Chapter 2

MULTIMEDIA INFORMATION NETWORKS IN SOCIAL MEDIA

Liangliang Cao, GuoJun Qi, Shen-Fu Tsai, Min-Hsuan Tsai, Andrey Del Pozo, Thomas S. Huang
Beckman Institute and Coordinated Science Lab
Department of ECE, University of Illinois at Urbana-Champaign
{cao4,qi4,stsai8,mtsa2,delpozo2,huang}@uiuc.edu

Xuemei Zhang and Suk Hwan Lim
Multimedia Interaction and Understanding Lab
HP Labs, Palo Alto, CA
{suk-hwan.lim,xuemei.zhang}@hp.com

Abstract
The popularity of personal digital cameras and online photo/video sharing community has lead to an explosion of multimedia information. Unlike traditional multimedia data, many new multimedia datasets are organized in a structural way, incorporating rich information such as semantic ontology, social interaction, community media, geographical maps, in addition to the multimedia contents by themselves. Studies of such structured multimedia data have resulted in a new research area, which is referred to as Multimedia Information Networks (MINet). MINets are closely related to social networks, but especially focus on understanding the topics and semantics of the multimedia files in the context of network structure. This chapter reviews different categories of recent MINet systems, summarizes the popular inference methods used in recent works, and discusses the applications related to MINets. We also discuss a wide range of topics including public datasets, related industrial systems, and potential future research directions in this field.

Keywords: Multimedia Information Networks, Social Media, Personal Photo Albums, Ontology, Geographical Annotation
1. Overview

A Multimedia Information Network (MINet) is a structured multimedia collection, where Multimedia documents, such as images and videos, play as nodes connected by various links. The upsurge of MINet is due to the flourish of web images and the popularity of online community. For example, Flickr and Facebook host billions of images, which are linked to each other via the users, groups, and tags. The network structure contains rich semantic information, such as semantic ontology, social relationships, and meta-information. It has become an attractive challenge to explore the network structure for the rich multimedia content in Multimedia Information Networks. An appropriate recent quote by John Munsell is as follows:

“If the content is king, then the conversion is queen”

Multimedia information networks can be viewed as a marriage of multimedia content and social networks. However, it conveys even richer information than each single topic. To understand MINets, we must consider not only the visual features for each node, but also explore the network structure associated with them. In recent years, there has been a lot of work on MINet, and this has begun to shape a new research topic in this area.

Although it has become a hot research topic in recent years, the subject of MINets is still in its early stage. This paper tries to summarize the recent development of MINets, with the aim of not covering as many contents as possible, but motivating more interesting problems in this field. This paper constitutes two parts: in the first part, we summarize four categories of popular link structures in MINets: semantic ontology, community media, personal photo albums, and geographical location. For each link structure, we review some typical methods together with their applications. In the second part, we discuss some specific topics in this field: data sets and industrial systems, inference methods, and future directions.

2. Links from Semantics: Ontology-based Learning

Ontology plays a pivotal role for representing concepts that we are concerned with and their relations. Ontology would contain whatever human knowledge pertaining to the domain of interests that could help better process and analysis of visual data and textual cues.

In [18], an ontology was constructed by hand to facilitate personal photo album management. In [106], the authors worked on image retrieval for animal domain. They extracted Scientific Classification information available in Wikipedia pages as their animal domain ontology. Next, in order to build textual ontology, they parsed relevant Wikipedia sections and found important keywords including concepts and relations. Finally, the relations in the ontology are further cross-verified via “is-a” relations in animal ontology.
In the case of ontology for more general subjects, Wikipedia and WordNet are two important knowledge bases. Wikipedia stores knowledge of generic terms and name entities in its countless articles. This results in difficulty in automatic and systematic knowledge extraction, while WordNet provides a lot of relational information between generic terms. The YAGO system developed in [101] bridges these two and provides a comprehensive ontology covering many generic subjects in daily life.

There have been some works focusing on hierarchical classification and retrieval. In [64], the ontology is more general as it exploits “is-a” and “part-of” relations and is in general a graph. In the resulting graph, each edge is associated with a binary classifier which computes conditional likelihood given the previous concept. Because the graph may not be a tree, there may be multiple paths from the root concept to any target concept. Each path is pessimistically associated with the minimum of the conditional likelihoods associated with all edges on the path. The marginal likelihood of any target concept is then optimistically set as the maximum of all path likelihoods from root to the target concept. Similarly for ImageNet constructed in [25], the “tree-max classifier” that propagates maximum of all descendant concept likelihoods one level up was proposed to estimate individual concept likelihood in the tree-structured ontology. In [126], the authors investigated the effects of combining label estimates of classifiers trained at various levels, as well as possibilities of transferring concept models to neighboring concepts like sibling concepts to address small sample problem. A more sophisticated hierarchical classification proposed in [11] is to encode the path from root concept to target concept by a fix-length binary vector and then treat it as a multilabel classification problem.

Effectively modeling structured concepts has become a critical ingredient for retrieving and searching multimedia data on the Web. Many sophisticated models have been proposed to recognize a wide range of multimedia concepts from our everyday lives to many specific domains, which lead to a collection of multimedia model warehouses such as Large-Scale Concept for Multimedia (LSCOM) [76] and 101 semantic concepts in multimedia [99]. To model the ontology, some researchers employ a flat correlative concept structure [81] [77], while others try to build hierarchical structures based on semantic correlations [114] [27] [64].

3. Links from Community Media

The center concept governing community media is that users play the central role in retrieving, indexing and mining media content. The basic idea is quite different from the traditional content-centric multimedia system. The web sites providing community media are not solely operated by the the owners but by millions of amateur users who provide, share, edit and index these
media content. In this section, we will review recent research advancements in community media systems from two aspects: (1) retrieval and indexing system for community media based on user-contributed tags, and (2) community media recommendations by mining user ratings on media content.

### 3.1 Retrieval Systems for Community Media

Recent advances in internet speed and easy-to-use user interfaces provided by some web companies, such as Flickr, Corbis and Facebook, have significantly promoted image sharing, exchange and propagation among users. Meanwhile, the infrastructures of image-sharing social networks make it an easy task for users to attach tags to images. These huge amount of user tags enable the fine understanding of the associated images and provide many research opportunities to boost image search and retrieval performance. On the other hand, the user tags somehow reflect the users’ intentions and subjectivities and therefore can be leveraged to build a user-driven image search system.

To develop a reliable retrieval system for community media based on these user contributed tags, two basic problems must be resolved. First of all, the
user tags are often quite noisy or even semantically meaningless [21]. More specifically, the user tags are known to be ambiguous, limited in terms of completeness, and overly personalized [32] [65]. This is not surprising because of the uncontrolled nature of social tagging and the diversity of knowledge and cultural background of the users [55]. To guarantee a satisfactory retrieval performance, tag denoising methods are required to refine these tags before they can be used for retrieval and indexing. Some examples of tag denoising methods are described below.

The work in [103] proposes to construct an intermediate concept space from user tags which can be used as a medium to infer and detect more generic concepts of interest in the future. The work in [112] proposes a probabilistic framework to resolve ambiguous tags which are likely to occur but appear in different contexts with the help of human effort. There also exist many tag suggestion methods [4] [69] [98] [115] which help users annotate community media with most informative tags, and avoid meaningless or low-quality tags. For all these methods, tag suggestion systems are involved to actively guide users to provide high-quality tags based on tags co-occurrence relations.

Secondly, the tags associated with an image are generally in a random order without any importance or relevance information, which limits the effectiveness of these tags in search and other applications. To overcome this problem, [59] proposes a tag ranking scheme that aims to automatically rank the tags associated with a given image according to their relevance to the image content. This tag ranking system estimates the initial relevance scores for the tags based on probability density estimations, and followed by a random walk over a tag similarity graph in order to refine the relevance scores. Another method was proposed in [55] that learns tag relevance by accumulating votes from visually similar neighbors. Treated as tag frequency, this learned tag relevance is seamlessly embedded into tag-based social image retrieval paradigms.

Many efforts have been made on developing the multimedia retrieval systems by mining the user tags. As a typical social image retrieval system illustrated in Figure 2.1, a distance metric is mined from these web images and their associated user tags, which can be directly applied to retrieve web images in a content-based image retrieval paradigm [83]. In [109], Wang et al. propose a novel attempt at model-free image annotation, which is a data-driven approach that annotates images by mining their search results based on user tags and surrounding text. Since no training data set is required, their approach enables annotating with unlimited vocabulary and is highly scalable and robust to outliers.
3.2 Recommendation Systems for Community Media

Developing recommendation systems for community media has attracted much attention, because of the popularity of Web 2.0 applications, such as Flickr, Youtube and Facebook. Users give their own comments and ratings on multimedia items, such as images, amateur videos and movies. However, only a small portion of the multimedia items have been rated and thus the available user ratings are quite sparse. Therefore, an automatic recommendation system is desired to be able to predict users’ ratings on multimedia items, so that they can easily find the interesting images, videos and movies from shared multimedia contents.

Recommendation systems measure the user interest in given items or products to provide personalized recommendations based on user taste [39] [6]. It becomes more and more important to enhance user experience and loyalty by providing them with the most appropriate products in e-commerce web sites (such as Amazon, eBay, Netflix, TiVo and Yahoo). Thus, there are many advantages in designing a user-satisfied recommendation system.

Currently, existing recommendation systems can be categorized into two different types. The content-based approach [1] creates a profile for each user or product which depicts its nature. User profiles can be described by their historical rating records on movies, personal information (such as their age, gender, or occupation) and their movie types of interest. Meanwhile, movie profiles can be represented by other features, such as their titles, release date, and movie genres (e.g., action, adventure, animation, comedy). The obtained profiles allow programs to quantify the associations between users and products.

The other popular recommendation systems rely only on the past user ratings on products with no need to create explicit profiles. This method is known as collaborative filtering (CF) [31], and it analyzes relationships between users and interdependencies among products. In other words, it aims at predicting a user rating based on user ratings on the same set of multimedia items. The only information used in CF is the historical behavior of users, such as their previous transactions or the way they rate products. The CF method can also be cast as two primary approaches - the neighborhood approach and latent factor models. Neighborhood methods compute the relationships between users [38] [9] [85] or items [26] [56] [93] or combination thereof [108] to predict the preference of a user to a product. On the other hand, latent factor models transform both users and items into the same latent factor space and measure their interactions in this space directly. The most representative methods of latent factor models are singular value decomposition (SVD) [80]. Evaluations on recommendation systems suggest that SVD methods have gained state-of-the-art performance among many other methods [80].
Some public data sets are available for comparison purpose among different recommendation systems. Among them, the most exciting and popular one is the Netflix data set for movie recommendation\(^1\). The Netflix data set contains more than 100 million ratings on nearly 18 thousand movie titles from over 480,000 randomly-sampled customers. These user ratings were collected between October 1998 and December 2005, and they are able to represent the user trend and preference during this period. The ratings are given on a scale from 1 to 5 stars. The date of each rating as well as the title and year of release for each movie are provided. No other data, such as customer or movie information, were employed to compute Cinematch’s accuracy values used in this contest. In addition to the training data set, a qualifying test set is provided with over 2.8 million customer-movie pairs and the rating dates. These pairs were selected from the most recent ratings from a subset of the same customers in the training data set, over a subset of the same movies.\(^2\)

4. **Network of Personal Photo Albums**

The proliferation of high quality and modestly priced digital cameras has resulted in an explosion of personal digital photo collections in the past ten years. It is not uncommon for a home user to take thousands of digital photos each year. These images are largely unlabeled, and often have minimal organization done by the user. Managing, accessing, and making use of this collection has become a challenging task.

The majority of current work on image annotation and organization makes use of photo collections available on the web, such as those from Flickr, Facebook, or Wiki pages. This provide many insights that are also applicable to personal collections. However, personal collections also have some distinctive properties, that are very different from professional or web collections. We will discuss research works that address the issues related to each of these unique aspects of personal photo collections.

4.1 **Actor-Centric Nature of Personal Collections**

Both identity of people and number of people in photos are important cues to users in how they remember, describe, and search personal photos \([72]\). There is a large body of research on face recognition. The accuracy of state-of-the-art face recognition algorithms under controlled-capture situation are reasonably high \([68]\), although accuracy can vary significantly for different persons and data-sets. While accuracy under uncontrolled capture may not be sufficient for fully-automatic annotation, it can be a baseline for very efficient semi-automatic annotation. Personal collections are usually dominated by a small number of subjects. Therefore, with an intelligent user interfaces, it is possible to achieve 80% recall rate and near 100% precision with minimal amount of
user input, such as a few seconds of simple drag-and-drop merging of similar clusters [122]. Subjects in personal photos within a collection are usually related by family ties or friendships. Discovering and understanding the relationship among identified people in personal collections has significant application impact. Sharing of personal photos on social networking sites such as Facebook makes it possible to discover relations beyond an individual’s collection. Recent work by Wu et al [113] took advantage of face clustering technology in order to discover social relationships of subjects in personal photo collections. Co-occurrence of identified faces as well as inter-face distances (inferred from in-image distance and typical human face size) are used to calculate link strength between any two identified persons. In such cases, social clusters as well as social importance of individuals can be calculated.

Relationship analysis of subjects in personal photos can be used to improve facial annotation. Gallagher et al [29] exploits pairwise connections between frequent faces in personal photo collections in order to build up prior probabilities of groups of faces appearing in a single image. This information is used to identify individuals from ambiguously-labeled group shots. While the emphasis of this paper is not on relationship analysis, it shows one of the many potential values in modeling relationships between people in personal photo collections.

4.2 Quality Issues in Personal Collections

The transition to digital photography encouraged much more trial-and-error in consumer photo capture. People often take a number of photos of the same subjects in sequence in the hope of picking out an ideal one later. This habit contributed to the increasing collection size of personal photos. Effective near-duplicate detection is highly useful both in browsing and in creation of photo-based products such as photo-books.

Near-duplicate detection can be done with a variety of approaches, each with different trade-offs in accuracy and efficiency. Detection methods based on global features such as color histograms are fast but are likely to introduce false-positives when two photos with very different contents have similar colors, so they are often used in combination with other methods, or as a pre-filter step [43]. Detection based on structural features are less prone to confusion, at a much higher computational cost [20, 121]. A hybrid approach [102] combines time, color, and local structural information in a cascade framework so that more “expensive” calculations are only done when necessary.

After duplicate clusters are identified, selection, ranking, or mashing of the duplicate photos are usually called for in consumer applications. Selection in a duplicate set can be done by image quality or image appeal [94], or by how
well an image represents the cluster [20]. Since duplicates in a personal collection are usually taken with the same camera at the same scene from slightly different viewpoints, it is also possible to combine the images in the duplicate set to form a new composite image that is a “bigger and better” representation of the scene [97].

4.3 Time and Location Themes in Personal Collections

Personal collections are generally captured using a small number of cameras with time stamp (and sometimes GPS information). Typically, users store these photos in folders labeled with time and event information [49]. Event and location information are also among the top cues people remember and use when searching for photos [72]. Due to the availability and importance of time and location information for personal collections, both cues are used in photo clustering and annotation work for personal photos.

Time clustering is one of the fastest and most effective method in organizing personal photos into meaningful groups. It can be done with a simple K-means clustering algorithm [62] or with more sophisticated multi-scale approach [22], or by identifying “bursts” of photo taken with roughly the same rate [33]. Time clustering is often used in combination with image content similarity. In addition to being an informative cue of event grouping, time cues can also be used to limit image similarity calculations to images within a certain time interval, therefore reducing the computation load for content-based clustering [102].

Location information is another key element in event clustering of personal photos. Naaman et al. [74] used both time and location information to create a photo organization system. A location clustering and naming algorithm is used to assign location tags to each photo using GPS information extracted from photo EXIF headers. Cao et al [17] showed that even without GPS tags in EXIF headers, it is possible to generate location tags with usable accuracy when combining image features with existing textual tags. These results indicate the feasibility and usefulness of location identification in personal collection analysis.

4.4 Content Overlap in Personal Collections

For personal photos, users often know more about the content than what can be inferred from a single image, such as who took the picture, who was standing nearby, what happened before/after that moment, and what other objects go together with the objects visible in the image. Looking beyond a single photo can often make a big difference in the annotation accuracy of personal collections.

Annotation of a collection instead of individual photos is one way to take advantage of the photo-to-photo correlation in personal collections. Cao et al
used conditional random fields to model correlations of different photos in a sub-collection to annotate photos by scene type and also annotate sub-collections by event type. This work made good use of the inherent event-based organization and strong inter-photo correlations that are characteristic of personal collections to improve image collection annotation.

In the Gallagher et al work mentioned earlier [29], face appearance and co-appearance frequencies within personal collections are used to build up prior probabilities of groups of faces appearing in a single image, and this in turn is used to disambiguate personal labels for individual photos. This is an example of using the correlation of photos within a collection to improve individual photo annotation.

One can also go beyond the personal collection itself, and use photos in web collections and their associated labels to improve personal photo annotation. Liu et al. [61] learns images corresponding to concepts from enormous images and their surrounding text on the web. The links between concepts and visual features are transferred from the Internet to personal photo collections together with concepts hierarchies provided by WordNet(http://wordnet.princeton.edu/), the latter giving “subclass-of” relation between concept in order to decide the negative training images for a certain concept – if concept “pool” has descendants(subclasses) “natatorium”, “cistern”, “lido”, and “sink”, then images irrelevant to “pool” will be those surrounded by text that do not contain any of these subclasses. In a word, both concept-to-image link and subclass-of link between concepts are collected and utilized to train the image retrieval system.

Correlations among keywords, visually similar images, and closely time-stamped images usually provide much implicit human knowledge for annotation. Roughly speaking, visually similar photos are supposed to have correlated annotations, and semantically close keywords are usually used to describe largely overlapping sets of images. Motivated by these ideas, Jia et al. [42] proposed to propagate correlations of these three domains around to improve the final annotation, where initial keyword correlation is obtained from Google in a statistical way. After steady-state keyword and visual correlation graphs are obtained based on their initial values and annotation of the images, they are used together with the temporal correlation graph to get the final annotation.

5. Network of Geographical Information

Geographical information is often represented in the form of a longitude-latitude pair in order to represent the locations where the images are taken. In recent years, the use of geographical information has become more and more popular. With advances in low-cost GPS chips, cell phones and cameras have become equipped with GPS receivers, and thus are able to record the locations while taking pictures. Many online communities, such as Flickr and Google
Earth, allow users to specify the location of their shared images either manually through placement on a map or automatically using image meta-data embedded in the image files. At the end of 2009, there have been approximately 100 million geo-tagged images in Flickr, and millions of new images continue to be added each month.

![Figure 2.2. GPS Network](image)

Given the geographical information associated with images, we can easily construct a network by grouping images taken from neighboring regions. Figure 2.2 illustrates such a network. The network shows that the visual information may be correlated with GPS positions. This makes it possible to infer the image semantics with geographical information, or to estimate the geographical information from visual content. By exploring the rich media such as user tags, satellite images and wikipedia knowledge, we can leverage the visual and geographical information for many novel applications.

Geographical annotation provides a rich source of information which can link millions of images based on the similarity from geographical measures. There has been a growing body of work in visual research community investigating geographical information for image understanding [73] [2] [16] [118] [45] [119] [47] [71] [84] [63] [95] [12]. One line of research utilizes geographical information for better understanding of the image semantics. Another line of research is devoted to estimating the geographical information from general images.

### 5.1 Semantic Annotation

Naaman et al. [73] proposed a system to suggest candidate identity labels based on the meta-data of a photo including its time stamp and location. The
system explores the information of events and locations, and the co-occurrence statistics of people. However, image analysis techniques are not used in the system.

After Naaman’s work, more recent lines of research aim to understand semantics from both visual information and geographical collections. Joshi and Luo [45] propose to explore the Geographical Information Systems (GIS) database using a given geographical location. They use descriptions of small local neighborhoods to form bags of geo-tags as their representation. The associations of Geo-tags and visual features are learned to infer the event/activity labels such as “beach” or “wedding”. The authors demonstrate that the context of geographical location is a strong cue for visual event/activity recognition.

Yu and Luo [118] propose another way to leverage nonvisual contexts such as location and time stamps. Their approach learns from rough location (e.g., states in the US) and time (e.g., seasons) information, which can be obtained through picture metadata automatically. Both visual and nonvisual context information are fused using a probabilistic graphical model to improve the accuracy of object region recognition. In [63], the authors explore satellite images corresponding to picture location data and investigate their novel uses to recognize the picture-taking environment. The satellite image functions as a third eye above the object. This satellite information is combined with classical vision-based event detection methods. Luo et al. [63] employed both color- and structure-based visual vocabularies for characterizing ground and satellite images, respectively. The fusion of the complementary views (photo and satellite) achieves significant performance improvement over the ground view baseline.

5.2 Geographical Estimation

The previous section discusses research relevant to understanding the semantics better using geographical information. An interesting question of a different nature is whether we can use the visual information to estimate the geographical locations even when they are not provided. As evidenced by the success of Google Earth, there is a great need for such geographic information from users. Many web users have high interests in not only the places they live but also other interesting places around the world. Geographic annotation is also desirable when reviewing travel and vacation images. For example, when a user becomes interested in a nice photo, he or she may want to know where exactly it is. Moreover, if a user plans to visit a place, he or she may want to find out the points of interest nearby. Recent studies suggest that geo-tags expand the context that can be employed for image content analysis by adding extra information about the subject or environment of the image.
Hays and Efros [37] were among the first to consider the problem of estimating the location of a single image using only its visual content. They collected millions of geo-tagged Flickr images. By using a comprehensive set of visual features, they employed nearest neighbor search in the reference set in order to locate the image. Results show that the approach is able to locate about a quarter of the images (from a test data set) to within approximately 750 km of their true location.

Motivated by [37], Gallagher et al. [30] incorporate textual tags to estimate the geographical locations of images. Their results show that textual tags perform better than visual content but they perform better in combination than either alone. Cao et al. [16] also recognizes the effectiveness of tags in estimating the geolocations. Similar to [30], they combine tags with visual information for annotation, however, they propose a novel model named logistic canonical correlation regression which explores the canonical correlations between geographical locations, visual content and community tags. Unlike [37], they argue that it is difficult to estimate the exact location at which a photo was taken and propose to estimate only the coarse location. A mean-shift based approach is employed to spatially cluster all the images into several hundreds of regions, and then estimate the most likely region for each image. The experimental results show that inference of the coarse location will lead to both meaningful and accurate annotations.

Similar to Cao et al. [16], Crandall et al. [23] only estimate the approximate location of a novel photo. Using SVM classifiers, a novel image is again geo-located by assigning it to the best cluster based on its visual content and annotations. At the landmark scale, both the text annotations and visual content perform better than chance while at the metropolitan scale, only the text annotations perform better than chance. Since landmarks are the most interesting locations, such geo-tagged images have potential to produce tourist maps using geographical annotation techniques [19]. In a recent research work [123] supported by Google, Zhen et al. focused on the landmark recognition. They build a web-scale landmark recognition engine named “Tour the world” using 20 million GPS-tagged photos of landmarks together with online tour guide web pages. The experiments demonstrate that the engine can deliver satisfactory recognition performance with high efficiency. However, it is still an open question whether it is possible to recognize non-landmark locations reliably.

5.3 Other Applications

In addition to the research works discussed above, there are other interesting directions. Agarwal et al. [3] develop an exciting system with a 500 core cluster which matches and reconstructs three dimensional scenes from an extremely large number of photographs collected from Flickr. The results show that three
dimensional reconstruction algorithms can scale with the size of the problem and the amount of available computation. Ji et al. [41] and Wang et al. [110] consider the problem of mining geographical information from web blogs and forums together with images. Quack et al. [84] develop a system for linking images from community photo collections to relevant Wikipedia articles. Cao et al. [12] propose to build a tourism recommendation system using web photos with GPS information.

6. Inference Methods

Inference is a central problem in Multimedia information networks. Since it is impossible to overview all the inferences algorithms, next we will discuss some issues which we believe are important with general interests.

6.1 Discriminative vs. Generative Models

Discriminative and generative models are two groups of machine learning algorithms with different ways of learning models from data. Given input $x$ and their label $y$, generative models aim to learn the joint distribution $P(x, y) = P(x|y)P(y)$. In contrast, discriminative classifiers model the posterior $P(y|x)$ directly, or learn a direct map from inputs $x$ to class labels. Examples of generative models are naive Bayes, Bayesian Network, GMM, HMM, and many graphical models. Discriminative models include logistic regression, SVM, Boosting, and Conditional Random Fields. Generally speaking, when there are enough training samples, discriminative models lead to more accurate classification results than generative models [78]. This is verified by the dominating success of discriminative models in web image retrieval and annotation systems [75] [81]. However, in recent years, there has been a renaissance of generative models for a wide range of multimedia information network applications. Based on our understanding, generative models share the following advantages which make them attractive in many application domains

- Generative models can generate more intuitive interpretation of data samples. Generative models provide an estimation of density functions, from which we can easily determine the marginal distribution in different scenarios. Starting from the $P(x, y)$, we can obtain the conditional distribution $P(x|y)$ and thus obtain the representative samples for each class. Moreover, we can employ the latent topic models [8] [40] to explore the multinomial distributions of coherent factors.

- Generative models can easily handle the missing value problem. In a multimedia information network, it is common that some attributes of a data sample are missing. Generative model can handle such samples easily by using Expectation-Maximization (EM) algorithms to estimate
missing values. In contrast, discriminative models usually have to neglect such samples.

- With generative models, it is possible to incorporate network structure with the model. For example, [13] proposes to model the photo correlations using both visual similarity and the temporal coherence, and construct a network for label propagation. Another example lies in network regularized topic model [67], which combines the topic model with a harmonic regularizer based on social network structure. It has been shown that generative models are flexible, and can handle multi-modal information embedded in the information network, which is preferable in many situations.

6.2 Graph-based Inference: Ranking, Clustering and Semi-supervised Learning

An information network can also be represented as a graph, in which nodes are connected by social network or other links. One of the most famous graph-based learning algorithm is PageRank [79], which has led to a revolution in web searching engines since 1998. The main idea of PageRank is to view a link from one page to another as an endorsement of the landing page. The more links point to a page, the more likely it is relevant. In theory, PageRank is closely related the power method

\[ b^{t+1} = \frac{Ab^t}{||Ab^t||} \]

where \( b \) will converge to the eigenvector of matrix \( A \). Power method is one of the most efficient ways of finding the dominant eigenvector. In recently years, there has been much effort in speeding up the computation of PageRank [46, 35, 66] and robustness to web spam [7] [34]. Bharat and Mihaila proposed the Hilltop algorithm [7], which selects webpages as “good expert” for certain queries, and generates query-dependent authority scores. With a similar motivation, Haveliwala presented Topic-Sensitive Page Rank (TSPR), which computes a set of topic-sensitive PageRank scores for pages using the topics of query keywords [36]. At the query time, TSPR uses linear combination of these scores to generate context-specific importance scores for pages [36].

Like other algorithms such as HITS [51] and SALSA [54], classic PageRank builds the adjacency matrix based on the hyperlinks between web documents. Hyperlinks on the web can be easily added or deleted by web content creators and the PageRank result can be affected by web spammers who purposely create a large number of hyperlinks such as link farms and exchanges. Some researchers proposed to build the adjacent matrix from other types of informa-
tion, including language models [53], page structural information [10], user browsing history [60], and image local patches [44].

Another popular iterative algorithm is also related to the power method. In [124], Zhou et al. proposes to combine partial labels $y$ to estimate the score $\pi$ of each document:

$$\pi^{t+1} = \alpha A \pi^t + (1 - \alpha) y$$  \hspace{1cm} (2.1)

A similar idea is also developed by Zhu et al. [125], where the value of $\pi$ on labeled samples are fixed as $y$ and the resulted method corresponds to the harmonic solution of the un-normalized Laplacian computation on the graph.

The idea of [124] [125] is especially useful for the image tagging problem. When the user-provided tags are noisy and not complete, equation (2.1) can be used to refine the tag and propagate the existing labels to the unlabeled images. Rui et al. generalized label propagation methods to a bipartite graph reinforcement model [88]. Liu et al. proposed a similar approach that aims to automatically rank the tags associated with a given image according to their relevance to the image content [58]. In [120], Zha et al. considered the problem of suggesting queries for joint text and image search.

The graph-based ranking approaches also motivate the research of summarizing photo collections, which often contain too many photos for the users to view in a short time. An intuitive solution is to cluster the entire collection before ranking. This can speed up the ranking algorithms and alleviate the computational burden. In [48], the authors cluster the images before ranking in order to obtain representative images or landmarks from Flickr photos. However, a statistical clustering algorithm often fails to distinguish different semantic groups and does not always improve the retrieval results [105]. In [15], Cao et al. proposes to consider ranking and clustering tasks simultaneously. By ranking the images in a structural fashion, [15] designed the RankCompete algorithm, which discovers the diverse structure embedded in photo collections, and ranks the images according to their similarity among local neighborhoods instead of across the entire photo collection.

Graph-related representations are also widely used in many manifold learning algorithms [87] [116]. These algorithms first construct the graph based on neighborhood similarity, and then project the high dimensional feature vectors into another space which keeps the graph structure.

6.3 Online Learning

Online learning has a long history which can be traced back to Rosenblatt’s work on the perceptron algorithm for linear discriminant functions [86]. Since then, many online learners have been proposed and have shown excellent results. Some examples include the Winnow family of algorithms [57] and the
milestone backpropagation (BP) [90] learning method on Artificial Neural Networks.

In this section, we briefly review some existing online learning algorithms that have been widely used to tackle large scale problems, with an emphasis on multimedia analysis in practice. Our discussion in this section includes a general online learning framework and the corresponding theoretical analysis of the worst case bound on total loss between the online algorithm and their offline counterparts. Following the general online learning framework, we instantiate it with some concrete examples of online learners, including binary classifiers, multi-label classifiers, conditional random fields [92], metric learning and PCA (principal component analysis). These online algorithms have been applied to practical large scale multimedia analysis and related fields and have achieved promising success both in terms of efficiency and effectiveness.

A General Framework for Online Learning. Given a sequence of trials $S = \{(x_i, y_i) | i = 1, \ldots, \ell\}$, online learning algorithms dynamically maintain a sequence of parameters $\{\theta_t\}$ over a hypothesis space $\mathcal{H}$ for $t = 1, \ldots, \ell$. Before trial $t$, the model with parameter $\theta_t$ is learned from the data in the past trials $\{(x_i, y_i) | i = 1, \ldots, t - 1\}$. After the $t$th trial $(x_t, y_t)$, the parameter $\theta_t$ is updated according to rules that gives a new model in $\mathcal{H}$. In general, the model observes a new instance $x_t$ and makes a prediction $\hat{y}(x_t, \theta_t)$ on it. Then the new parameter $\theta_{t+1}$ is obtained, depending on the true outcome $y_t$, the prediction $\hat{y}(x_t, \theta_t)$ and the learning rate $\eta_t$.

To design a proper updating rule, two criteria should be obeyed [50] [82].

- Conservativeness. It ought to preserve the old knowledge that has already existed in the current model, because this knowledge has the rich historical information about previous trials;
- Correctiveness. The performance of the model should be improved on the new trial(s).

Following this idea, the new model parameter $\theta_{t+1}$ is updated by minimizing an objective function

$$F(\theta_{t+1}) = d(\theta_{t+1}, \theta_t) + \eta_t L(y_t, \hat{y}_t(x_t, \theta_{t+1}))$$

where $d$ is a distance measure between two parameterized models with $\theta_{t+1}$ and $\theta_t$. $L$ is a loss function between the true outcome $y_t$ and the prediction $\hat{y}_t$, and $\eta_t$ is the learning rate.

There are many options for the distance measure $d$. For example, when $\mathcal{H}$ is a hypothesis space consisting of linear predictors, $\theta_t$ is a weighting vector. Then $d$ can be a squared Euclidean distance or the relative entropy if $\theta_t$ is a probability vector (i.e., each entry of $\theta_t$ is nonnegative and the sum of all entries is equal to 1). Notice that the latter choice of the relative entropy indicates
that the distance measure $d$ needs not to satisfy the triangle inequality. On the other hand, the loss function $L$ also has many variant forms, depending on the specific problems of interests. Take the example of the aforementioned linear predictors, $L$ can be the squared loss between $y_t$ and $\hat{y}_t$ for regression problems, or the $0/1$ loss for the classification problems. By choosing different $d$ and $L$, different online learners can be defined. In the next section, we introduce some online learners that have already been proposed in the literature.

**Some Examples of Online Learners.** We illustrate some examples of recently developed online learners to help illustrate the principles discussed above in a more concrete way.

**Multi-Label Classifier:** The multi-label classifier was proposed in [82]. In this work, the Kullback-Leibler divergence was used as $d$ in (2.2), and a submanifold $H$ satisfying multi-label constraints set by the current trial $t$ is formed which acts as the loss function in (2.2). Mathematically, the new model is solved by minimizing

$$D_{KL}(p^{t+1}(y|x_t)||p^t(y|x_t)) + \infty(p^{t+1} \in H)$$

where $\infty(p^{t+1} \notin H)$ is $\infty$ if $p^{t+1} \notin H$, otherwise it is 0. $D_{KL}$ is the kullback-leibler divergence.

**Adaptive Support Vector Machine** Another example of an online learner for binary classification problem is given in [117]. In this work, the squared Euclidean distance between the parameters of two successive linear models $w_{t+1}$ and $w_t$ was used as $d$ and the hinge loss on the new trial $(x_t, y_t)$ served as $L$ in (2.2). Hence, the new parameter $w_{t+1}$ was solved by minimizing the following

$$||w_{t+1} - w_t||^2 + \eta_t(1 - y_t w_{t+1} \cdot x_t)^+$$

where $(a)_+ = \max(0, a)$ and $|| \cdot ||_2$ is the $l_2$ norm for a vector.

**Metric Learning** Besides the online learners that involve a sequence of single training examples represented by $(x_t, y_t)$, the work in [24] proposed another form of online learner which updates a metric model with a pair of examples $(x^{(1)}_t, x^{(2)}_t, d_t)$, where $d_t$ is the target distance metric for learning at trial $t$. In their method, the Bregman-divergence over the positive definite convex cone is used to measure the progress between two successive metric models parameterized by Mahalonobis matrix $M_{t+1}$ and $M_t$. Meanwhile, the squared loss between the prediction $\hat{d}_t = (x^{(1)}_t - x^{(2)}_t)^T M_{t+1} (x^{(1)}_t - x^{(2)}_t)$ and the target $d_t$ was used to measure the correctiveness of new model on trial $t$. Formally, $M_{t+1}$ was computed by minimizing

$$D_{ld}(M_{t+1}, M_t) + \eta_t(\hat{d}_t - d_t)^2$$

where $D_{ld}$ is the Bregman distance between two positive definite matrices.
Besides the above concrete examples, many other online learners have been proposed recently. Some examples are the online PCA algorithm in [111], a new stochastic gradient method for SVM [96] and CRF (conditional random fields) [92]. These online learners follow the main idea in (2.2) but may use some approximations to update the model instead of exactly minimizing it.

7. Discussion of Data Sets and Industrial Systems

Multimedia search has been a hot research topic for more than one decade. However, for a long time most of the systems were constructed based on limited amount of data [89], [107]. In recent years, the popularity of online photo sharing community and the development of image search engine makes it possible to collect a large scale dataset. To get a concrete idea, Flickr hosted more than 4 billions of photos in 2009 while Facebook hosted more than 15 billion photos. Researchers have found it easy to build million-scale datasets using web crawler or simply the web service APIs. For example, Torralba et al. [104] have collected 80 Million tiny images in 2008.

On the other hand, it is a tedious work to annotate large scale dataset, which makes the research in this field limited to datasets of relatively small size. For example, Corel image dataset include only 68K images, while the recently popular Caltech 101 [28] dataset owns less than 60 images for most concepts. To overcome this difficulty, two effective approaches are employed to obtain rich annotations. The first approach is to develop an open annotation tool so that the researchers in the field can contribute the labels. Russell et al. in MIT have built such an open annotation interface named “LabelMe” [91], which has earned 111K object-level annotations on 30K images. The second approach is to employ the Amazon Mechanical Turk as the platform for annotation, where internet users contribute labels based on the instructions. Since many internet users treat the labeling process as a fun task, they are happy to label the images with low pay. Sorokin and Forsyth are the first to employ the Amazon Mechanical Turk for image annotation [100]. More recently, with the aid of the Amazon Mechanical Turk, Deng et al. [25] are able to build a much larger dataset named ImageNet with about 10 million images and 15K synsets.

Despite the increasing research interests on social networks, there is no public image data set that contains both image and network structure. However, we can crawl network information from Flickr, Youtube, Facebook, or even Wikipedia. Mislove et al. have built such a network database [70] with no image involved. It will be interesting to enrich such network dataset with visual information.

Motivated in part by the success in academia, more and more industrial systems begin to take multimedia information into account when developing search engines. Although none of the major search engines (e.g., Google, Mi-
crosoft Bing Search) utilized visual feature before 2008, nowadays there have been quite a few attempts to develop large scale visual search systems. TwitPic was launched to allow users to post pictures to follow Twitter post. Tiltomo was developed to search the Flickr dataset based on tags and similar themes. Tineye and Gazopa allow users to provide their own pictures and find similar peers in the internet. Such similar image searching functions have also been supported by Google Image and Bing Image. Moreover, Google has built a beta version of “Swirl” search, which organizes the image search results into groups by hierarchically clustering the visual features. In addition, more and more companies have target the searching problem in mobile platform, and quite a few systems have been developed including Google Goggles, kooaba, snaptell, etc. Another group of companies focus on vertical visual search, which considers a specific segment of visual search, for example, Paperboy considers on searching news articles or books, while Plink focuses on art works search. Table 2.1 summarizes these industrial systems. Although these industrial engines are not mature enough and only index a small portion of the photos all over the internet, they keep improving themselves and provide excellent testbed for evaluating new techniques in multimedia applications.

<table>
<thead>
<tr>
<th>Name</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Image</td>
<td><a href="http://images.google.com/">http://images.google.com/</a></td>
</tr>
<tr>
<td>Bing Image</td>
<td><a href="http://www.bing.com/images/">http://www.bing.com/images/</a></td>
</tr>
<tr>
<td>TwitPic</td>
<td><a href="http://twitpic.com/">http://twitpic.com/</a></td>
</tr>
<tr>
<td>Tiltomo</td>
<td><a href="http://www.tiltomo.com/">http://www.tiltomo.com/</a></td>
</tr>
<tr>
<td>Tineye</td>
<td><a href="https://www.tineye.com/">https://www.tineye.com/</a></td>
</tr>
<tr>
<td>Gazopa</td>
<td><a href="http://www.gazopa.com/">http://www.gazopa.com/</a></td>
</tr>
<tr>
<td>Google Swirl</td>
<td><a href="http://image-swirl.googlelabs.com/">http://image-swirl.googlelabs.com/</a></td>
</tr>
<tr>
<td>Google Goggles</td>
<td><a href="http://www.google.com/mobile/goggles/landmark">http://www.google.com/mobile/goggles/landmark</a></td>
</tr>
<tr>
<td>kooaba</td>
<td><a href="http://www.kooaba.com/">http://www.kooaba.com/</a></td>
</tr>
<tr>
<td>snaptell</td>
<td><a href="http://www.snaptell.com/">http://www.snaptell.com/</a></td>
</tr>
<tr>
<td>Paperboy</td>
<td><a href="http://www.paperboytool.com/">http://www.paperboytool.com/</a></td>
</tr>
<tr>
<td>Plink</td>
<td><a href="http://www.plinkart.com/">http://www.plinkart.com/</a></td>
</tr>
</tbody>
</table>

In summary, we are witnessing a new age when large scale datasets replace the small ones, and when multimedia search engines begin to take shape. We can expect that the research on multimedia systems will obtain significant advance in the next decade.

8. Discussion of Future Directions

It is usually difficult to judge which research direction is most important or will become more popular, especially in an area as fast evolving as multimedia information networks. However, in this paper we still make a few
audacious estimations, with the hope of not being 100% correct, but reflecting some thoughts we wish to be useful for the audience.

8.1 Content-based Recommendation and Advertisement

An impulse force for the fast development of social network and multimedia retrieval lies in the success of web companies such as Flickr, Facebook, Google, etc. For these companies, the most important source of their revenue lies in the online advertisement. We can expect that new advertisement techniques will earn wide market recognition and will attract many research efforts.

Content based recognition and recommendation systems have the potential of locating users’ interests, and connecting advertisement providers with their potential customers. This task is challenging because we need to take into account not only the semantics of the web multimedia, but also the user’s information such as friends, groups, and even viewing history.

The obstruction of content-based research is also closely related to its advantages: since this task is strongly motivated by business profits, the success of recommendation and advertisement system is highly affected by the evolved business model. It is possible that an elegant research algorithm might not be as successful in the field of business.

8.2 Multimedia Information Networks via Cloud Computing

The explosion of multimedia data has made it impossible for single PCs or small computer clusters to store, index, or understand real multimedia information networks. In U.K., there are about 4.2 million surveillance cameras, which means, there is one surveillance camera for every 14 residents. On the other hand, both videos and photos have become prevalent on popular websites such as Youtube and Facebook. Facebook has collected the largest photo bank in the history (15 billion photos in total, with an increasing rate of 220 million new photos per week). It has become a serious challenge to manage or process such an overwhelming amount of multimedia files. Fortunately, we are entering the era of "cloud computing", which provides the potential of processing huge multimedia and building large-scale intelligent interface to help us understand and manage the media content.

Cloud computing, conceptually speaking, describes the new computing interface whereby details are abstracted from the users who no longer have need of, expertise in, or control over the technology infrastructure “in the cloud” that supports them. Cloud computing system includes a huge data storage center and compute cycles nearby. It constitutes front ends for users to submit their jobs using the service provided. It incorporates multiple geographically distributed sites where the sites might be constructed with different structure and
services. From the developer’s viewpoint, cloud computing also reduces developing efforts. For example, working on MapReduce (a programming paradigm in cloud computing) is much easier than using classical parallel message passing interface (MPI). In the past few years, many successful cloud computing systems have been constructed, including Amazon EC2, Google App, IBM SmarterPlanet, and Microsoft Windows Azure.

Despite the fast developing of cloud computing system, most of the current research efforts are spent on designing high performance distributed computing systems and infrastructures of distributed database. There has been only a small portion of work considering on using cloud computing for pattern recognition and machine learning, however, these works are not fit for multimedia data analysis. There are following challenges in employing cloud computing for multimedia information networks:

- **Fast indexing techniques**: unlike the text features which can be effectively retrieved using inverted index, multimedia data usually involves high dimensional features. How to design a fast indexing scheme is crucial for large scale applications. Many researchers have worked on locality sensitive hashing (LSH) [5]. However, LSH is not efficient enough for extremely high dimensional features. Moreover, LSH for optimal performance costs much more space than classical index structures like KD-tree. Another possible approach is to create a large dictionary for visual features, for which inverted index can be applied as well as text words. The limitation of such approach is that much information are lost during the quantization step. It is still an open question how to design the optimal indexing technique.

- **Effective metric learning and feature selection**: The difficulty of multimedia search lies in the gap between low level feature and high level semantics. How to learn the effective feature for different tasks is among the most important problems in the field of multimedia. Moreover, how to learn the metric for low level features is also a crucial problem. In the new era of massive online multimedia data, cloud computing are expected to do a better job by exploring large datasets and associated rich network structure.

**Acknowledgement**

This research was supported in part by HP Innovation Research Program. This research sponsored in part by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government.
The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. Cao is supported in part by CSE fellowship in University of Illinois at Urbana-Champaign. Tsai is supported in part by HP Innovation Research Program and the National Science Council, Taiwan, R.O.C. under contract NSC-095-SAF-I-564-035-TMS.
References


REFERENCES


REFERENCES


