give us more than 40000+ points. The node-feature vector was in binary form and had high dimensionality. We had 0 and 1 as the value for each field in this vector corresponding to the presence of the word in the image data or not. We further converted this raw binary 0 and 1 form, to usable social features with the use of 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 on these feature vectors. For text processing, we can consider each image as a document and the current node feature vector as a representation of the dictionary. This vector represents the presence of

a word in the document. Now, we have a corpora of word features and in the 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 Dirichlet 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 show a similar contextual presence.In Deerwester [1988], 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. Deerwester [2007].

In typical information retrieval methods information is retrieved by literally matching terms in in the search space with those of the search query. However, this method, which depends purely on lexical matching, can be inaccurate. Since, there are many ways to express a given concept literal matching may not provide us the relevant information. A better approach will be to create a basis for conceptual topic in 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 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 prove to be more robust indicators of information than individual terms.

**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 projects 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 the GENSIM library (Řehůřek and Sojka [2010]) in Python for Latent Semantic Indexing of our data. It was developed by Rehurek and Sojka [2010] for topic modeling with large corpora.

#### 4.2.3.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is another frequently used process in natural language processing. It is a generative model that allows a set of observations to be depicted by unobserved groups explaining why some parts of the data are quite similar. For example in a general natural language processing scenario, when the observations are words associated with a document, it assumes a document is a mixture of a small number of topics and that

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

each word's presence is dedicated to one of the document's concepts. LDA was actually a graphical model presented in Blei [2003] for topic discovery. LDA has connection with image classification because in Li [2005] a variation on LDA was used to automatically split the natural images into categories, such as forest and mountain, by assuming the images as words. Latent Dirichlet allocation uses a 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. In Blei [2003], 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 plate model of Linear Discriminant Analysis. With plate notation, the dependencies among the many variables can be captured concisely. The boxes are plates representing replicates. The outer plate represents the documents, while the inner plate represents the repeated choice of topics and words within a documents. M denotes the number of documents. N the number of words in a document. $\alpha$ is the parameter of the Dircihlet prior on the per-document topic distributions, $\beta$ is the parameter of the dirichlet prior on the per-topic: word distribution, $\theta_i$ is the topic distribution for document i, $\phi_k$ is the is the word distribution for topic k,$z_{ij}$ is the topic for the jth word in document $i$, and $w_{ij}$ is the specific word.The $w_{ij}$ are the only observable variables, We can then mathematically describe the random variables as follows:

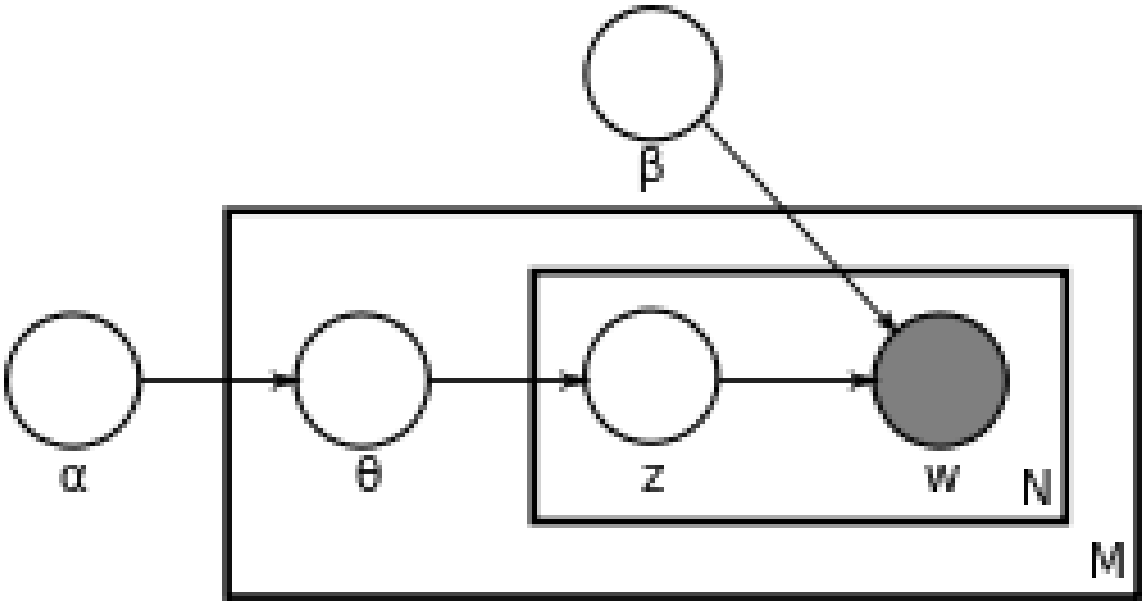


FIGURE 4.9: Plate Model for LDA Blei [2003]

- $\phi_{k=1.....K} \sim Dirichlet_V(\beta)$
- $theta_{d=1...M} \sim Dirichlet_K(\alpha)$
- $z_{d=1.....M,w=1.....N_d} \sim Categorical_K(\theta_d)$
- $w_{d=1.....M,w=1.....N_d} \sim Categorical_V(\phi_{z_{dw}})$

In our case the LDA did not provide good results because we already have textual data which is oriented to solve a binary classification problem. This binary nature of a topic leads to too sparse feature generation from Latent Dirichlet allocation (LDA). This sparseness of features around images lead to low quality classification. LDA also fails if discriminatory information is not in the mean but in the variance of the data Blei [2003]. We, therefore, discard this method for our computations.

#### 4.2.3.3 Random Projections

Random Projections is a powerful methods for dimensionality reductions in image and text data (Bingham and Mannila). In Bingham and Mannila [2001], Bingham introduced random projections as a simpler and less erroneous dimensionality reduction tool for information retrieval from text and processing of images. It is very useful in cases where reduction of high dimensional data to low dimensions is essential. Which if not done leads to heavy computation penalty without 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,

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

is the projection of the data onto a lower k-dimensional subspace. The key idea of random mapping arises from the Johnson-Lindenstrauss lemma [Johnson et al., 1986]: 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 [Johnson et al., 1986] 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 [Řehůřek and Sojka, 2010] in Python. It has been found that even though it is computationally light, Random projections is a sufficiently accurate method for dimensionality reduction of high dimensional data Dasgupta [2000].

### 4.2.4 Implementation of Dimensionality reduction

Considering that we are using a large database and we need to do dimensionality reduction for such data. We use an online version of aforementioned techniques. So that we don't have to bother about loading the whole data into memory.

Both LSI and RP rely on TF-IDF (term frequency - inverse document frequency) as a fast pre-processing step. Rehurek and Sojka [2010] gives a framework Řehůřek and Sojka [2010] for doing all this text processing on large corpora in memory independent fashion. We use this as a tool for doing LSI and RP Computation. On varying the number of dimensions 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 use kernel methods which would be too costly and too susceptible to over-fitting

with thousands of binary features. We directly converted the node features of dimension 40000+ in social features to dimension 300 (LSI) and dimension 400 (RP). This conversion was done in Python using GENSIM and the converted files 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 then we discuss the classifiers tested. In end, 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 chapter **??** Apart from these visual features, we constructed 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 lead to two different feature vectors. First, where dimensionality reduction was done by Latent Semantic Indexing and the second, where dimensionality reduction was by Random Projections.

## 5.2 Classifiers Used

After experimenting with various classifiers like Random Forest, MLP (Multi Layer Perceptron), libSVM (with various kernels linear, RBF, $\chi^2$, histogram intersection), we found that in case of image features libSVM with $\chi^2$ kernel worked best and in case of social features libSVM with the linear kernel gave us the best results. We,therefore, used libSVM with $\chi^2$ kernel as the classifier for visual feature vectors and libSVM with Linear Kernel as the 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 the four data-sets. First, we have shown the classification results from all the features extracted. We first experimented with the classifiers based on individual features, means separate classifier from SIFT, HoG, GIST , GLCM, COLOR, social Features generated through LSI and social Features generated through LSI. In next step, we did ensemble of classifiers obtained by different image and social meta data classifiers to learn the linear combination of weighted classifier.

We have also given qualitative and quantitative conclusions/observations of our results. We have shown a comparison with published results on each of these four benchmarks or with results in associated competitions.

The goal of all these comparisons is to assess the improvement that can be obtained by using social meta-data for images. We reported the mean average precision (MAP) for the sake of comparison with published materials and competition results. We also gave the accuracy for the binary prediction/classification of labels.

All these results are for tenfold cross validation. For ensemble methods, we divided the data in three parts: training, testing and validation. We randomly chose 10% instances

for validation set and learned weights for the linear combination of the classifiers, which provided best results. After learning these weights, we used the optimized combination of the classifiers to do testing on test set.

For better visualization, we have broken our result in three tables, for each data set.

- In first table, we have shown results using only the image descriptors. In this table, we have only considered the visual features and shown the classification result for each of extracted visual feature.
- In second table, we have shown results using only the social meta-data descriptors. In this table, we have only considered the social meta-data features . We have shown the classification result for each of the method LSI and Random Projections.
- In third table, we have compared result with the published paper. In this comparison, we have taken the result of the ensemble of social and visual classifiers, and best of visual and best of social descriptor

We have created such tables for comparing mean average precision and accuracy for each of the data-set.

### 5.3.1 MIR Flickr collection

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

In [Huiskes, 2010], a combination of social data and low-level content-based descriptors improve the accuracy of visual concept classifiers. We use the results of this paper as a comparison metric for our results.

In [Huiskes, 2010], 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 used flickr tags of visual concepts. A set consisting of 293 binary features was developed using these tags. These tags were chosen such that every tag corresponds with at least 50 images in the MIR Flickr collection.

They have used two classifiers one is Linear Discriminant Analysis and other is support vector machines. For each of these classifier, they have first tested with classifier using only the image descriptors and then they tested using the image desciptors + Flickr tags as features. The classifiers were trained on the 24 potential labels, and 14 regular (subjective) annotations. We chose the results of the labels, we have considered. We could not cover all the labels because for other labels, we could not get much data on flickr. That data might be available at that time but now those image URLs are either denying access or do not exist. So, we refrain our self to a subset of those 38 labels.

We have compared the classification accuracy and mean average precision between classifiers based on low-level features only and classifiers that additionally use the Flickr tags as features. We have shown our result in following tables.

- In table 5.1, we have shown accuracy using only the image descriptors. In this table, we have only considered the visual features and shown the classification result for each of extracted visual feature.
- In table 5.2, we have shown accuracy using only the social meta-data descriptors. In this table, we have only considered the social meta-data features . We have shown the classification result for each of the method visual feature, LSI and Random Projections.

- In table 5.3, we have compared our accuracy with the published paper. In this comparison, we have taken the result of the ensemble of social and visual classifiers, and best of visual and best of social descriptor. We have divided the result of [Huiskes, 2010] in two parts. One is for SVM Classifier and second is for LDA Classifier. Each of this has two sub parts, one is for only the image descriptors and second sub part is for combined descriptor of image and flicker tags.

- In table 5.4, we have shown mean average precision using only the image descriptors. In this table, we have only considered the visual features and shown the classification result for each of extracted visual feature.

- In table 5.5, we have shown mean average precision using only the social meta-data descriptors. In this table, we have only considered the social meta-data features . We have shown the classification result for each of the method visual feature, LSI and Random Projections.

- In table 5.6, we have compared our mean average precision with the published paper. In this comparison, we have taken the result of the ensemble of social and visual classifiers, and best of visual and best of social descriptor.

**Observations**

- The classifications based on visual only features gave an average precision of 76.15%, which outperformed the low level image descriptor based classification with 40.43% in case of LDA and 44.38% in case of SVM. The result was also better than classification based on the combination of low level image descriptors and Flickr tags. Here we saw a precision increment of 28% as compared to SVM results in paper and 26.40% as compared to LDA results in paper. This result is an indication that our choice of image descriptors gives a better semantic analysis of image then the low level image descriptors used in image.

- When we do classification on the basis of social features computed using LSI, we got an average precision of 87.59%. This was 39.71% more compared to SVM classification of combined features of Flickr tags and low level image descriptors and 37.83% more compared to LDA classification of this combined feature set. This might be because we covered more tags and sources of meta-data then the taken by the paper.

- LSI based social feature classification outperformsed 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 consistent with the results shown in Julian McAuley [2012]. 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 related to these images. For example, we can see that in 5.1, first image is a non-cloud image and second image is image labeled as cloud. The second image has most of the comments saying nice view, beautiful place etc. Very less textual cues actually indicate the cloud part. Apart from this, these images contains various visual descriptors , which is more emphasized intensity gradients or distributed edge directions. Therefore, HOG-LBP and GLCM features works quite better for this label.

- 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

FIGURE 5.1: MIR Cloud Examples

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

- The classifications based on visual only features give an average accuracy of 76.03%. ***** You have not ensembled visual only features. See what you get when you ensemble visual only features and compare. *****
- 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 ensemble 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.*

TABLE 5.1: MIR Precision Comparison: Only Visual Feature

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GLCM |
| Flower | 78.94 | 41.60 | 73.19 | 70.37 | 64.40 |
| Car | 83.27 | 56.54 | 71.70 | 64.35 | 76.65 |
| Bird | 70.11 | 47.96 | 61.77 | 60.63 | 71.17 |
| Dog | 73.68 | 47.39 | 68.80 | 65.31 | 67.67 |
| Night | 82.40 | 58.00 | 70.40 | 79.16 | 81.37 |
| Tree | 76.92 | 54.59 | 74.97 | 61.64 | 76.11 |
| Clouds | 90.91 | 50.88 | 79.78 | 67.31 | 76.55 |
| Portrait | 69.31 | 50.61 | 65.30 | 59.99 | 65.85 |
| Female | 63.72 | 52.56 | 62.27 | 54.11 | 58.77 |
| Male | 60.16 | 50.62 | 60.96 | 52.88 | 61.79 |
| People | 68.36 | 60.37 | 58.19 | 55.56 | 59.18 |
| Sea | 88.48 | 53.04 | 77.20 | 73.42 | 81.05 |
| River | 77.14 | 46.68 | 69.73 | 68.55 | 71.10 |
| Baby | 77.58 | 47.71 | 72.29 | 73.86 | 80.06 |

### 5.3.2 ImageCLEF

ImageCLEF has 99 labels. For some labels we have less than 20 instances. Learning a visual bag of words for so few instances is not practical. Such small set of images would not be able to deliver a bag words, exhaustive enough to bag all the characteristics of that label. We, therefore, discard such labels. After analysing the availability instances of images per label and their respective meta data, we restricted ourselves to following labels.

*'Adult','Aesthetic_ Impression','Animals','Autumn','City life','cute','Day','Flowers','Food','Graffiti' ,'Landscape_ Nature', 'Painting', 'Portrait', 'Single_ Person', 'Sky', 'Street', 'Summer', 'Sunset/Sunrise', 'Vehicle', 'Winter'*

Nowak [2010] has the best comparative published results for judging our hypothesis as it already has results based on Flickr user tags and multi-modal approaches that consider

TABLE 5.2: MIR Precision Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Flower | 91.37 | 74.34 |
| Car | 92.21 | 72.68 |
| Bird | 93.37 | 79.26 |
| Dog | 95.94 | 73.38 |
| Night | 87.81 | 71.85 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Tree | 81.43 | 69.18 |
| Clouds | 82.75 | 74.12 |
| Portrait | 83.54 | 71.20 |
| Female | 81.52 | 69.90 |
| Male | 74.51 | 67.04 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| People | 90.77 | 73.32 |
| Sea | 93.34 | 73.85 |
| River | 82.39 | 73.58 |
| Baby | 95.24 | 73.24 |
| Average | 87.59 | 72.64 |

visual information and/or Flickr user tags and/or Exif Information.In their classification, textual information is represented as a binary presence/absence vector of the most common 698 Flickr user tags. An approach applied to visual words, a visual words baseline with SIFT descriptors, colour and texture features, which was similar as bag of words model on these features. They further combined these features with flickr User Tag system using tf-idf, which is very much similar to our LSI method. But they used small set of user tags compared to us and other data like comments, gallery etc. was not included in their test data. In following section, we describe what are the results of this additional information.

- In Table 5.10, we have shown mean average precision using only the image descriptors. In this table, we have only considered the visual features and shown the classification result for each of extracted visual feature.

- In 5.11, we have shown mean average precision using only the social meta-data descriptors. In this table, we have only considered the social meta-data features .

TABLE 5.3: MIR Precision Comparison: Comparison of Published Results, results of ensemble and best of social and visual features

| Labels | Ensemble | Published Results | | | | Best of Social and Visual Features | |
|---|---|---|---|---|---|---|---|
| | | SVM | | LDA | | Social | Visual |
| | | Flickr tags+ Image Descriptors | Image Descriptors Only | Flickr tags + Image Descriptors | Image Desciptors Only | | |
| Flower | 92.96 | 48.00 | 46.90 | 56.00 | 30.10 | 91.37 | 78.94 |
| Car | 96.91 | 33.90 | 17.90 | 29.70 | 14.20 | 92.21 | 83.27 |
| Bird | 95.77 | 44.30 | 12.80 | 42.60 | 9.70 | 93.37 | 71.17 |
| Dog | 98.15 | 60.70 | 15.50 | 62.10 | 10.80 | 95.94 | 73.68 |
| Night | 90.53 | 58.80 | 55.40 | 61.50 | 51.50 | 87.81 | 82.40 |
| Clouds | 92.00 | 69.50 | 65.10 | 65.10 | 57.70 | 82.75 | 90.91 |
| Portrait | 84.08 | 48.00 | 49.30 | 54.30 | 43.20 | 83.54 | 69.31 |
| Female | 83.55 | 46.40 | 46.10 | 49.40 | 40.40 | 81.52 | 63.72 |
| Male | 76.34 | 41.30 | 40.70 | 43.40 | 35.60 | 74.51 | 61.79 |
| People | 93.91 | 74.80 | 36.10 | 73.10 | 62.80 | 90.77 | 68.36 |
| Sea | 98.01 | 52.90 | 36.60 | 47.70 | 25.50 | 93.34 | 88.48 |
| River | 83.35 | 15.80 | 17.90 | 31.70 | 13.00 | 82.39 | 77.14 |
| Baby | 98.55 | 20.00 | 8.40 | 28.50 | 6.90 | 95.24 | 77.58 |
| Average | 90.58 | 47.88 | 35.72 | 49.76 | 31.77 | 87.59 | 75.79 |

TABLE 5.4: MIR Accuracy Comparison: Only Visual Features

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| Flower | 79.50 | 44.00 | 73.17 | 70.50 | 64.17 |
| Car | 83.48 | 54.78 | 70.65 | 64.13 | 77.17 |
| Bird | 70.00 | 48.21 | 62.14 | 59.82 | 69.82 |
| Dog | 73.33 | 47.83 | 68.00 | 64.17 | 67.67 |
| Night | 82.50 | 53.00 | 70.33 | 79.33 | 82.00 |
| Tree | 77.00 | 53.33 | 74.00 | 61.17 | 75.00 |
| Clouds | 89.33 | 47.33 | 77.67 | 67.50 | 75.67 |
| Portrait | 70.33 | 49.33 | 65.00 | 59.83 | 66.83 |
| Female | 63.00 | 52.17 | 61.50 | 52.50 | 59.00 |
| Male | 60.00 | 47.67 | 60.67 | 52.00 | 60.50 |
| People | 68.33 | 47.33 | 57.83 | 55.17 | 59.17 |
| Baby | 88.33 | 53.75 | 76.67 | 72.08 | 81.25 |
| Average | 79.38 | 46.88 | 68.75 | 66.25 | 69.38 |

TABLE 5.5: MIR Accuracy Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Flower | 89.17 | 73.33 |
| Car | 91.52 | 71.09 |
| Bird | 92.14 | 78.57 |
| Dog | 93.33 | 72.83 |
| Night | 86.33 | 71.00 |
| Tree | 81.33 | 68.33 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Clouds | 83.33 | 73.17 |
| Portrait | 84.33 | 71.33 |
| Female | 82.83 | 69.83 |
| Male | 76.67 | 66.00 |
| People | 88.33 | 72.50 |
| Baby | 89.17 | 72.50 |
| Average | 78.13 | 72.50 |

We have shown the classification result for each of the method visual feature, LSI and Random Projections.

- In Table 5.12, we have compared our mean average precision with the published paper. In this comparison, we have taken the result of the ensemble of social and

TABLE 5.6: MIR Accuracy Comparison: Comparison of Published Results, Results of ensemble and Best of social and visual features

| Labels | Ensemble Method | Social Only | Visual Only |
|---|---|---|---|
| Flower | 89.64 | 89.17 | 79.5 |
| Car | 94.10 | 83.48 | 83.48 |
| Bird | 94.70 | 91.52 | 70.00 |
| Dog | 95.92 | 92.14 | 73.33 |
| Night | 89.22 | 93.33 | 82.50 |
| Tree | 85.31 | 86.33 | 77.00 |
| Clouds | 89.76 | 81.33 | 89.33 |
| Portrait | 84.59 | 83.33 | 70.33 |
| Female | 84.63 | 84.33 | 63.00 |
| Male | 79.72 | 82.83 | 60.67 |
| People | 91.42 | 76.67 | 68.33 |
| Baby | 93.06 | 88.33 | 88.33 |
| Average | 79.93 | 89.17 | 79.38 |

visual classifiers, and best of visual and best of social descriptor. From the Nowak [2010], we took the best results for that label and compared our results with that.

- In Table 5.7, we have shown accuracy using only the image descriptors. In this table, we have only considered the visual features and shown the classification result for each of extracted visual feature.

- In 5.8, we have shown accuracy using only the social meta-data descriptors. In this table, we have only considered the social meta-data features . We have shown the classification result for each of the method visual feature, LSI and Random Projections.

- In 5.9, we have compared our accuracy with the published paper. In this comparison, we have taken the result of the ensemble of social and visual classifiers, and best of visual and best of social descriptor. From the Nowak [2010], we took the best results for that label and compared our results with that.

.

**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 5.9. The reason of this is that we used more exhaustive set of social meta-data. We used tags, galleries, group, comments etc., whereas the Nowak [2010] used only handful of tags. This enriched social meta-data improved our result.
- Ensembling of all the features outperforms the published results in precision with average of 6.93%.5.9
- Ensembling of all the features outperforms the published results in accuracy with average of 1.26%.5.12
- 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 was sparse.
- For the labels like Autumn, Landscape, CityLife, Paintings, Sky, Street and Sunset Sunrise we see that visual feature based computation works quite well because these images are more visual concept centric and social data has lesser role to play there. The comments and tags on these images are more generic like "beautiful view", 'serene' etc. These textual information are not centric to the label or class of images like Autumn , Landscape etc. An image of 'sea', 'waterfall' etc can also have similar tags and comments. This weakness the distinguishing power of meta-data.
- 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 an important role because it is quite different from the normal natural photos.

TABLE 5.7: ImageCLEF Precision Comparison: Only Visual Features

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| Adult | 60.94 | 56.88 | 54.7 | 56.45 | 60.24 |
| Aesthetic Impression | 59.58 | 53.33 | 54.94 | 55.60 | 50.82 |
| Animals | 65.67 | 54.52 | 58.87 | 58.65 | 60.09 |
| Autumn | 76.01 | 69.43 | 67.71 | 58.74 | 65.18 |
| Citylife | 69.87 | 44.81 | 62.13 | 52.6 | 62.63 |
| cute | 55.11 | 51.49 | 52.67 | 52.41 | 51.13 |
| Day | 67.39 | 54.79 | 59.98 | 63.3 | 61.41 |
| Flowers | 73.75 | 52.97 | 65.75 | 65.94 | 57.58 |
| Food | 73.3 | 68.41 | 72.84 | 67.89 | 59.68 |
| Grati | 66.02 | 61.5 | 61.32 | 56.33 | 56.63 |
| Landscape Nature | 77.95 | 52.49 | 66.03 | 62.66 | 68.3 |
| Painting | 60.26 | 51.96 | 55.7 | 53.26 | 62.49 |
| Portrait | 64.98 | 57.47 | 60.8 | 61.9 | 63.16 |
| Single Person | 58.97 | 56.14 | 56.37 | 54.8 | 60.28 |
| Sky | 80.42 | 54.78 | 73.07 | 69.15 | 67.28 |
| Street | 67.86 | 60.37 | 62.71 | 53.38 | 63.46 |
| Summer | 62.96 | 52.3 | 56.92 | 59.51 | 59.18 |
| Sunset Sunrise | 85.49 | 57.4 | 71.96 | 71.15 | 82.35 |
| Vehicle | 74.42 | 51.68 | 62.11 | 58.25 | 69.1 |
| Winter | 69.89 | 60.94 | 58.3 | 58.47 | 64.1 |
| Average | 68.54 | 56.18 | 61.74 | 59.52 | 62.25 |

### 5.3.3 PASCAL

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

TABLE 5.8: ImageCLEF Precision Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Adult | 89.56 | 75.24 |
| Aesthetic Impression | 63.91 | 59.30 |
| Animals | 92.83 | 74.05 |
| Autumn | 85.89 | 65.67 |
| Citylife | 78.47 | 63.79 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| cute | 63.11 | 61.01 |
| Day | 84.15 | 68.67 |
| Flowers | 93.54 | 69.79 |
| Food | 87.51 | 71.04 |
| Grati | 81.17 | 67.22 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Landscape Nature | 84.45 | 75.04 |
| Painting | 70.51 | 59.1 |
| Portrait | 85.82 | 75.59 |
| Single Person | 91.25 | 76.48 |
| Sky | 87.3 | 76.27 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Street | 79.2 | 65.91 |
| Summer | 71.8 | 65.56 |
| Sunset Sunrise | 87.09 | 78.51 |
| Vehicle | 87.98 | 69.92 |
| Winter | 85.02 | 75.9 |
| Average | 82.53 | 69.7 |

- 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

- In 5.13, we have shown mean average precision using only the image descriptors. In this table, we have only considered the visual features and shown the classification result for each of extracted visual feature.
- In **??**, we have shown mean average precision using only the social meta-data descriptors. In this table, we have only considered the social meta-data features . We have shown the classification result for each of the method visual feature, LSI and Random Projections.

TABLE 5.9: ImageCLEF Precision Comparison: Comparison of Published Results, Results of ensemble and Best of social and visual features

| Labels | Ensemble Method | CLEF Best Results | Social Only | Visual Only |
|---|---|---|---|---|
| Adult | 91.76 | 77.21 | 89.56 | 60.94 |
| Aesthetic Impression | 67.74 | 60.66 | 63.91 | 59.58 |
| Animals | 93.76 | 84.34 | 92.83 | 65.67 |
| Autumn | 88.12 | 83.51 | 85.89 | 76.01 |
| Citylife | 82.02 | 78.37 | 78.47 | 69.87 |
| cute | 64.49 | 59.71 | 63.11 | 55.11 |
| Day | 87.43 | 80.75 | 84.15 | 67.39 |
| Flowers | 94.13 | 82.72 | 93.54 | 73.75 |
| Food | 92.3 | 85.2 | 87.51 | 73.3 |
| Grati | 84.09 | 66.21 | 81.17 | 66.02 |
| Landscape Nature | 88.21 | 88.68 | 84.45 | 77.95 |
| Painting | 73.04 | 72.43 | 70.51 | 62.49 |
| Portrait | 90.27 | 81.34 | 85.82 | 64.98 |
| Single Person | 93.99 | 76.41 | 91.25 | 60.28 |
| Sky | 88.05 | 89.26 | 87.3 | 80.42 |
| Street | 83.41 | 76.76 | 79.2 | 67.86 |
| Summer | 75.87 | 74.08 | 71.8 | 62.96 |
| Sunset Sunrise | 91.74 | 91.83 | 87.09 | 85.49 |
| Vehicle | 88.97 | 78.64 | 87.98 | 74.42 |
| Winter | 88.1 | 80.71 | 85.02 | 69.89 |
| Average | 85.37 | 78.44 | 82.72 | 69.13 |

TABLE 5.10: ImageCLEF Accuracy Comparison: Only Visual Features

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| Adult | 60.33 | 55.5 | 53.83 | 56.33 | 60.67 |
| Aesthetic Impression | 59.33 | 45.33 | 54.5 | 54.83 | 48.83 |
| Animals | 65.67 | 53.17 | 58.67 | 58 | 59.33 |
| Autumn | 74.29 | 66.43 | 63.57 | 57.86 | 63.57 |
| Citylife | 69.83 | 47.67 | 62.83 | 52.33 | 62.83 |
| cute | 55.17 | 49.5 | 52.17 | 52 | 47 |
| Day | 66.83 | 49.33 | 58.67 | 62.67 | 62 |
| Flowers | 72.25 | 52.25 | 65.75 | 65.25 | 56.75 |
| Food | 73.33 | 51.67 | 72.33 | 67 | 60.33 |
| Grati | 56.43 | 50.71 | 57.14 | 53.57 | 56.43 |
| Landscape Nature | 77 | 51.5 | 66 | 60.83 | 67.5 |
| Painting | 60 | 51.67 | 50.28 | 53.06 | 61.39 |
| Portrait | 66 | 52 | 60.33 | 62 | 63.33 |
| Single Person | 58.33 | 47.17 | 55.67 | 54.33 | 59.67 |
| Sky | 79.5 | 52.67 | 71.83 | 67.33 | 65.5 |
| Street | 67.83 | 50.83 | 63.33 | 53 | 64.17 |
| Summer | 62.67 | 52 | 56.83 | 58.33 | 59.67 |
| Sunset Sunrise | 85 | 55 | 72.37 | 70.53 | 81.05 |
| Vehicle | 73.5 | 48.33 | 61.83 | 57.83 | 69.33 |
| Winter | 70 | 57.5 | 58.75 | 56.25 | 64.17 |
| Average | 67.66 | 52.01 | 60.83 | 58.67 | 61.68 |

- In 5.15, , we have compared our mean average precision with the VOC Competition results. In this comparison, we have taken the result of the ensemble of social and visual classifiers, and best of visual and best of social descriptor

- In Table 5.16, we have shown accuracy using only the image descriptors. In this table, we have only considered the visual features and shown the classification result for each of extracted visual feature.

- In 5.17, we have shown accuracy using only the social meta-data descriptors. In

TABLE 5.11: ImageCLEF Accuracy Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Adult | 90.03 | 80.73 |
| Aesthetic Impression | 68.01 | 60.89 |
| Animals | 92.13 | 88.43 |
| Autumn | 86.16 | 88.31 |
| Citylife | 81.02 | 82.67 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| cute | 65.72 | 59.24 |
| Day | 82.82 | 85.17 |
| Flowers | 91.74 | 87.03 |
| Food | 87.36 | 88.7 |
| Grati | 75.9 | 69.85 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Landscape Nature | 84.69 | 91.54 |
| Painting | 71.96 | 76.28 |
| Portrait | 90.5 | 85.59 |
| Single Person | 91.89 | 80.04 |
| Sky | 87.36 | 91.75 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| Street | 79.52 | 81.2 |
| Summer | 71.97 | 78.5 |
| Sunset Sunrise | 88.24 | 93.24 |
| Vehicle | 88.5 | 82.38 |
| Winter | 86.48 | 85.36 |
| Average | 83.1 | 81.84 |

this table, we have only considered the social meta-data features . We have shown the classification result for each of the method visual feature, LSI and Random Projections.

- In 5.18 we have shown the result of the ensemble of social and visual classifiers, and best of visual and best of social descriptor. These results shows the accuracy of classifying various labels in binary prediction environment.

**Observations**

- Observations:

TABLE 5.12: ImageCLEF Accuracy Comparison: Results of ensemble and Best of social and visual features

| Labels | Ensemble Method | CLEF Best Results | Social Only | Visual Only |
|---|---|---|---|---|
| Adult | 90.03 | 80.73 | 88.5 | 60.67 |
| Aesthetic Impression | 68.01 | 60.89 | 64.83 | 59.33 |
| Animals | 92.13 | 88.43 | 90.17 | 65.67 |
| Autumn | 86.16 | 88.31 | 83.57 | 74.29 |
| Citylife | 81.02 | 82.67 | 78 | 69.83 |
| cute | 65.72 | 59.24 | 63 | 55.17 |
| Day | 82.82 | 85.17 | 82.17 | 66.83 |
| Flowers | 91.74 | 87.03 | 89.75 | 72.25 |
| Food | 87.36 | 88.7 | 86 | 73.33 |
| Grati | 75.9 | 69.85 | 75 | 57.14 |
| Landscape Nature | 84.69 | 91.54 | 83.67 | 77 |
| Painting | 71.96 | 76.28 | 69.17 | 61.39 |
| Portrait | 90.5 | 85.59 | 86.67 | 66 |
| Single Person | 91.89 | 80.04 | 91.33 | 59.67 |
| Sky | 87.36 | 91.75 | 86.33 | 79.5 |
| Street | 79.52 | 81.2 | 78.5 | 67.83 |
| Summer | 71.97 | 78.5 | 71 | 62.67 |
| Sunset Sunrise | 88.24 | 93.24 | 86.84 | 85 |
| Vehicle | 88.5 | 82.38 | 87.5 | 73.5 |
| Winter | 86.48 | 85.36 | 84.58 | 70 |
| Average | 83.1 | 81.84 | 81.33 | 67.66 |

- 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%.
- Th ensemble of social and visual classifiers 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.
- 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% compared to the visual only features.
- The ensemble of social and visual features gives an accuracy of 72.49 % which is 2.71% more than only social features based classification and 5.81% more than only visual features based classification.
- The accuracy for visual only features for most labels are comparable with social features and do not have a difference more than 3%. This difference is much higher for the other three data sets. 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 and do not lead to much human interaction leading to less social data. We verified that PASCAL has the least number tags and comments per image compared to all other data-sets used in our thesis.

TABLE 5.13: PASCAL Precision Comparison: Only Visual Features

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| aeroplane | 75.42 | 56.53 | 72.82 | 65.2 | 78.79 |
| bicycle | 67.51 | 55.38 | 60.6 | 53.57 | 62.41 |
| bird | 62.49 | 55.64 | 65.89 | 50.86 | 65.3 |
| boat | 67.14 | 57.14 | 63.19 | 57.79 | 62.99 |
| bottle | 57.54 | 48.67 | 55.7 | 57.8 | 62.44 |

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| | LSI | RP | | | |
| bus | 74.26 | 56.43 | 66.24 | 57.97 | 66.13 |
| car | 61.25 | 54.2 | 58.41 | 54.69 | 60.63 |
| cat | 69.36 | 56.27 | 64.01 | 59.97 | 68.32 |
| chair | 61.19 | 54.38 | 51.94 | 57.86 | 61.14 |
| tvmonitor | 66.9 | 57.38 | 67.69 | 59.13 | 67.84 |

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| cow | 68.83 | 56.62 | 61.2 | 60.39 | 64.33 |
| diningtable | 67.67 | 62.7 | 65.11 | 57.66 | 62.59 |
| dog | 65.07 | 49.69 | 61.33 | 55.64 | 64.52 |
| horse | 67.49 | 46.98 | 58.99 | 56.93 | 62.26 |
| motorbike | 66.52 | 58.15 | 68.11 | 55.44 | 62.67 |

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| pottedplant | 59.6 | 53.57 | 55.97 | 59.72 | 57.1 |
| sheep | 73.34 | 49.68 | 67.7 | 62.14 | 68.27 |
| sofa | 64.89 | 53.21 | 56.27 | 54.22 | 62.71 |
| train | 71.61 | 56.76 | 65.79 | 57.61 | 63.24 |
| Average | 66.26 | 54.6 | 61.9 | 57.19 | 63.61 |

### 5.3.4 NUS

In Chua [2009], the authors have used six types of low-level features. These six type of image features include 64-D color histogram, 144-D color auto-correlogram, 73-D edge

TABLE 5.14: PASCAL Precision Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| aeroplane | 81.76 | 70.6 |
| bicycle | 67.65 | 60.58 |
| bird | 79.22 | 69.8 |
| boat | 63.67 | 54.88 |
| bottle | 66.3 | 61.95 |
| bus | 76.79 | 63.74 |
| car | 68.71 | 56.02 |
| cat | 77.73 | 68.49 |
| chair | 71.76 | 61.5 |
| cow | 71.63 | 63.67 |
| diningtable | 82.07 | 70.89 |
| dog | 78.93 | 64.97 |
| horse | 78.8 | 65.44 |
| motorbike | 76.8 | 59.73 |
| pottedplant | 64.34 | 57.02 |
| sheep | 78.75 | 65.73 |
| sofa | 70.16 | 63.35 |
| train | 83.48 | 69.57 |
| tvmonitor | 69.25 | 61.77 |
| Average | 73.67 | 63.28 |

direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT descriptions.

Color histogram serves as an effective representation of the color content of an image. It is defined as the distribution of the number of pixels for each quantized bin.

Color auto-correlogram (HSV) was proposed to characterize the colordistributions and the spatial correlation of pairs of colors together.

TABLE 5.15: PASCAL Precision Comparison: Comparison of Published Results, Results of ensemble and Best of social and visual features

| Labels | Ensemble Method | PASCAL Best Results | Social Only | Visual Only |
|---|---|---|---|---|
| aeroplane | 85.83 | 77.5 | 81.76 | 78.79 |
| bicycle | 68.29 | 63.6 | 67.65 | 67.51 |
| bird | 82.38 | 56.1 | 79.22 | 65.89 |
| boat | 68.54 | 71.9 | 63.67 | 67.14 |
| bottle | 71.09 | 33.1 | 66.30 | 62.44 |
| bus | 77.58 | 60.6 | 76.79 | 74.26 |
| car | 73.5 | 78 | 68.71 | 61.25 |
| cat | 81.73 | 58.8 | 77.73 | 69.36 |
| chair | 73.87 | 53.5 | 71.76 | 61.19 |
| cow | 75.59 | 42.6 | 71.63 | 68.83 |
| diningtable | 85.35 | 54.9 | 82.07 | 67.67 |
| dog | 83.18 | 45.8 | 78.93 | 65.07 |
| horse | 82.19 | 77.5 | 78.80 | 67.49 |
| motorbike | 80.51 | 64 | 76.80 | 68.11 |
| pottedplant | 67.87 | 36.3 | 64.34 | 59.72 |
| sheep | 80.13 | 44.7 | 78.75 | 73.34 |
| sofa | 70.64 | 50.9 | 70.16 | 64.89 |
| train | 86.96 | 79.2 | 83.48 | 71.61 |
| tvmonitor | 74 | 53.2 | 69.25 | 67.84 |
| Average | 76.86 | 56.93 | 73.67 | 66.26 |

Edge direction histogram encoded the distribution of the directions of edges. It was comprises a total of 73 bins, in which the first 72 bins were the number of edges with directions quantized at five degrees interval, and the last bin was the count of number of pixels that do not contribute to an edge.

Wavelet texture transform provided a multi-resolution approach for texture analysis.Wavelet transform performed on image involved recursive filtering and sub-sampling

TABLE 5.16: PASCAL Accuracy Comparison: Only Visual Features

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| aeroplane | 73.67 | 47.33 | 70.67 | 64.17 | 74.83 |
| bicycle | 66.17 | 54.67 | 60.17 | 53.17 | 62.5 |
| bird | 62 | 55.67 | 64.83 | 51 | 63.33 |
| boat | 65.67 | 52 | 62.67 | 57 | 61 |
| bottle | 56.33 | 48.67 | 55.33 | 52.17 | 61.83 |
| bus | 73.5 | 55.5 | 66 | 57.5 | 65.83 |
| car | 61.5 | 49.5 | 57.67 | 54.17 | 60.5 |
| cat | 68.33 | 55.83 | 63.83 | 59.5 | 67 |
| chair | 60.67 | 54 | 51.33 | 56.83 | 60.5 |
| cow | 68.83 | 51.5 | 61.33 | 60 | 63.67 |
| diningtable | 67 | 47.83 | 63.33 | 57.33 | 63 |
| dog | 64.5 | 47 | 60.83 | 54.67 | 64.83 |
| horse | 67.83 | 48 | 59.17 | 56.33 | 61.17 |
| motorbike | 66.33 | 57.17 | 66.67 | 55.67 | 62.67 |
| person | 59.17 | 52 | 52.33 | 51.33 | 55.33 |
| pottedplant | 59.33 | 53 | 55.17 | 52.5 | 56.5 |
| sheep | 72.83 | 50.17 | 67.67 | 61.5 | 67.17 |
| sofa | 63.83 | 52.17 | 56 | 52.5 | 63.5 |
| train | 70.5 | 49.5 | 65.17 | 57.33 | 63.33 |
| tvmonitor | 65.83 | 51.33 | 67 | 57.83 | 67.83 |
| Average | 65.69 | 51.64 | 61.36 | 56.13 | 63.32 |

Block-wise color moments, were, the first (mean), the second (variance) and the third order (skewness) color moments in the image. These have been found to be efficient and effective in representing the color distributions of images.Thus for the data set, they extracted the block-wise color moments over $5 \times 5$ fixed grid partitions, giving rise to a block-wise color moments with a dimension of 225.

For creating bag of visual words they followed following steps:

TABLE 5.17: PASCAL Accuracy Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| aeroplane | 74.5 | 69 |
| bicycle | 65.17 | 59.17 |
| bird | 74.83 | 68.33 |
| boat | 62.5 | 54.83 |
| bottle | 63.83 | 60.33 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| bus | 71.5 | 62.83 |
| car | 67.83 | 56 |
| cat | 73.5 | 68 |
| chair | 68.33 | 60 |
| cow | 69.33 | 62.83 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| diningtable | 75.67 | 68 |
| dog | 74.17 | 63.17 |
| horse | 73.67 | 64 |
| motorbike | 72.17 | 58.67 |
| person | 56.33 | 52.33 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| pottedplant | 61.67 | 56.17 |
| sheep | 74.67 | 64.5 |
| sofa | 65.67 | 62.17 |
| train | 78.83 | 68.5 |
| tvmonitor | 66.33 | 59.83 |
| Average | 69.53 | 61.93 |

- At first, they had applied the difference of Gaussian filter on the gray scale images to detect a set of key-points and scales respectively.
- In second step, they computed the Scale Invariant Feature Transform (SIFT) over the local region defined by the key-point and scale.
- In third step, they created the visual vocabulary by exploiting the k-means clustering. Here they generated 500 clusters, and thus the dimension of the bag of visual words is 500.

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 5.18: PASCAL Accuracy Comparison: Comparison of Published Results, Results of ensemble and Best of social and visual features

| Labels | Ensemble Method | Social Only | Visual Only |
|---|---|---|---|
| aeroplane | 78.46 | 74.50 | 74.83 |
| bicycle | 69.82 | 65.17 | 66.17 |
| bird | 75.22 | 74.83 | 64.83 |
| boat | 67.85 | 62.50 | 65.67 |
| bottle | 67.69 | 63.83 | 61.83 |
| bus | 77.38 | 71.50 | 73.50 |
| car | 69.77 | 67.83 | 61.50 |
| cat | 74.07 | 73.50 | 68.33 |
| chair | 72 | 68.33 | 60.67 |
| cow | 73.17 | 69.33 | 68.83 |
| diningtable | 75.81 | 75.67 | 67.00 |
| dog | 77.9 | 74.17 | 64.83 |
| horse | 76.7 | 73.67 | 67.83 |
| motorbike | 73.74 | 72.17 | 66.67 |
| person | 59.85 | 56.33 | 59.17 |
| pottedplant | 61.79 | 61.67 | 59.33 |
| sheep | 74.85 | 74.67 | 72.83 |
| sofa | 68.96 | 65.67 | 63.83 |
| train | 80.1 | 78.83 | 70.50 |
| tvmonitor | 67.97 | 66.33 | 67.83 |
| Average | 72.16 | 69.53 | 65.69 |

**Observations**

- The classification based on visual only features gives an average precision of 74.63% which is better by the large margin of 10.13% from the published results in Chua [2009].

- The classifications based on social only features (with LSI pre-computation) gives an average precision of 90.56% outperforming the results obtained from the visual

only features by a margin of 15.93%. This improvement over only visual features is beacuse of rich social meta-data. The NUS data set favors the highly popular images and we see that it has the highest tag vs image ratio of 19.4. This ample amount of auxiliary data apart from visual features improvises the enhancement over visual only descriptors.

- ensemble of all the features and ensemble classification leads to an average precision of 93.61%, which is better than the precision of social only features with a 3.05% margin. This result is way better than the baseline results mentioned in Chua [2009] and exceeds them by a margin of 29.11% in precision.

- The classifications based on visual only features give an average accuracy of 74.84%.

- The classifications based on social only features (with LSI pre-computation) gives an average accuracy of 88.61% outperforming the results obtained from the visual only features by a margin of 13.77%.

- Ensemble of all the features leads to the average accuracy of 91.23%, which is better than accuracy of social only features by a 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 in these images, compared to social data.

Table 5.19: NUS Precision Comparison: Only Visual Features

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| animal | 70.44 | 73.46 | 51.62 | 66.21 | 57.18 |
| coral | 83.1 | 79.45 | 53.11 | 65.22 | 69.23 |
| dancing | 74.07 | 75.59 | 59.13 | 67.39 | 63.3 |
| harbor | 77.58 | 81.4 | 53.22 | 79.32 | 72.4 |
| military | 79.28 | 70.67 | 56.42 | 61.35 | 65.02 |
| mountain | 72.42 | 84.3 | 54.58 | 78.03 | 78.62 |
| snow | 66.07 | 69.11 | 56.66 | 60.82 | 64.92 |
| statue | 73.67 | 57.66 | 54.79 | 59.57 | 52.99 |
| tattoo | 71.95 | 83.43 | 58.64 | 70.74 | 71.13 |
| temple | 69.59 | 64.42 | 54.15 | 57.92 | 58.67 |
| waterfall | 82.59 | 86.45 | 54.65 | 76.68 | 76.56 |
| wedding | 76.31 | 66.43 | 54.44 | 56.91 | 61.5 |
| AVERAGE | 74.76 | 74.36 | 55.12 | 66.68 | 65.96 |

Table 5.20: NUS Precision Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| animal | 90.01 | 70.44 |
| coral | 94.82 | 83.1 |
| dancing | 92.82 | 74.07 |
| harbor | 92.24 | 77.58 |
| military | 93.07 | 79.28 |
| mountain | 89.26 | 72.42 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| snow | 87.54 | 66.07 |
| statue | 84.56 | 73.67 |
| tattoo | 86.81 | 71.95 |
| temple | 89.72 | 69.59 |
| waterfall | 95.88 | 82.59 |
| wedding | 89.98 | 76.31 |
| AVERAGE | 90.56 | 74.76 |

TABLE 5.21: NUS Precision Comparison: Comparison of Published Results, Results of ensemble and Best of social and visual features

| Labels | Ensemble Method | NUS Best Results | Social Only | Visual Only |
|---|---|---|---|---|
| animal | 94.09 | 83.46 | 90.01 | 73.46 |
| coral | 95.45 | 45.73 | 94.82 | 79.45 |
| dancing | 95.98 | 53.66 | 92.82 | 75.59 |
| harbor | 93.63 | 36.47 | 92.24 | 81.40 |
| military | 97.86 | 55.72 | 93.07 | 71.94 |
| mountain | 90.05 | 34.23 | 89.26 | 84.30 |
| snow | 92.32 | 42.43 | 87.54 | 69.11 |
| statue | 88.56 | 54.63 | 84.56 | 59.57 |
| tattoo | 88.92 | 56.24 | 86.81 | 83.43 |
| temple | 93.68 | 43.78 | 89.72 | 64.42 |
| waterfall | 98.58 | 57.38 | 95.88 | 86.45 |
| wedding | 94.22 | 72.34 | 89.98 | 66.43 |
| AVERAGE | 93.61 | 53.01 | 90.56 | 74.36 |

Table 5.22: NUS Accuracy Comparison: Only Visual Features

| Labels | Visual Feature Based Classification | | | | |
|---|---|---|---|---|---|
| | HOG-LBP | SIFT | GIST | COLOR | GCM |
| animal | 72.86 | 50.54 | 66.07 | 56.96 | 64.29 |
| coral | 79.63 | 52.59 | 65.56 | 68.33 | 67.41 |
| dancing | 75.58 | 58.85 | 67.88 | 62.88 | 60.96 |
| harbor | 82.96 | 53.15 | 80.37 | 72.59 | 73.15 |
| military | 69.81 | 55.74 | 61.11 | 64.26 | 70.93 |
| mountain | 85 | 53.33 | 78.33 | 76.85 | 71.85 |
| snow | 70.19 | 47.59 | 60.93 | 64.63 | 65 |
| statue | 58.15 | 54.07 | 59.26 | 52.59 | 55.93 |
| tattoo | 84.07 | 57.04 | 70.93 | 71.67 | 77.41 |
| temple | 65 | 53.52 | 58.15 | 58.52 | 57.96 |
| waterfall | 86.67 | 53.7 | 75.93 | 76.48 | 84.07 |
| wedding | 65.93 | 53.33 | 56.67 | 60.74 | 60.93 |
| Average | 74.65 | 53.62 | 66.76 | 65.54 | 67.49 |

Table 5.23: NUS Accuracy Comparison: Social Features based on LSI and RP Methods

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| animal | 88 | 70.17 |
| coral | 92.33 | 80.83 |
| dancing | 92 | 74.33 |
| harbor | 90.83 | 75.83 |
| military | 90.83 | 77 |

| Labels | Social Features | |
|---|---|---|
| | LSI | RP |
| mountain | 86.67 | 72.67 |
| snow | 86.5 | 66 |
| statue | 84.67 | 72.5 |
| tattoo | 85.17 | 71.33 |
| temple | 87.67 | 68.83 |
| waterfall | 93.83 | 81.17 |
| wedding | 84.83 | 74.67 |
| Average | 88.61 | 73.78 |

TABLE 5.24: NUS Accuracy Comparison: Comparison of Published Results, Results of ensemble and Best of social and visual features

| Labels | Ensemble Method | Social Only | Visual Only |
|---|---|---|---|
| animal | 91.62 | 88.00 | 72.86 |
| coral | 95.99 | 92.33 | 79.63 |
| dancing | 92.39 | 92.00 | 75.58 |
| harbor | 93.02 | 90.83 | 82.96 |
| military | 94.69 | 90.83 | 70.93 |
| mountain | 90.55 | 86.67 | 85.00 |
| snow | 88.44 | 86.50 | 70.19 |
| statue | 85.23 | 84.67 | 59.26 |
| tattoo | 88.83 | 85.17 | 84.07 |
| temple | 91.5 | 87.67 | 65.00 |
| waterfall | 93.98 | 93.83 | 86.67 |
| wedding | 88.57 | 84.83 | 65.93 |
| Average | 91.23 | 88.61 | 74.65 |

# Chapter 6

# Conclusion and Future Work

In this work, we have demonstrated that if we use social networking meta data related to images present on social networks like Flickr, Instagram, Facebook, Google Plus etc. We achieve much better image classification results. We started with 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 a weighted combination of these feature models. This analysis shows that for image classification problems 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 images to feature vectors. We used text processing methods to extract the semantic structure of these social feature vectors in low dimensional spaces. These social feature vectors gave better classification for most labels 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 similar to textual information about the image. The comments, the tags, the groups and the galleries of an image has direct textual information of that image class. Therefore, correlating this textual information with the class of the image is not so complex. Where as in case of visual features, we have to do a lot of processing to get features, which embody some visual property like texture, objects, color

pattern and spatial envelope. But none of these properties directly tell the exact label of the image. In contrast in social meta data we might have the label as a tag or as a word in a comment.

Several machine learning techniques were explored for analysis and classification including Spatial Pyramid, Bag of Words Model, Latent Semantic Indexing and LibSVM. The results from 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 social- meta data gives that extra information, which proves to be instrumental in image classification. The fusion of social features and visual features enhances the accuracy of image classification in all cases that we examined and sometimes the difference is very significant.

## 6.1 Future Work

Some of the labels in our data-set did not have enough images like fox, lake, mountain, protest, earthquake etc. In such cases, when the number of images available were less then 50, we discarded that label for classification because we did not have enough positive instances. In such cases we can try 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 used classifiers 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 techniques.

We have currently used MATLAB and Python in our implementation. This implementation was done on single thread on a local system. We can change our implementation to faster platforms C++, with parallel processing for computationally difficult parts of

feature extraction. This can considerably speed up model creation. We used the approach of weighted features, we could use better fusion techniques to ensemble all the features.

Our social feature vector will be efficient only if the social meta data of an 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 based on the network. This technique can also be used to retrieve images from social networking sites on the basis of linkages. The system can be extended to do automated tag recommendations.

# Bibliography

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.

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.

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`.

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`.

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`.

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`.

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`.

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`.

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

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`.

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.

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`.

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`.

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.

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.

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.

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.

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.

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.

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.

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`.

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

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

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

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

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

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.

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

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.

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

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

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

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

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`.

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

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

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`.

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.

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`.

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.

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.

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, 2009a.

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`.

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.

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

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`.

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

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.

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

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`.

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`.

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

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

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`.

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

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

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

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.

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`.

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

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.

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 2009b. doi: 10.1109/CVPR.2009.5206816.

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`.

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`.

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`.