UNIVERSITY OF CALIFORNIA,
IRVINE

Event Recognition in Photo Streams

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Networked Systems

by

Shonali Balakrishna

Thesis Committee:
Professor Ramesh Jain, Chair
Professor Nalini Venkatasubramanian
Professor Athina Markopoulou

2016

DEDICATION

To my parents,
who are always a boundless source of love, inspiration and wisdom.

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **vi**

ACKNOWLEDGMENTS **vii**

ABSTRACT OF THE THESIS **viii**

1 Introduction **1**

2 Related Work **6**
2.1 Current Technology Review 6
2.2 Literature Review 8
2.2.1 Event Based Image Organization 9
2.2.2 Clustering 10
2.2.3 Probabilistic Learning 11

3 System Model **15**
3.1 Preliminaries 15
3.1.1 Dataset 15
3.1.2 Event Modeling 16
3.2 Components 17
3.2.1 Features 17
3.2.2 Clustering 17
3.2.3 Probabilistic Models 18
3.3 Overall Model 18

4 Feature Modeling **19**
4.1 Spatio-Temporal Metadata 19
4.2 Derived Metadata 20
4.2.1 Location tags 20
4.2.2 Concept tags 21
4.2.3 Temporal tags 23

5 Clustering **24**
5.1 Hierarchical Clustering 25
5.2 Modularity Optimization 26
5.3 Flat Clustering 27

6 Probabilistic Learning **29**
6.1 Naive Bayes 31
6.2 Conditional Random Fields 33

7 Experiments and Evaluation **35**
7.1 Clustering 35
7.2 Probabilistic Learning 38
7.3 Discussions 42

8 Conclusion **44**
8.1 Future Work 45

Bibliography **46**

LIST OF FIGURES

Page

2.1 Image Search and Organization with Google Photos 7

3.1 System Model 16

4.1 Location tags for Sightseeing Event label 20
4.2 Correlation matrix of concept tags and event labels 21
4.3 Concept tags for Sightseeing Event label 22
4.4 Time tags for Sightseeing Event label 23

7.1 Comparison of Hierarchical Clustering for various distance metrics and attributes 36
7.2 Comparison of Purity, ARI, AMI for Hierarchical Clustering and Modularity Optimization 37
7.3 Naive Bayes Precision-Recall for individual event labels 38
7.4 Confusion matrix for Naive Bayes Event Prediction 39
7.5 CRF Precision-Recall for individual event labels 40
7.6 CRF Average Precision-Recall comparison 40
7.7 General Comparison of Naive Bayes and CRF 41
7.8 F1-Measure Comparison of Naive Bayes and CRF for individual event labels 42

LIST OF TABLES

Page

5.1 An example cluster with spatio-temporal attributes 25
5.2 Clustering of an example Photo Stream into three clusters 28

6.1 Event Prediction for an example cluster 30

ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Ramesh Jain, for being a constant source of encouragement and inspiration. His excellent guidance and sharp insights were invaluable to my thesis.

I am indebted to Dr. Setareh Rafatirad for her unfailing guidance, which has hugely influenced my research direction.

I am sincerely grateful to my committee members, Professor Nalini Venkatasubramanian and Professor Athina Markopoulou, for their thoughtful feedback and support throughout this process.

I would like to thank my fellow lab members for their timely help and support during the course of my thesis.

I would like to express my heartfelt gratitude to my family and friends, for their encouragement and love over the years.

ABSTRACT OF THE THESIS

Event Recognition in Photo Streams

By

Shonali Balakrishna

Master of Science in Networked Systems

University of California, Irvine, 2016

Professor Ramesh Jain, Chair

Visual information is now an increasingly important part of the data ecosystem. While photos are captured and shared widely over the internet, methods for image search and organization remain unintuitive. The current state of the art in image organization extracts people, objects, location and time information from images and provides options for the automatic grouping of photos based on one of these attributes. As a next step in image organization, we believe that images need to be grouped based on the events they represent. However, such an event-based grouping of images is presently very primitive and imprecise.

In this thesis, we present an event-based image organization approach, where the key-idea is to leverage visual concepts and spatio-temporal metadata of images in order to automatically infer their representative event; this approach combines clustering with probabilistic learning methods. Clustering is performed on images based on spatio-temporal metadata, where each cluster represents an event that occurred at a particular spatio-temporal point/region. We build probabilistic models that learn the associations between the different features and the predefined event labels of each cluster, and use the learned models to automatically infer the events in incoming photo streams. We evaluate efficiency of the proposed method by using a personal image data set, using metrics such as precision and recall.

The contributions of this thesis are two-fold. First, we use several web based sources for semantically augmenting the spatio-temporal metadata corresponding to the images, and second, we combine clustering with probabilistic learning to identify and annotate events in the photo stream, using the augmented metadata and visual image concepts.

Chapter 1

Introduction

Visual information is now an increasingly important part of the data ecosystem. The Web has become exceedingly visual, with the wide proliferation of photos and videos from phone cameras. To put numbers in perspective for photos, in 2015 alone, for every minute on average, Instagram users liked more than 1.7 million photos, Pinterest pinners pinned nearly 10,000 images and Facebook users uploaded nearly 136,000 photos [1, 3].

While photos are captured and proliferated widely on social media websites and over the internet in general, the methods for access and retrieval of desired images are still limited. It is difficult to relive memories through photographs due to the lack of automatic image organization and efficient image retrieval methods. Traditional image organization methods of manually adding metadata (like captions and keywords) to photo collections is laborious and time consuming. There is a pressing need for a means to automatically organize and annotate such photo collections.

The current state of the art in image organization extracts people, objects, location and time information from images and provides options for automatic grouping of photos based on one of these attributes. Photo organization features in Google Photos, Apple's Photos, Flickr use these attributes for grouping photos. Such a grouping of images along any one of these dimensions provides a good first step in image organization; but these groupings are unintuitive as the groups are based on only one of the attributes at any given time and are hence not representative of the event they capture.

As a next step in image organization, we believe that images need to be grouped based on the events they represent. In this context, an event answers questions such as: where (the place), when (the time), who (the people involved), what (the type of event), how (the attributes that characterize the event) [50] and hence, is a broad encapsulation of all of these attributes of location, time, people and concepts. Events are an effective grouping mechanism, because humans organize their memories around concrete events that they have experienced [51]; hence events are semantically connected to the information that multimedia captures. This sort of event-based organization is also known to influence how people organize their personal media [44]. However, such an event-based grouping of images is presently very primitive and imprecise. Leveraging visual concepts to infer and annotate events will take image organization to the next level, by grouping images into event-based semantically meaningful collections.

In this thesis, we present an event-based image organization approach that combines clustering with probabilistic inference, by exploiting image metadata and image visual concepts. Events are an efficient paradigm for unified multimedia indexing, and provide a powerful framework for multimedia search and retrieval [50]. Moreover, events can be key cues of identification for image search and retrieval in an image corpus [34]. In this context, personal photographs have become a rich source of information to keep track of the events occurring in a person's life. Our goal in this thesis is to automatically classify photo streams with high level events (like *lunch*, *entertainment*).

Our approach combines contextual spatio-temporal metadata - including augmented location categories and time of day tags - with visual concepts, to infer the event represented by a cluster of images. It has been proven that an approach integrating both contextual metadata and visual content is effective in bridging the existing multimedia semantic gap in understanding what an image represents [30, 18]. Such an approach would be the first step in image search, querying based on context semantic event cues, by automatically annotating social media with events from public event directories [27] or by automatically identifying events that have occurred based on social media [7].

We perform clustering on the images in order to group them into semantically meaningful collections. Smartphones and modern digital cameras are equipped with various sensors and record rich contextual metadata like location, time and image EXIF parameters for the images captured on them. The importance of contextual metadata - especially time and location - in multimedia search is emphasized in [34]. Clustering based on both time and location has proven to be more effective, with clusters that are more accurate representations of events [15]. Therefore, for the purposes of clustering, we focus on the spatio-temporal attributes, given that photos from an event are generally taken in close proximity in time and location, with small variability. We define all photos captured at a particular spatio-temporal point/region to correspond to an event that has occurred. We cluster the images based on spatio-temporal EXIF metadata like latitude, longitude, and timestamp, to obtain event-based clusters.

We augment the contextual metadata with visual concepts, location information and time tags. Recent advances in deep learning have now made it computationally feasible to mine visual concepts from image content, making it possible for us to use a web-based deep learning API, ClarifAI [2], on the images in the dataset. Furthermore, we query a location web service, Foursquare [4], to obtain information about the nearest known places at the image location, ranging from restaurants to historic monuments. Moreover, we also use the temporal metadata in a modified form, where the local time in the specified timezone is classified into a time of day tag.

We predict the event represented by the image cluster by using these three key forms of image metadata - visual concept information, location categories and time of day attribute - derived from content, location and time, respectively. We use probabilistic models, such as Conditional Random Fields (CRF) [24] and Naive Bayes [14], to learn the correlations between these attributes and the corresponding event label. A probabilistic classifier predicts, given a sample input, a probability distribution over the set of output classes. We train probabilistic models to learn from the training data and the learned model is then used to predict the event labels given an image cluster.

As the number of images on our smartphones, hard disks, and cloud hosting accounts increase exponentially, methods for their efficient organization and access needs to scale up accordingly. Organization of images based on single attributes like time, location, people or objects is a competent first step in this direction. However, such an organization does not provide an intuitive image organization and search experience, due to the lack of holistic representation of life experiences in images. Events are an efficient paradigm for grouping of images, as it provides a natural abstraction of human experiences as captured by images. Being able to infer the high-level event a group of images represent is crucial in providing precise and detailed event annotations answering the 5 Ws of image search - what, where, when, who and why, eventually providing an event perspective for searching and browsing through image content. Event recognition, in this perspective, is a critical component in the larger framework of image organization/search/exploration, which enables mining of intelligent insights from visual data. Automatic event annotation in photo streams is an important research problem, since it provides an avenue not just for automatic image organization, but also for efficient indexing of images based on these annotated events, allowing for a search experience in images as effective as existing document search systems.

This thesis is organized as follows: In Chapter 2, we present, firstly, a review of the existing technology and features in image organization and secondly, a literature review of event-based image organization, clustering and probabilistic learning. In Chapter 3, we describe the overall system model and its components and explain the event and dataset modeling. In Chapter 4, we describe the feature modeling in detail and illustrate using some examples. In Chapter 5, we describe the clustering algorithms explored and our clustering approach along with example illustrations. In Chapter 6, we detail the probabilistic learning algorithms explored and our modeling for each algorithm. In Chapter 7, we present the experiments and evaluations conducted for our work, and discuss the findings. We conclude this thesis and address future work in Chapter 8.

Chapter 2

Related Work

We review firstly, the existing technology and features in image organization and secondly, relevant literature pertaining to event-based image organization, clustering and probabilistic learning. While most of the literature that is reviewed does not solve the exact research problem addressed in this thesis, it focuses on the individual components that we use in our model. The work discussed in this section does not synthesize these components into a comprehensive event-based image organization model, but it has in some form influenced our research. Therefore, we will survey the relevant research on each component to construct our system model.

2.1 Current Technology Review

We review the features in the current state of the art technology for image search and organization to better address the needs of the existing framework. We review the features in the three most commonly used image storage and organization applications: Google Photos, Apple's Photos and Flickr.

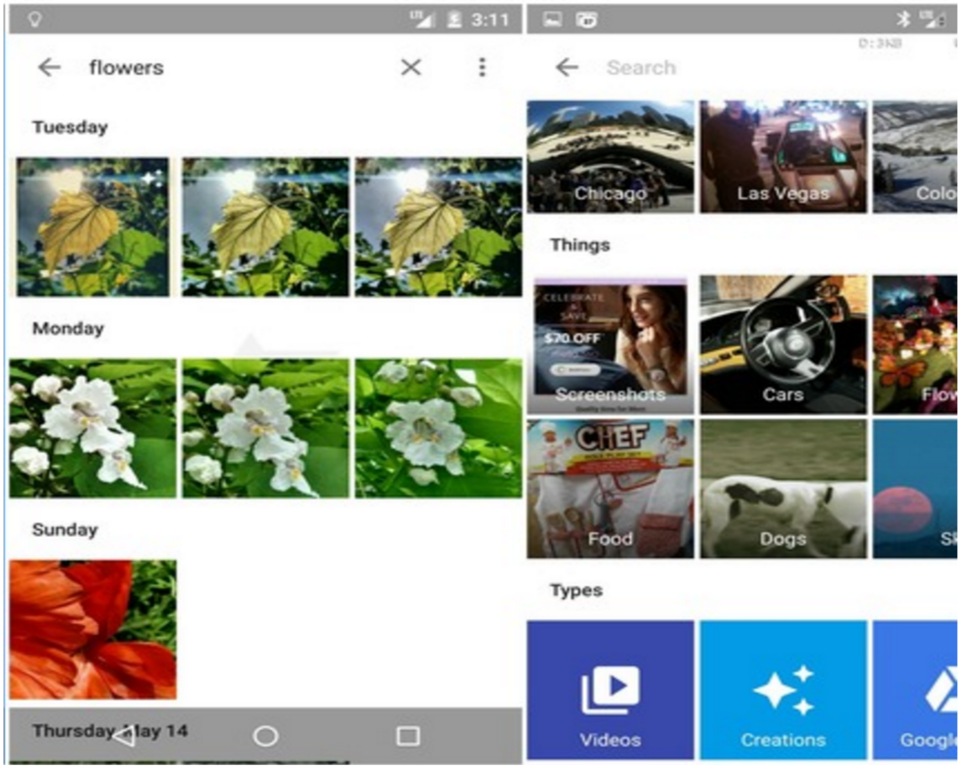


Figure 2.1: Image Search and Organization with Google Photos

Google Photos automatically organizes photos and makes them searchable; by default, photos are organized based on the time that it was captured. Once you click the search button, it provides options for searching based on people and places relevant to the user's photos. On selecting the Albums option, it displays groups of photos organized in three categories - People, Places and Things. In the People section, Google groups together all photos containing the face of a person, as identified by its face recognition system, making it possible to search for all photos using the name of the person. In the Places section, Google utilizes geo-tagging information wherever such data is available, and groups photos based on the GPS information. When this GPS information is not available, it detects known landmarks. In the Things section, it recognizes objects or concepts from images - for example, beaches or cars or sky or buildings or food or concerts, graduations, birthdays, to name a few - and organizes images based on these categories. Searching for any of these objects or concepts results in pictures with these objects being displayed. Discrepancies in recognition of people or objects do exist, but the system is reasonably accurate. Furthermore, while searching, it is possible to combine two attributes, for example, a search for 'snowstorm in Toronto' would display relevant results. But image organization in albums is based on only one of the attributes of time, location, people or objects, at a time. An example of image search and organization in Google Photos is displayed in Figure 2.1

Flickr is a photo social network and provides automatic tagging of images based on what is present in them, but this remains primitive for personal image collections. Camera Roll, for personal photographs, provides two options for automatic organization - chronological, based on date and time of upload, and Magic View, an organization based on a fixed number of content types ranging from people to landscapes to animals. However, for public collections of images, Flickr has powerful features for searching using objects or concepts. Search options include object attributes and filtering based on color, size, media type etc. There is also a map view to organize images based on location. No automatic face recognition options are present in Flickr - manual tagging by the user is required.

Apple's Photos is a photo management application developed by Apple. Photos organizes images chronologically into moments and also specifies the location wherever the GPS information is available. It also has limited functionality in using face recognition to group photos based on the people in them. Newer features include grouping based on the nature of the photos, such as selfies, etc or based on the application source, for example Instagram etc.

2.2 Literature Review

We discuss relevant literature related to event-based image organization, clustering and probabilistic learning and discuss how this work relates to our approach.

2.2.1 Event Based Image Organization

Studies in neuroscience have shown that humans remember their life experiences by structuring past experiences into events [51]. This event structure is also known to influence how people organize their personal media [44]. Events address questions of what, where, who, when and how and hence, are semantically connected to the information that multimedia captures [50].

Work has been done previously in detecting events in social media and relating these events to the media corresponding to them [27, 46, 41, 21, 28]. Towards building stories using online social media profiles, life events are detected by leveraging the spatio-temporal metadata of the images to build a relationship strength model for each user profile [46]. Events and their properties are automatically detected by temporally monitoring media shared on social media websites, and these events are used to structure online media and likewise, the media is used to illustrate events in event directories [28]. Event and location semantics are extracted from Flickr tags based on temporal and spatial distribution of tag usage, as derived from photo metadata [41]. EventMedia creates an infrastructure for unifying event centric information derived from event directories, media platforms, social networks using linked data technologies and integrates related media descriptions with event descriptions [21]. Events are integrated with the media corresponding to them by using linked data technologies for semantically enriching descriptions of both events and media, using a minimalist LODE ontology and Media ontology respectively; public event directories are scraped, semantic web technologies are used to provide a means for searching and browsing interlinked media collections using an event perspective [27].

Event recognition in images have been explored in past literature. Event and sub-event recognition for single images is performed using a pipeline of classifiers, each successively classifying the image into a event and a sub-event respectively [31]; furthermore, [31] uses time constraints with clustering in a Bag of Features classification approach to recognize sub-events, to further improve accuracy. A probabilistic graphical model is used with Variational Message Passing, to classify into 8 sports events using object and scene category features [25]. High level concepts are used as features for semantic event detection, with event-level Bag of Features representation for modeling events [19]. Temporal information and visual features such as scene and object classifier outputs (such as *indoor/outdoor*, *nature*, *sky*, *faces*, etc.) is used with a Bayesian belief network for event classification [12]; broad event classes like *Vacation*, *Party*, *Family* and *Sports* were used and features that are effective for each event label are computed manually for the dataset. A contextual meaningful hierarchy of events is built in [48] by utilizing simple contextual cues of time, space and visual appearance; Multi-modal clustering based on space, time and color distribution is performed.

Therefore, while past literature exists on utilizing image visual concepts or scene labels in event-based image organization and event detection, this work was done at a time when such visual concepts were primitive and limited in number and/or quality. More recent advances in deep learning have not yet been leveraged to this end, in order to obtain more accurate and extensive event annotations.

2.2.2 Clustering

Clustering has been widely used for grouping images using contextual metadata [11, 15, 29, 7, 40]. The importance of contextual metadata - especially time and location - in multimedia search is emphasized in [34]. Clustering based on both time and location has proven to be more effective, with clusters that are more accurate representations of events [15].

Clustering is used in [7] to identify events in social media documents(photos, videos etc), with contextual features in a weighted cluster ensemble algorithm. A variety of algorithms are used in [11] on time and/or image content attributes to cluster the images in order to group them based on events. Images are clustered in [15] based on spatio-temporal data to detect high level events that occurred in reconnaissance missions in scenarios where sensor data may be missing or unavailable. Images are clustered in [29], using temporal metadata and image content features derived from image understanding, for classifying into events. An algorithm for clustering of evolving data streams, using an online and offline clustering component is proposed in [6]. Agglomerative clustering is performed in [40] on image EXIF metadata as a first step and the finest possible sub-event represented by an image in a photo stream is extracted using image metadata, user information, ontological model and other web and external data sources; They use an event ontology, and provide flexibility in specifying models by using contextual information to augment the model on the fly.

Graph community detection algorithms are also a form of clustering and are more commonly called community detection. A detailed description is given in [35] about Modularity and the relation between Modularity Optimization and community partitioning. Modularity is used in a hierarchical manner in biochemical networks in [42]. A method to optimize the modularity of a graph is described in [8], which is the basis of our implementation of modularity. The metrics used to estimate the quality of a community partitioning are described in [10], which we have implemented for our clustering performance evaluation.

2.2.3 Probabilistic Learning

There are two kinds of probabilistic learning models - generative and discriminative. A generative classifier learns a model of the joint probabililty $P(x, y)$ of the inputs x and output labels y, and then makes predictions using the Bayes rule to compute the posterior $P(x|y)$. A discriminative model learns the conditional probability distribution or posterior $P(x|y)$ directly. A binary classification task is used in [20] to examine the conditions in which discriminative and generative classifiers are each effective; it proves that contrary to widely held notions of the superiority of discriminative classifiers over generative classifiers, there are distinct regimes where generative classifiers are known to perform better.

Naive Bayes

Naive Bayes is a generative probabilistic model based on the Bayes theorem and assumes that the features are independent given the label. Despite this unrealistic assumption, Naive Bayes has proven to be immensely successful in a number of domains. Towards measuring end user perceptions of performance in distributed systems, the labeling of a remote procedure call sequence with the correct transaction type is addressed using Naive Bayes, which works well with an accuracy of 87% [17].

An empirical study of the Naive Bayes classifier is performed in [43], proving the effectiveness of classification for both completely independent features and features with functional dependencies; it also shows that the accuracy of the Naive Bayes classifier is dependent on the amount of class information lost due to its independence assumptions. The optimality of the Naive Bayes and the reason for its success in classification is discussed in [52]. The classifier performs well even when dependencies exist among attributes, as classification depends on the distribution of the dependencies across all attributes - the dependencies might cancel each other out across labels and across atrributes. The accuracy of Naive Bayes is due to the usage of the zero one loss function which does not penalize inaccurate posterior probabilities as long as the class with the highest posterior is correctly estimated [13].

Two kinds of first order Naive Bayes probabilistic models exist for text-based classification [33] - One, a multivariate Bernoulli model, which is a Bayesian network with no dependencies between words and binary word features. A binary attribute vector indicating which words occur is created and probability is computed by multiplying attribute values. Second, a multinomial model, uses unigram language models with integer word counts. The text is represented by word occurrences with frequencies of words recorded. When computing the probability of the label, probabilities of word occurrences are multiplied.

Conditional Random Fields

CRF has been used in a variety of disparate domains for modeling the correlation between sequential data. CRF was first presented for the purpose of segmenting and labeling of sequential data [24] and has been successfully used in text processing applications, like named entity extraction [32] and shallow text parsing [47].

CRF has also been used previously for modeling the dependencies in sensor data for activity detection [26], in videos to infer semantic event labels [49] and for motion tracking in video sequences [45]. Hierarchical CRF is used in [26] for activity recognition, to learn patterns of human behaviour from sensor data to infer high level activities and places. CRF is used in [49], in videos to infer semantic event labels from multiple sequences to retrieve specific video segments.

Furthermore, CRF has been used in classification of image regions [23], image labeling [16] and object recognition in cluttered unsegmented images [39]. Image regions are classified in [23], by modeling the neighborhood dependencies in observed data as well as labels using CRF. CRF is applied to the image labeling problem in [16] with label features operating at different scales, and predictions of various features are combined multiplicatively. Object recognition in cluttered unsegmented images is done in [39] by modeling the assignment of parts to local features for each object class by a CRF.

This indicates that CRF is an effective framework for modeling the correlation between sequential data in a variety of disparate domains. With regards to contextual image metadata, CRF has been used in finding correlations between event labels, images' visual content through scene labels, GPS and time [9]; But [9] differs from our work in that they employ single scene labels to create a multi-level annotation hierarchy, while we propose a flat prediction approach using augmented metadata based on temporal, spatial and multiple visual concepts, derived from external web sources.

Therefore, in summation, while some work has been done in the past on event detection and annotation using clustering and probabilistic methods, this work was conducted at a time when the spectacular advances in deep learning that we've seen over the past few years had not yet happened. Hence, highly accurate and descriptive visual concepts derived from images have not been leveraged for event recognition, which is the key novelty of our work.

Chapter 3

System Model

We discuss the overall system model, its various components and the role they play in this chapter. We first discuss the preliminaries - the dataset that we use for the evaluation of this approach, and how it ties into the modeling of the high-level events that we have worked with. Next, we discuss each component and the part it plays in the overall model.

3.1 Preliminaries

The preliminary modeling in this thesis includes modeling of the dataset used and the modeling of the events. We discuss each of these in detail below.

3.1.1 Dataset

We use a personal dataset to evaluate the performance of our algorithm, containing 392 clusters, and manually labeled with 7 predefined event classes as ground truth. In order to aid the efficient labeling of ground truth for events, an event labeling web interface was

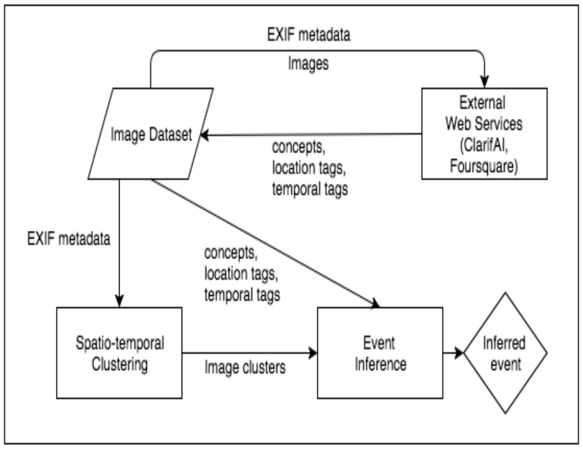


Figure 3.1: System Model

created. This dataset is primarily a travel dataset, captured over a period of two years, over various locations across the world. This dataset includes image EXIF metadata such as latitude, longitude, localtime and timezone. Furthermore, several external sources, such as Foursquare and ClarifAI, have been used to augment this dataset with attributes of visual image concepts, time of day tags and foursquare location properties.

3.1.2 Event Modeling

As a first step in the direction of event annotation in images, we have worked with 7 high level events suited to our dataset. These seven event labels are: *Breakfast*, *Lunch*, *Dinner*, *Meeting*, *Sight Seeing*, *Urban Walk*, *Entertainment*.

Breakfast, *Lunch* and *Dinner* correspond to images taken during each of the three meals during the day. *Meeting* relates to images taken around a large gathering of people, typically in a formal work-oriented setting or a family gathering; *Sightseeing* relates to images taken during travel, from outdoor landscapes to historic monuments in cities; *Urban walk* are images taken within the city, on an everyday basis, while walking, driving, etc; *Entertainment* relates to images taken during leisure or entertainment, either at the beach, resort, mall etc.

3.2 Components

The overall system model has been illustrated in Figure 3.1. Each of the components of the system model are elaborated on below, and the overall workflow is explained.

3.2.1 Features

The time of capture of images (in local time, along with the corresponding time zone) and the location of capture(in latitude and longitude) is available to us through the image metadata. Besides the features available to us through the image metadata, we have used web-based external sources to augment the dataset. We have extracted the most important visual concepts present in the images by using a deep learning API (ClarifAI [2]) on the image content. We have also obtained the closest places(using Foursquare [4]) in proximity with the longitude and latitude of the image cluster, and used the broad location category of these places as a means of classification of location information. We have also converted the time information into categorical data buckets, thus classifying the time into the time of day when it was captured. These three features have been added to the dataset, and have been primarily used in the prediction model.

3.2.2 Clustering

We have explored multiple clustering algorithms - agglomerative clustering, a graph community detection algorithm called modularity optimization and a flat clustering algorithm for streaming data - for finding an optimal grouping of images, based on image spatio-temporal EXIF metadata. Since photos have been treated as an incoming stream in our work, for the prediction phase we use a flat clustering approach for streaming data [6] to obtain clusters. We cluster based on location(latitude and longitude) and time information.

3.2.3 Probabilistic Models

The nature of our dataset is such that the attributes of an image are probabilistic, with multiple concepts and location categories, each of which might be noisy, uncertain or irrelevant to the cluster and event label that it represents. Therefore, we use probabilistic learning models to model the relationships between the attributes and the event labels. We have explored probabilistic learning methods such as Conditional Random Fields[24] and Naive Bayes [14] to perform structured prediction of event labels.

3.3 Overall Model

The overall system is shown in Figure 3.1. Image content and Image EXIF metadata is sent to External Web Services like ClarifAI and FourSquare. The corresponding visual image concepts and location categories are returned from ClarifAI and FourSquare respectively and are augmented to the image metadata. The spatio-temporal EXIF metadata is sent to a clustering block, where it is clustered according to location and time. The concepts, location categories, time of day attributes and the resultant image clusters are passed to the Event Prediction subsystem. At the Event Prediction block, a learning model was previously trained on labeled cluster data containing attributes of concepts, location categories, time of day and labeled with event tags. This learned model is used to predict the label of the input image clusters and return the predicted event label.

Chapter 4

Feature Modeling

We first discuss the raw spatio-temporal metadata that is available to our dataset, and then discuss the features that we have derived and modeled from this raw data - concept tags, location tags and temporal tags.

4.1 Spatio-Temporal Metadata

Temporal information, in the context of the capture of images, is available to us through the Image EXIF metadata that accompany images taken through digital cameras or mobile phones. Spatial metadata similarly accompany images captured, through the use of GPS sensors which record location information. As such, the timestamp(including timezone information), latitude, longitude at which the image was captured is the raw data that is available to us through the image dataset. This data is commonly present for all images taken on digital cameras or smartphones. We use this raw spatio-temporal data for the clustering phase of our approach.

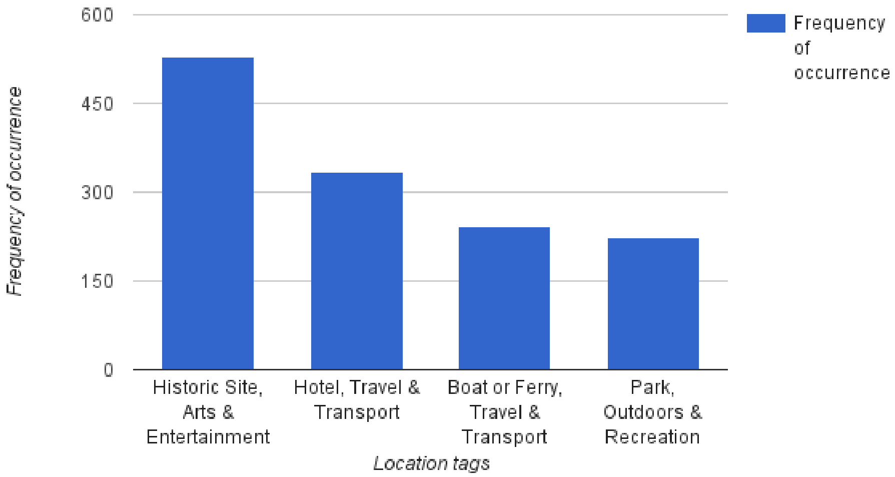


Figure 4.1: Location tags for Sightseeing Event label

4.2 Derived Metadata

Using the raw spatio-temporal metadata and the images themselves, we derive our main features for the prediction phase of our approach. We use the temporal data to derive time of day tags, the location information to derive the location tags and the image content to glean the visual concepts.

4.2.1 Location tags

Foursquare [4], a local search and discovery service, has been queried with the latitude and longitude of the images, to obtain further information on the places on Foursquare closest to the specified latitude and longitude. We obtain detailed information like name, address, distance from the queried location etc about each of these possible places. Among these attributes, we use the location category label, which ranges from *restaurant* to *historic monument*, as a way to classify the image locations into a category. The broad location category of these places is used as a means of classification of location information. This

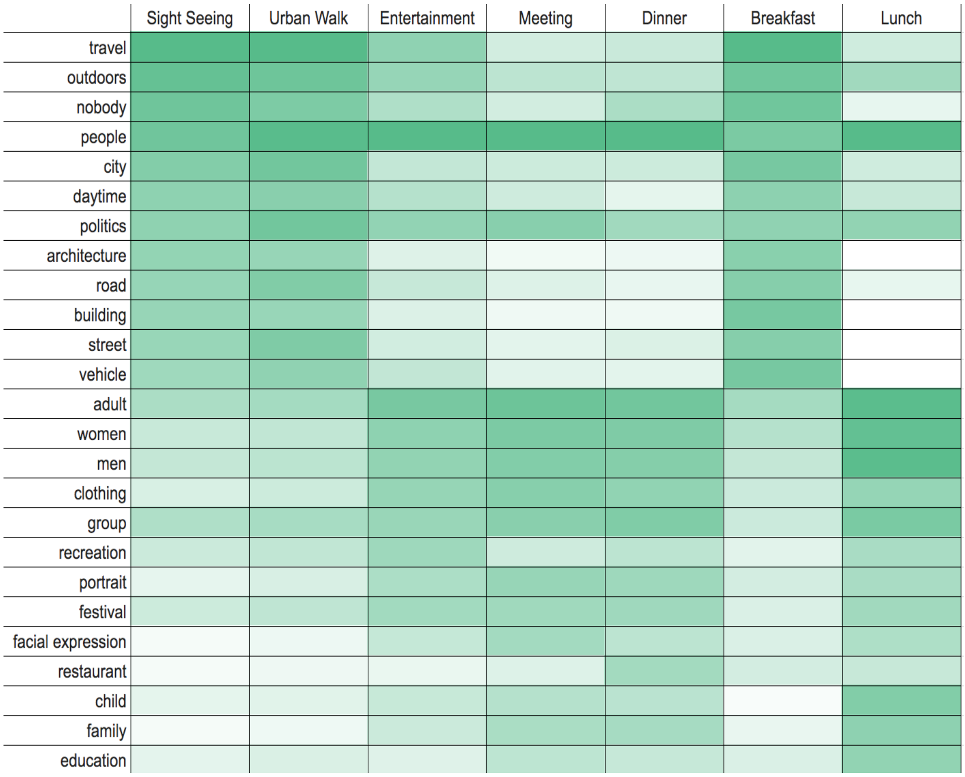


Figure 4.2: Correlation matrix of concept tags and event labels

information provides us with context on the location and hence the possible event that occurred at the location.

We further analyze the location tags by using event labels. As illustrated in Figure 4.1, location categories like *Historic Site, Arts and Entertainment*, *Travel and Transport*, *Outdoors and Recreation* are the most frequently occurring for the *Sight Seeing* label.

4.2.2 Concept tags

ClarifAI [2], a web based deep learning API, has been used to extract multiple visual concepts from each of the images, each concept returned with a probability of occurrence in the image. ClarifAI detects objects and concepts representing the content of the images. The ClarifAI vocabulary is extensive and able to detect over 11,000 different things in over 20 languages

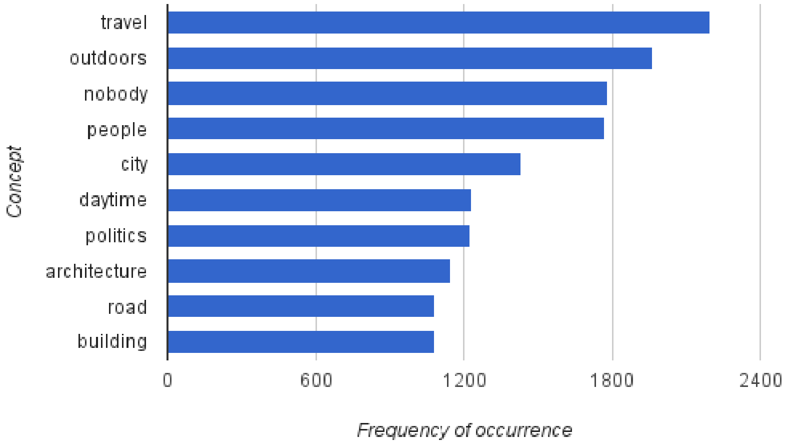


Figure 4.3: Concept tags for Sightseeing Event label

[2]. Concepts range from *outdoors* to *windows* to *family* to *politics* and are an indicator of the objects and semantic representation of the image. Such information is useful in that it expresses the key concepts present in the image content in text, thus allowing us to leverage the powerful advances in deep learning over the recent years to solve the image organization problem.

We further analyze the concepts that we work with by using event labels. A correlation matrix depicting the different concepts and their correlations with event labels is displayed in Figure 4.2. The intensity of the green indicates the frequency of occurrence of each concept for the corresponding event label, with darker shades indicating a higher frequency.

We further elaborate on concept tags by using an example event label - *Sight Seeing*. As evident in Figure 4.3, visual concept tags like *travel*, *outdoors*, *people*, *nobody*, *city*, *daytime*, *architecture* occur very frequently for the *Sight Seeing* event label.

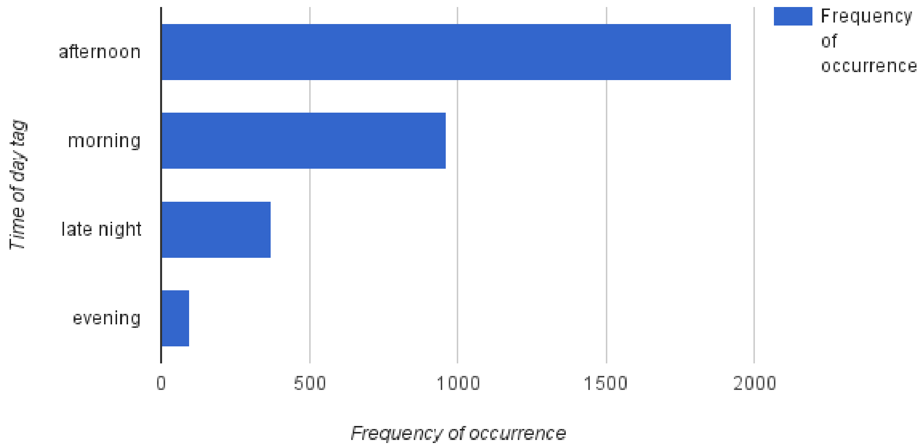


Figure 4.4: Time tags for Sightseeing Event label

4.2.3 Temporal tags

From the temporal metadata of the image, the local time and timezone has been used to classify it into a time of day tag - either *morning*, *afternoon*, *evening* or *late night* based on the time range it falls under. The time information is therefore converted into categorical data buckets, thereby classifying the time information into the time of day when it was captured. This information is useful, as there exists correlations between events and the time of day that it occurs during. For example, *Meeting* is likely to occur in the *morning* or *afternoon*, *Breakfast*, *Lunch* and *Dinner* are likely to occur in the *morning*, *afternoon* and *evening* respectively. By converting the time information into these tags, we are able to leverage these correlations.

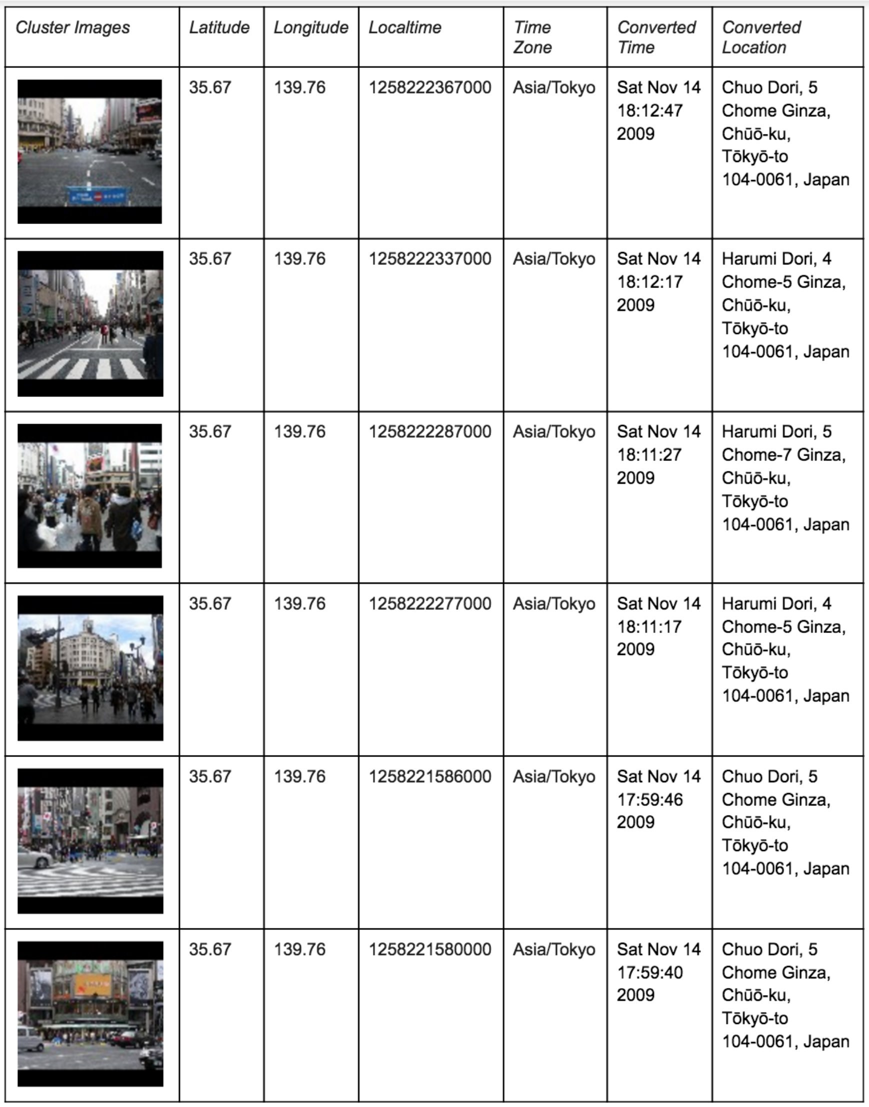
We further analyse the time of day tags that we have derived by using event labels. As illustrated in Figure 4.4, most of the *Sight Seeing* activities occur during the afternoon, indicated by the high frequency of occurrence of the *afternoon* time of day tag in the training data.

Chapter 5

Clustering

We discuss the clustering algorithms that were explored in our work - Agglomerative Clustering, Modularity Optimization and Flat Online Clustering for photo streams. The goal is to cluster images into groups, where all images in the group are similar in terms of a similarity metric, and all dissimilar images are grouped in different clusters. The distance metric, in our case, computes the similarity of spatial and temporal values of captured images.

We illustrate the general attributes used by the clustering algorithm using an example resultant cluster, as shown in Table 5.1. We illustrate the general working of our clustering algorithm through Table 5.2, which displays a photo stream that was grouped into three clusters. As seen in Table 5.1 and 5.2, images that are similar (or close together) in latitude, longitude and localtime are grouped together. The location in terms of address, and time in terms of date is also displayed for easy comprehension.



Cluster Images	*Latitude*	*Longitude*	*Localtime*	*Time Zone*	*Converted Time*	*Converted Location*
	35.67	139.76	1258222367000	Asia/Tokyo	Sat Nov 14 18:12:47 2009	Chuo Dori, 5 Chome Ginza, Chūō-ku, Tōkyō-to 104-0061, Japan
	35.67	139.76	1258222337000	Asia/Tokyo	Sat Nov 14 18:12:17 2009	Harumi Dori, 4 Chome-5 Ginza, Chūō-ku, Tōkyō-to 104-0061, Japan
	35.67	139.76	1258222287000	Asia/Tokyo	Sat Nov 14 18:11:27 2009	Harumi Dori, 5 Chome-7 Ginza, Chūō-ku, Tōkyō-to 104-0061, Japan
	35.67	139.76	1258222277000	Asia/Tokyo	Sat Nov 14 18:11:17 2009	Harumi Dori, 4 Chome-5 Ginza, Chūō-ku, Tōkyō-to 104-0061, Japan
	35.67	139.76	1258221586000	Asia/Tokyo	Sat Nov 14 17:59:46 2009	Chuo Dori, 5 Chome Ginza, Chūō-ku, Tōkyō-to 104-0061, Japan
	35.67	139.76	1258221580000	Asia/Tokyo	Sat Nov 14 17:59:40 2009	Chuo Dori, 5 Chome Ginza, Chūō-ku, Tōkyō-to 104-0061, Japan

Table 5.1: An example cluster with spatio-temporal attributes

5.1 Hierarchical Clustering

Hierarchical clustering seeks to build a hierarchy of clusters, and is useful when such hierarchies naturally exist in the data. There are two kinds of hierarchical clustering algorithms - agglomerative and divisive. Agglomerative clustering is a bottom up approach, where each observation begins in a cluster of its own, and clusters are iteratively merged based on a distance metric, until one big cluster is formed. The resulting dendrogram is sliced at the desired level of the hierarchy to obtain the corresponding clusters. Divisive clustering is a top down approach, where all observations begin in one big cluster and splitting is performed recursively based on a metric, until you reach the end of the hierarchy.

In order to decide which clusters to merge together, a distance metric is used infer the measure of dissimilarity or of distance between the two clusters. This distance metric could be Euclidean, Manhattan, Hamming etc. There are also several linkage criteria based on which clusters are merged together - either the minimum distance (single linkage) between any two points in the two clusters, or the maximum distance (complete linkage), or the averaged distance (average linkage).

Single linkage agglomerative clustering was performed for our work. Clusters were obtained based on minimum distance between any two points in the two clusters. Distance metrics like hamming, jaccard and euclidean were used and evaluated against each other. Hierarchical clustering was evaluated for a combination of features in the dataset - Image EXIF attributes like time, location, focal length, flash, ISO and exposure time. The number of clusters was set manually, to the optimal number.

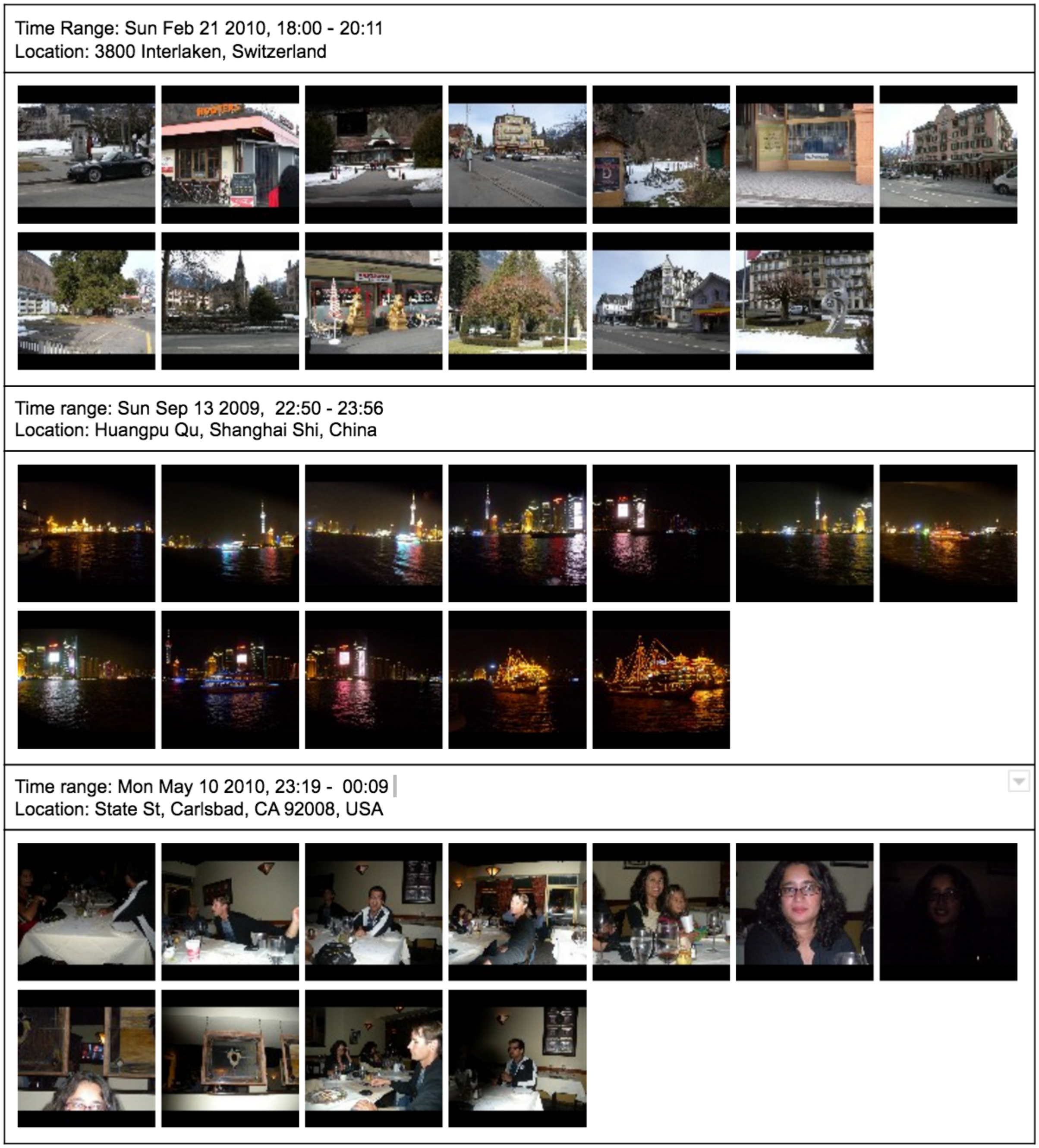
5.2 Modularity Optimization

Modularity optimization is a graph community structure detection method. Louvain method using the Community API [8] has been implemented in our work. The working of this algorithms is as follows: Each node starts in a community of its own. Phase 1: For every node, find a neighbour whose community assignment maximizes the modularity of that node. Assign the node with the community of the neighbour which maximizes modularity, only if the gain is positive. If the modularity gain is not positive, the community assignments remain unchanged. Each node can be considered multiple times, until a local modularity maxima is reached. Phase 2: Build a network whose nodes are the communities detected in phase 1, and the links between these nodes are assigned a weight given by the sum of weights of all links between these two communities. These two phases are repeated iteratively until convergence is reached, wherein the maximum modularity is obtained and no further gain is possible.

We represent the spatio-temporal data from our image dataset as a graph and implement community detection algorithms on this graph, so as to effectively cluster them based on their spatio-temporal attributes. Spatio-temporal (time, latitude, longitude) metadata was used to model the graph, where the nodes represent images and edges are drawn between nodes when two images have their spatio-temporal distance within a threshold (for example, 7 days, 13km). The constraint over the edges was varied for optimal results. In addition, weights are assigned to edges based on the magnitude of spatio-temporal distance between nodes - more the distance, lesser the weights. In order to model the inverse relationship between weights assigned and magnitude of spatio-temporal distance, the following function was used: $W = \frac{K}{(1+d)}$ where W is the weight, K is a constant, d is the magnitude of spatio-temporal distance.

5.3 Flat Clustering

Since we focus on photo streams, for the prediction phase of our work, we treat photos as data streaming in and use a basic flavor of sliding window based clustering algorithm, inspired by [6]; We cluster based on time, latitude and longitude (three dimensional clustering). We start with a random seed of sliding window and then adapt it as we see more data. The photo stream is processed in sliding windows of fixed length and outdated data is discarded. The clustering process consists of an offline component that computes clustering statistics, and an online component that uses these statistics for the clustering for each sliding time window.



Time Range: Sun Feb 21 2010, 18:00 - 20:11
Location: 3800 Interlaken, Switzerland

Time range: Sun Sep 13 2009, 22:50 - 23:56
Location: Huangpu Qu, Shanghai Shi, China

Time range: Mon May 10 2010, 23:19 - 00:09
Location: State St, Carlsbad, CA 92008, USA

Table 5.2: Clustering of an example Photo Stream into three clusters

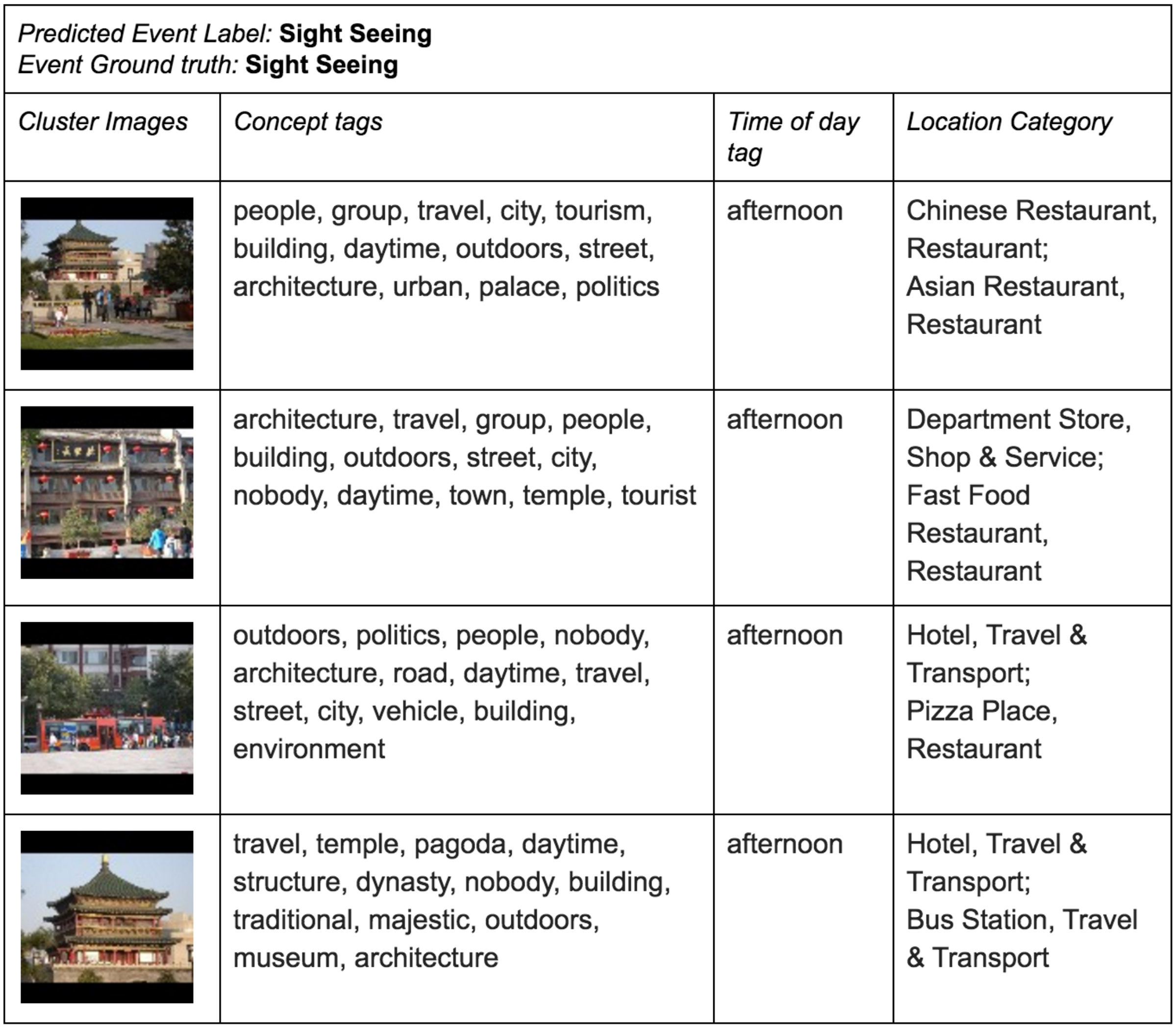
Chapter 6

Probabilistic Learning

Our data is of probabilistic nature, with some of the concepts, location categories possibly noisy, uncertain or irrelevant to the images that they represent. Furthermore, the relationship between event labels and these cluster attributes, relative to our ability to model them, are also not deterministic. Therefore, we use probabilistic models to represent the correlations between the event labels and these features. In this thesis, we explore a generative probabilistic method, Naive Bayes and a discriminative probabilistic graphical model, CRF.

Naive Bayes is a generative probabilistic model based on the Bayes theorem. A generative classifier learns a model of the joint probabililty $P(x, y)$ of the inputs x and output labels y, and then makes predictions using the Bayes rule to compute the posterior $P(x|y)$, by picking the most probable y label. A discriminative model learns the conditional probability distribution or posterior $P(x|y)$ directly.

CRF is a discriminative probabilistic graphical model which relaxes strong independence conditions. Probabilistic graphical models use a graph based representation to compactly model complex dependence relationships between random variables over a high dimensional space [22]. They encode relationships between multiple variables using a joint or conditional



Predicted Event Label: **Sight Seeing**
Event Ground truth: **Sight Seeing**

Cluster Images	*Concept tags*	*Time of day tag*	*Location Category*
	people, group, travel, city, tourism, building, daytime, outdoors, street, architecture, urban, palace, politics	afternoon	Chinese Restaurant, Restaurant; Asian Restaurant, Restaurant
	architecture, travel, group, people, building, outdoors, street, city, nobody, daytime, town, temple, tourist	afternoon	Department Store, Shop & Service; Fast Food Restaurant, Restaurant
	outdoors, politics, people, nobody, architecture, road, daytime, travel, street, city, vehicle, building, environment	afternoon	Hotel, Travel & Transport; Pizza Place, Restaurant
	travel, temple, pagoda, daytime, structure, dynasty, nobody, building, traditional, majestic, outdoors, museum, architecture	afternoon	Hotel, Travel & Transport; Bus Station, Travel & Transport

Table 6.1: Event Prediction for an example cluster

probability distribution, such that given observations, we are able to make predictions of the solutions with a probability distribution over all feasible solutions [37]. These probability distributions are specified by means of a graph. There are two types of graphical models - Bayesian networks and Markov networks. Bayesian networks are directed graphical models, where the family of probability distributions are specified by directed acyclic graphs, representing a factorization of conditional probability distributions over the random variables. Markov Random Fields are undirected graphical models, which defines a family of joint probability distributions using an undirected graph, which maybe cyclic [37]. Factor graphs are undirected graphical models that explicitly factorize the probability distributions. Once a model has been specified and parameterized, its parameters are learned using training instances, and the resulting model is used to solve inference tasks on future instances of data.

To illustrate the working of our general prediction model, Table 6.1 contains a example image cluster with each of its features and whose ground truth event label is *Sight Seeing.* This cluster consists of four images, with their respective concepts, location category tags and time of day tags. We notice that the cluster consists of frequently occurring tags for this event label, as evident in Figure 4.2, and the probabilistic model which was trained to capture these correlations correctly predicts the event label as *Sight Seeing.* We implement prediction using Naive Bayes and CRF.

6.1 Naive Bayes

Naive Bayes is a generative probabilistic model based on the Bayes theorem and assumes that the features are independent given the label. We approach the classification task here as a variety of event classification in a Bayesian learning model. We assume that the data was generated by a parametric model and use training data to compute the optimal values of the parameters of the Bayes model. Using this learned model, incoming feature data is classified using Bayes rule to calculate the posterior probability for each event class, hence selecting the most probable class as the output label.

In an image cluster, the visual concepts for the cluster are represented by $c = \{c_i\}$ where $i = 1, 2, ..N$ and N is the total number of images, the time of day tags are represented by $t = \{t_i\}$, and the location categories are represented by $l = \{l_i\}$. The set of possible event labels is represented by $E = \{e_1..e_m\}$, where m is the total number of event labels.

Given an test event cluster instance j with tags for time of day t^j, location category l^j and visual concepts c^j, we seek the event label e_k that maximizes $P(e_k|c^j, t^j, l^j)$.

Applying Bayes rule gives us,

$$P(e_k|c^j, t^j, l^j) = \frac{P(c^j, t^j, l^j|e_k)P(e_k)}{P(c^j, t^j, l^j)} \tag{6.1}$$

Since Naive Bayes assumes conditional independence between features given class, we get

$$P(c^j, t^j, l^j|e_k) = \prod^{i} P(c^j_i, t^j_i, l^j_i|e_k) \tag{6.2}$$

where j is the cluster instance, and i represents the ith image of the cluster. Therefore, given an image cluster j with attributes c^j, t^j, l^j we maximize for k in

$$P(e_k|c^j, t^j, l^j) = \frac{\prod^{i} P(c^j_i, t^j_i, l^j_i|e_k)P(e_k)}{P(c^j, t^j, l^j)} \tag{6.3}$$

For the inference, a *maximum a posteriori* or MAP decision rule is used, where we pick the class that is most probable. The corresponding classifier is called a Bayes classifier, which assigns an event label e_k for some k as follows:

$$\hat{y} = \arg\max_{k} \prod^{i} P(c^j_i, t^j_i, l^j_i|e_k)P(e_k) \tag{6.4}$$

Since $P(c^j, t^j, l^j)$ is a constant, it is ignored in the inference stage. The prior probabilities $P(c^j_i, t^j_i, l^j_i|e_k)$ and $P(e_k)$ are learned from the training data. Our implementation is based on MonkeyLearn [5]. A 75:25 split of the dataset was used as the training and testing datasets respectively.

6.2 Conditional Random Fields

We use Conditional Random Fields (CRF) [24], an undirected probabilistic graphical model, to model the correlation between the event labels and the derived attributes from the dataset. CRF allows us to relax strong independence assumptions in the state transition and directly models the conditional probability of labels given features. A 1st-order Markov CRF with state and transition features was modeled using CRFSuite [38], with state features conditioned on combinations of attributes and event labels, and transition features conditioned on event labels.

In an image cluster, the visual concepts for the ith image of a cluster are represented by $c = \{c_i\}$ where $i = 1, 2, ..N$, the time of day tags are represented by $t = \{t_i\}$, and the location categories are represented by $l = \{l_i\}$. We denote the event label using s_i^k, where $s_i^k = 1$ if the ith image is represented by the kth event label and $s_i^k = 0$, if not.

Given the tags for time of day, location category and visual concepts, we model the correlation between the event label and these features using the conditional probability of the kth event label as

$$P(s^k|c,t,l) = \frac{1}{Z_s}\exp(\sum_{i=1}^{N}\beta^k.f^k(c_i,t_i,l_i,s_i^k) + \sum_{i=1}^{N-1}\lambda^k.r^k(c_i,c_{i+1},t_i,t_{i+1},l_i,l_{i+1})) \tag{6.5}$$

where Z_s is the normalization constant, f_s^k is the feature function for individual images in a cluster with event label k and r_s^k models the correlation between features of consecutive images in the cluster.

The log-likelihood function for the kth event label is given by

$$L^k = -\log(Z_s) + \sum_{i=1}^{N}\beta^k.f^k(c_i,t_i,l_i,s_i^k) + \sum_{i=1}^{N-1}\lambda^k.r^k(c_i,c_{i+1},t_i,t_{i+1},l_i,l_{i+1}) \tag{6.6}$$

During the training phase, the parameter vectors λ^k, β^k are learned so as to maximize the objective function L^k in (6.6). While testing, given c, l, t, the event label s^k which maximizes (6.6) is returned as the predicted event label. The model is trained using Gradient Descent and the maximization of the logarithm of likelihood of the training data with L1 regularization term was computed using the L-BFGS [36] method. A 75:25 split of the dataset was used as the training and testing datasets respectively.

Chapter 7

Experiments and Evaluation

We discuss the experiments conducted and their evaluation results for the clustering and prediction components of our system model. We then discuss the overall implications of our evaluation results in the discussion section.

7.1 Clustering

Metrics for Evaluation

The clustering algorithms were evaluated with the following metrics:

Purity: Each cluster is assigned to the class which is most frequent in the cluster, and the purity is evaluated by finding the fraction of correctly assigned points. This metric is skewed when number of clusters is large, or each point is assigned to its own cluster.

Adjusted Random Index: Measures the similarity of each combination of two clusters in the dataset. This metric ignores permutations and has chance normalization. It also penalizes

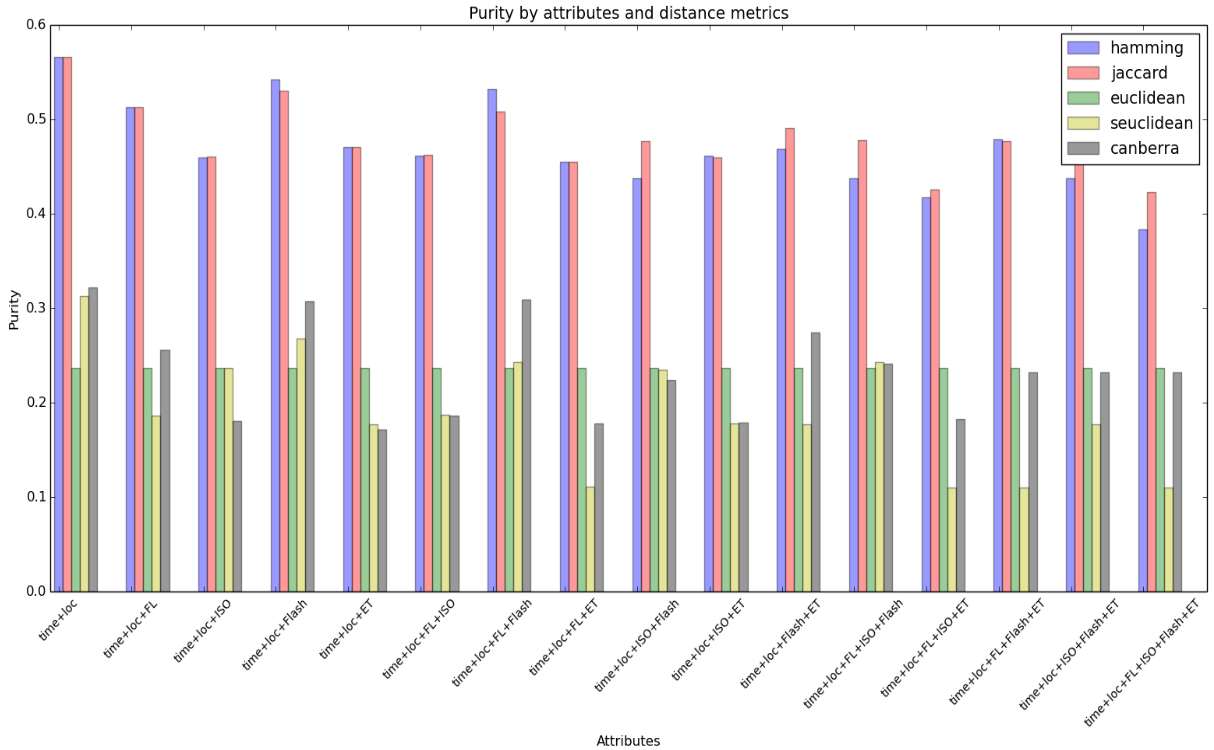


Figure 7.1: Comparison of Hierarchical Clustering for various distance metrics and attributes

false positives and false negatives.

Adjusted Mutual Information: Measures the agreement of the two assignments. This metric ignores permutations and is normalized against number of clusters, hence can be used to evaluate clusterings with different number of clusters.

Evaluation

An evaluation was done of the efficacy of hierarchical clustering for various combinations of the Image EXIF attributes like time, location, focal length, flash, ISO and exposure time. For each of these attributes, hierarchical clustering was implemented using various distance metrics like Hamming, Jaccard, Euclidean, Squared Euclidean, Canberra. The results are displayed in Figure 7.1. We find that the best quality of clustering in terms of the purity metric, was obtained using only the time and location attributes, for distance metrics of

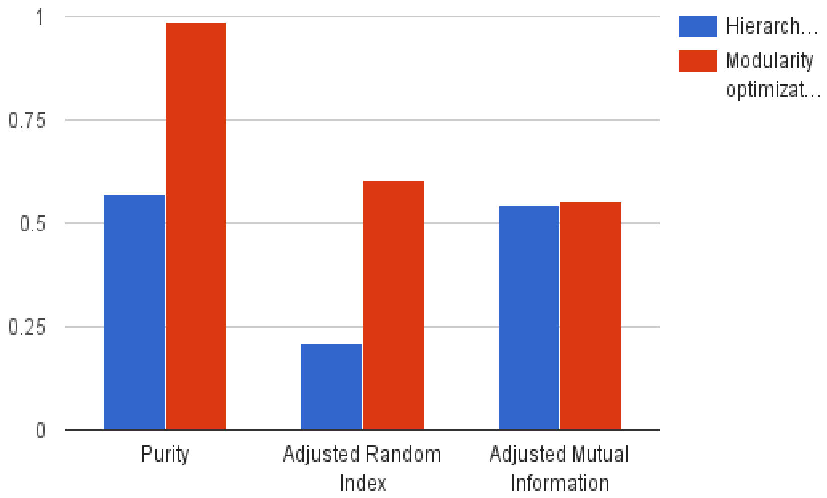


Figure 7.2: Comparison of Purity, ARI, AMI for Hierarchical Clustering and Modularity Optimization

hamming and jaccard. The other attributes that were found to provide decent clustering results when used along with time and location were ISO and Flash. Based on these findings, for the prediction phase, we clustered the data based on only time and location.

The clustering algorithms that were evaluated were Hierarchical clustering and Modularity optimization. These algorithms were evaluated based on metrics like Purity, Adjusted Ramdon Index and Adjusted Mutual Information. The evaluation results and comparison is illustrated by Figure 7.2. Modularity Optimization seems to perform better in terms of ARI and Purity, but these metrics are skewed and don't account for the large number of single node clusters formed. The AMI metric, which does normalize against these flaws, provides a more just evaluation between the two algorithms. According to the AMI metric, both algorithms perform almost equally well. Hierarchical clustering has the added benefit of allowing us to control the number of clusters formed.

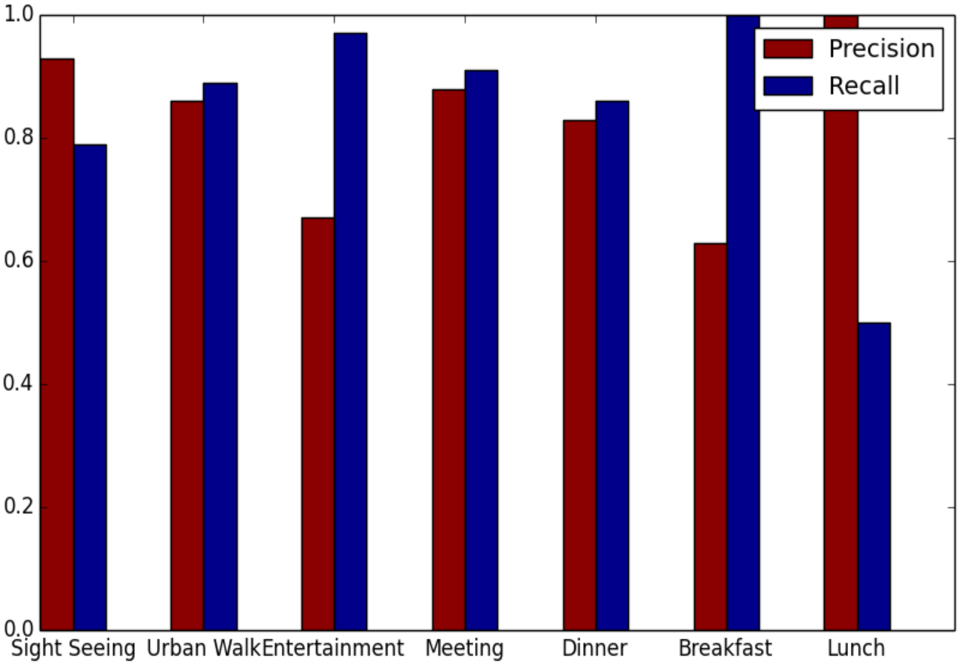


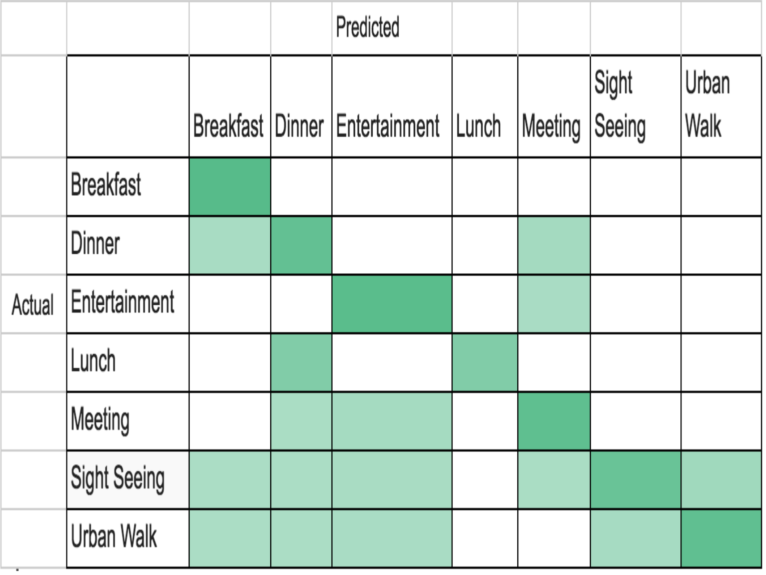
Figure 7.3: Naive Bayes Precision-Recall for individual event labels

7.2 Probabilistic Learning

Naive Bayes

The performance of our proposed prediction approach is evaluated using metrics like precision and recall. Figure 7.3 displayed the precision and recall performance of Naive Bayes for each of the event labels. The Naive Bayes classifier provides consistently good performance for all labels, and manages to classify event labels of *Breakfast* and *Lunch*, despite the low number of data points pertaining to these labels. Excluding the outlier event labels of *Breakfast* and *Lunch*, the best performing event labels are *Sight Seeing* with a precision of 0.93 and *Entertainment* with a recall of 0.97.

Figure 7.4 is a confusion matrix for the Naive Bayes classifier, which shows the classification performance of each event label with respect to the others - it displays the false positives and



				Predicted				
		Breakfast	Dinner	Entertainment	Lunch	Meeting	Sight Seeing	Urban Walk
	Breakfast	🟩						
	Dinner	🟩	🟩			🟩		
Actual	Entertainment			🟩		🟩		
	Lunch		🟩		🟩			
	Meeting		🟩	🟩		🟩		
	Sight Seeing	🟩	🟩	🟩		🟩	🟩	🟩
	Urban Walk	🟩	🟩	🟩			🟩	🟩

Figure 7.4: Confusion matrix for Naive Bayes Event Prediction

true negatives as classified by the Naive Bayes. It is evident from Figure 7.4 that the Naive Bayes classifier correctly predicts the event labels on most instances. The misclassification that occurs most commonly is that of predicting *Urban Walk* instead of the correct label of *Sight Seeing*. Some other common misclassifications are predicting *Entertainment* and *Sight Seeing* instead of the correct label of *Urban Walk*, and *Meeting* instead of *Dinner*.

Conditional Random Fields

Figure 7.5 depicts the CRF precision and recall values for each of the event labels. The event labels of *Meeting* and *Dinner* were best performing, with a best precision of 0.7333 for *Meeting* and best recall of 0.7143 for *Dinner*. Due to the low number of training and testing data points for the event labels of *Breakfast* and *Lunch*, the precision and recall of these two event labels dropped to zero. A comparison of the average precision and recall including and excluding these two outliers, as shown in Figure 7.6 shows that the average precision increased considerably from 0.439 to 0.615, and the average recall increased from 0.417 to 0.584, after excluding these two worst performing event labels.

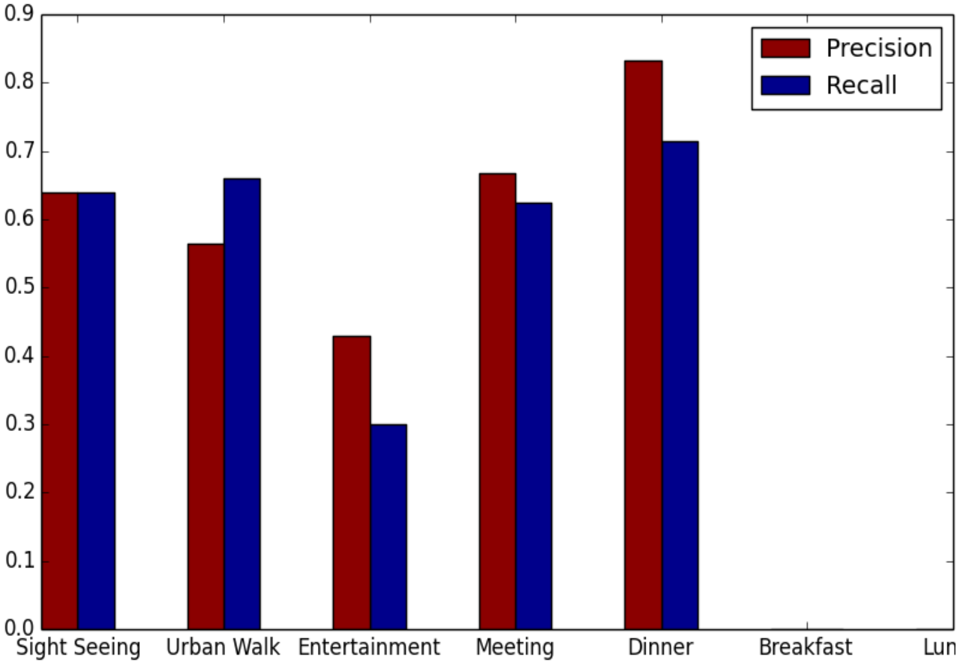


Figure 7.5: CRF Precision-Recall for individual event labels

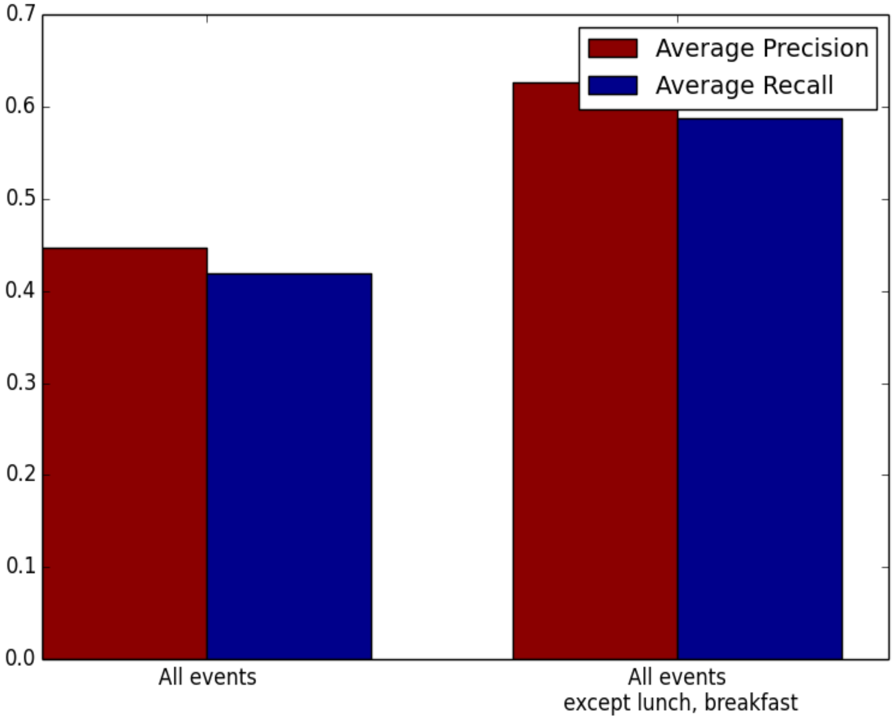


Figure 7.6: CRF Average Precision-Recall comparison

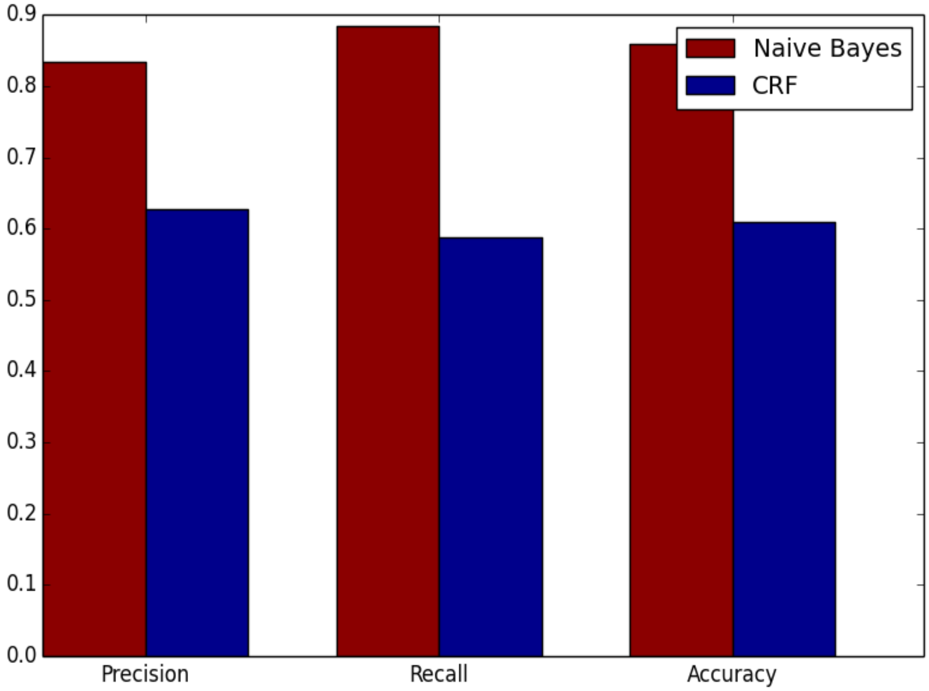


Figure 7.7: General Comparison of Naive Bayes and CRF

Comparison and Evaluation

Figure 7.7 compares the Naive Bayes classifier with the CRF classifier, in terms of metrics of precision, recall and accuracy. The precision and recall are averaged over all the event labels. Naive Bayes is found to perform better in terms of all three metrics, achieving an average precision of 0.834 and an average recall of 0.884 over CRF's 0.626 and 0.588 respectively. These averages were computed for the best performing five labels, to the exclusion of *Breakfast* and *Lunch*, for a fair comparison for events with decent number of training data. In terms of accuracy, Naive Bayes achieves an accuracy of 87% over CRF's 61%. This shows that Naive Bayes outperforms CRF comprehensively and provides accurate prediction results.

Figure 7.8 displays the comparison between CRF and Naive Bayes for each event label using a metric called F1-measure, which is the harmonic mean of precision and recall. Naive Bayes

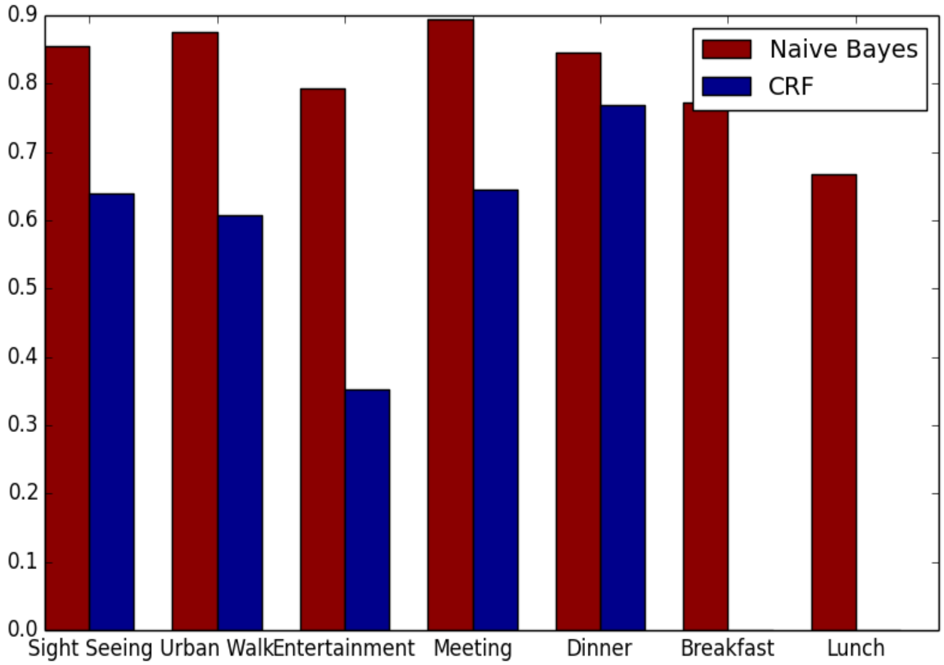


Figure 7.8: F1-Measure Comparison of Naive Bayes and CRF for individual event labels

outperforms CRF for all event labels. Naive Bayes is able to predict even outlier labels of *Lunch* and *Breakfast*, unlike CRF, despite the low number of training data available for these two events. This shows the robustness of the simpler Naive Bayes prediction model over a more complex model like CRF which accounts for dependencies in the data, even with such dependencies existing between attributes in our data.

7.3 Discussions

Our approach was evaluated for a personal image dataset that was manually labeled with ground truth for clusters as well as with ground truth of event labels. We discuss our findings from the evaluations for each component of our model - the clustering component and the prediction component, and then discuss the overall performance of the approach.

The results obtained from clustering was evaluated against the ground truth with metrics such as purity, AMI and ARI, and the results obtained were promising. The results of clustering using various combination of attributes was compared, and spatio-temporal attributes were found to provide the best clustering. Finding metrics which justly evaluated clusterings against each other and against the ground truth was challenging, and hence multiple such metrics for evaluation were considered. We obtained clusters of AMI of over 0.5 for both algorithms evaluated, which are encouraging results, but also leaves room for further experimentation and improvements.

The results obtained for prediction were considerably good, with Naive Bayes giving us an accuracy of 87% over the entire test dataset. These prediction results prove that the features that were used to augment the dataset are of a relevant nature, and these features when used with probabilistic models provide for accurate results in event prediction.

Our overall approach performs well together, firstly giving us image groupings according to events (represented as a spatio-temporal point/region) and secondly providing us high-level event annotations for these image groups with high accuracy.

Chapter 8

Conclusion

We present an event-based image organization approach, where we combine clustering with probabilistic learning methods, to detect and infer events in a photo stream; we use multiple web-based sources for semantically augmenting the metadata of the images. Most notably, we leverage recent advances in deep learning to obtain highly detailed visual concepts from images. The efficiency of our event-based image organization approach was evaluated using a personal image dataset, and the results have been promising. Grouping of images based on events would allow for organization of images into semantically meaningful collections, and automatically inferring these events by exploiting visual concepts and image metadata is the first step in this direction.

Current state of the art in image organization and features group photos based on a single attribute - either time or location or people or objects. Events encapsulate all of these attributes and provide a natural abstraction of specific life experiences for users. Hence, events are an effective and intuitive mechanism for grouping of images. Automatic annotation of events in a photo stream is essential for effective search and organization of images, saving users valuable time and effort of manual annotation. Our approach when applied to photo streams provides an efficient framework for automatic event detection and annotation, providing for semantically meaningful organization of images.

8.1 Future Work

Implementing this approach on publicly available photo streams, such as those from Flickr, would be a natural next step, so as to allow us to evaluate this approach on real-world streaming data. These public data collections would provide for more training data, thus improving the precision and recall. Publicly available datasets contain user provided descriptions of images and collections, providing another avenue of data from which event-related information can be gleaned. This would allow us to build on the set of derived features we already have, thus creating a powerful framework for inferring events in photo streams.

Another future direction is utlizing agglomerative clustering along with event ontologies for predicting more fine-grained event annotations using similar probabilistic models. Such an approach not only provides annotation for events, but also allows for the semi-automatic creation of event ontology models which can be generated from the extracted correlations between features and event labels. Extending this framework to include these ontologies would make for a powerful automatic event annotation framework.

Bibliography

[1] Big Data Usage Per Minute. http://www.inc.com/larry-kim/15-mind-blowing-statistics-reveal-what-happens-on-the-internet-in-a-minute.html.

[2] ClarifAI. https://developer.clarifai.com/.

[3] Facebook Data Statistics. https://zephoria.com/top-15-valuable-facebook-statistics/.

[4] FourSquare. https://developer.foursquare.com/.

[5] MonkeyLearn. http://www.monkeylearn.com/.

[6] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proceedings of the 29th international conference on Very large data bases-Volume 29*, pages 81–92. VLDB Endowment, 2003.

[7] H. Becker, M. Naaman, and L. Gravano. Event identification in social media. In *WebDB*, 2009.

[8] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[9] L. Cao, J. Luo, H. Kautz, and T. S. Huang. Image annotation within the context of personal photo collections using hierarchical event and scene models. *Multimedia, IEEE Transactions on*, 11(2):208–219, 2009.

[10] M. Chen, T. Nguyen, and B. K. Szymanski. On measuring the quality of a network community structure. In *Social Computing (SocialCom), 2013 International Conference on*, pages 122–127. IEEE, 2013.

[11] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox. Temporal event clustering for digital photo collections. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 1(3):269–288, 2005.

[12] M. Das and A. C. Loui. Event classification in personal image collections. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1660–1663. IEEE, 2009.

[13] P. Domingos and M. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130, 1997.

[14] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[15] B. Gong, U. Westermann, S. Agaram, and R. Jain. Event discovery in multimedia reconnaissance data using spatio-temporal clustering. In *Proc. of the AAAI Workshop on Event Extraction and Synthesis (EES'06)*, 2006.

[16] X. He, R. S. Zemel, and M. Á. Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on*, volume 2, pages II–695. IEEE, 2004.

[17] J. L. Hellerstein, T. Jayram, I. Rish, et al. *Recognizing end-user transactions in performance management.* IBM Thomas J. Watson Research Division, 2000.

[18] R. Jain and P. Sinha. Content without context is meaningless. In *Proceedings of the international conference on Multimedia*, pages 1259–1268. ACM, 2010.

[19] W. Jiang and A. C. Loui. Semantic event detection for consumer photo and video collections. In *Multimedia and Expo, 2008 IEEE International Conference on*, pages 313–316. IEEE, 2008.

[20] A. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002.

[21] H. Khrouf and R. Troncy. Eventmedia: A lod dataset of events illustrated with media. *Semantic Web*, (Preprint):1–7, 2012.

[22] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

[23] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1150–1157. IEEE, 2003.

[24] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[25] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[26] L. Liao, D. Fox, and H. Kautz. Hierarchical conditional random fields for gps-based activity recognition. In *Robotics Research*, pages 487–506. Springer, 2007.

[27] X. Liu, R. Troncy, and B. Huet. Finding media illustrating events. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 58. ACM, 2011.

[28] X. Liu, R. Troncy, and B. Huet. Using social media to identify events. In *Proceedings of the 3rd ACM SIGMM international workshop on Social media*, pages 3–8. ACM, 2011.

[29] A. C. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *Multimedia, IEEE Transactions on*, 5(3):390–402, 2003.

[30] J. Luo, M. Boutell, and C. Brown. Pictures are not taken in a vacuum-an overview of exploiting context for semantic scene content understanding. *Signal Processing Magazine, IEEE*, 23(2):101–114, 2006.

[31] R. Mattivi, J. Uijlings, F. G. De Natale, and N. Sebe. Exploitation of time constraints for (sub-) event recognition. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 7–12. ACM, 2011.

[32] A. McCallum. Efficiently inducing features of conditional random fields. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 403–410. Morgan Kaufmann Publishers Inc., 2002.

[33] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION*, pages 41–48. AAAI Press, 1998.

[34] M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 196–203. ACM, 2004.

[35] M. E. Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[36] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of computation*, 35(151):773–782, 1980.

[37] S. Nowozin and C. H. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends® in Computer Graphics and Vision*, 6(3–4):185–365, 2011.

[38] N. Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.

[39] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Advances in neural information processing systems*, pages 1097–1104, 2004.

[40] S. Rafatirad, R. Jain, and K. Laskey. Context-based event ontology extension in multimedia applications. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pages 278–285. IEEE, 2013.

[41] T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110. ACM, 2007.

[42] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[43] I. Rish. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[44] K. Rodden and K. R. Wood. How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 409–416. ACM, 2003.

[45] D. A. Ross, S. Osindero, and R. S. Zemel. Combining discriminative features to infer complex trajectories. In *Proceedings of the 23rd international conference on Machine learning*, pages 761–768. ACM, 2006.

[46] M. K. Saini, F. Al-Zamzami, and A. E. Saddik. Towards storytelling by extracting social information from osn photo's metadata. In *Proceedings of the First International Workshop on Internet-Scale Multimedia Management*, pages 15–20. ACM, 2014.

[47] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 134–141. Association for Computational Linguistics, 2003.

[48] I. Tankoyeu, J. Paniagua, J. Stöttinger, and F. Giunchiglia. Event detection and scene attraction by very simple contextual cues. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 1–6. ACM, 2011.

[49] T. Wang, J. Li, Q. Diao, W. Hu, Y. Zhang, and C. Dulong. Semantic event detection using conditional random fields. In *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pages 109–109. IEEE, 2006.

[50] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE MultiMedia*, (1):19–29, 2007.

[51] J. M. Zacks, T. S. Braver, M. A. Sheridan, D. I. Donaldson, A. Z. Snyder, J. M. Ollinger, R. L. Buckner, and M. E. Raichle. Human brain activity time-locked to perceptual event boundaries. *Nature neuroscience*, 4(6):651–655, 2001.

[52] H. Zhang. The optimality of naive bayes. *AA*, 1(2):3, 2004.