A SEMINAR REPORT

on

WEB AND PERSONAL IMAGE ANNOTATION BY MINING LABEL CORRELATION WITH RELAXED VISUAL GRAPH EMBEDDING

Submitted by

GLAXY GEORGE

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)

ANGAMALY-683577, ERNAKULAM (DIST)

Affiliated to

MAHATMA GANDHI UNIVERSITY

Kottayam-686560

2013

FEDERAL INSTITUTE OF SCIENCE AND TECHNOLOGY (FISAT)

Mookkannoor(P.O), Angamaly-683577

MAHATMA GANDHI UNIVERSITY, Kottayam- 686560

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that this report entitled **Web And Personal Image Annotation By Mining Label Correlation With Relaxed Visual Graph Embedding** is a bonafide report of the seminar presented during 1^{st} semester by **Glaxy George**, in partial fulfillment of the requirements for the award of the degree of Master of Technology (M.Tech) in Computer Science & Engineering during the academic year 2013 - 2015.

Head of the Department

Date :

Place : Mookkannoor

ACKNOWLEDGMENT

If the words were considered as symbols of approval and token of acknowledgement, then let the words pay the heralding role of expressing my gratitude. First and Foremost i praise the God almighty for the grace he showered on me during my studies as well as my day-to-day life activities.

I would like to take this chance to thank the Principal Dr.K.S.M.Panikar, and Chairman of fisat, Mr.P.V.Mathew for providing me with such an environment, where students can explore their creative ideas. Equally eligible is the Head of the department of computer science, Dr. J.C.Prasad for encouraging the students to make these nations true.

I am extremely grateful to the seminar guide,Dr.Arun Kumar, Professor in Computer Science and Engineering Department, for his valuable suggestions for the seminar. I sincerely thank the computer science and engineering faculty for providing us with invaluable help.
Last but not the least, i thank all my families and friends for giving me the help, strength and courage for accomplishing the task.

ABSTRACT

The number of digital images rapidly increases, and it becomes an important challenge to organize these resources effectively. As a way to facilitate image categorization and retrieval, automatic image annotation has received much research attention. Considering that there are a great number of unlabeled images available, it is beneficial to develop an effective mechanism to leverage unlabeled images for large-scale image annotation. Meanwhile, a single image is usually associated with multiple labels, which are inherently correlated to each other.A new inductive algorithm is proposed for image annotation by integrating label correlation mining and visual similarity mining into a joint framework.A graph model is constructed according to image visual features. A multilabel classifier is trained by simultaneously uncovering the shared structure common to different labels and the visual graph embedded label prediction matrix for image annotation.The globally optimal solution of the proposed framework can be obtained by performing generalized eigen-decomposition. The proposed framework is applied to both web image annotation and personal album labeling using the NUS-WIDE, MSRAMM 2.0, and Kodak image data sets, and the AUC evaluation metric. Extensive experiments on large-scale image databases collected from the web and personal album show that the proposed algorithm is capable of utilizing both labeled and unlabeled data for image annotation and outperforms other algorithms.

List of Figures

List of Tables

Contents

Chapter 1

INTRODUCTION

With the development of computer networks and storage technologies,web images has increased. There are large amounts of digital images generated, shared, and accessed on different websites, e.g., Flicker. With the popularity of digital cameras, we are able to create personal photos easily. Consequently, the size of personal albums is getting larger. The growing number of web and personal images requires an effective retrieval and browsing mechanism in either a content- or keyword-based manner. Much research effort has been focused on this area during recent years, resulting in remarkable achievements . Among others, automatic image annotation technology, which associates images with labels or tags, has received much research interest. Automatic image annotation enables conversion of image retrieval into text matching. Indexing and retrieval of text documents are faster and usually more accurate than that of raw multimedia data. Image annotation thus brings several benefits in image retrieval, such as high efficiency and accuracy.

As a way to facilitate image categorization and retrieval, automatic image annotation has received much research attention. Considering that there are a great number of unlabeled images available, it is beneficial to develop an effective mechanism to leverage unlabeled images for large-scale image annotation. Meanwhile, a single image is usually associated with multiple labels, which are inherently correlated to each other. A straightforward method of image annotation is to decompose the problem into multiple independent single-label problems, but this ignores the underlying correlations among different labels. In this paper, we propose a new inductive algorithm for image annotation by integrating label correlation mining and visual similarity mining into a joint framework.

Image annotation is essentially a classification problem. In the field of multimedia and computer vision, many researchers have proposed a variety of machine learning and data mining algorithms for automatic image annotation recently [1],[2]. These works have shown promising achievements in overcoming the well-known semantic gap by applying machine learning algorithms to image annotation. Generally speaking, these approaches can be roughly divided into the following two groups:

The approaches in the first group are usually referred to as a tagging or retrieval-based paradigm. Image tagging approaches usually annotate images by leveraging web images, which are associated with user-defined tags. Typically, tagging approaches can be divided into two phases, i.e., a searching phase and a mining-for-tags phase. Tagging approaches first search for similar images from web-scale data sets and then mine the textual information associated with the retrieved images for image annotation.

Generally, there are three major research issues in image tagging: First, how to design an efficient indexing and matching algorithm for fast search over large-scale web image data sets; second, how to define accurate metrics for the retrieval process; and, third, how to utilize the search results for image tagging. For example, in [2], an efficient hashing scheme is proposed for image tagging. The system in [2] first searches for semantically and visually similar images from the web and then annotates images by mining the search results. In [3], a multiple- feature distance metric learning algorithm was proposed for cartoon image retrieval. Wu et al. proposed a probabilistic distance metric learning scheme for retrieval-based image annotation [4]. Because web images with user-generated tags are comparatively easy to obtain, image tagging has the advantage that less human labor is required. However, the automatically acquired images and tags are essentially noisy and incomplete [5]. Considering that the performance directly depends on the quality of user-generated tags, which are always unseen to the system, the performance of the tagging system is not stable. It remains unclear how retrieval-based image annotation systems will perform when the number of incorrect tags grows.

The approaches in the second group are usually referred to as labeling- or learning-based algorithms. Different from tagging, labeling usually requires some

Figure 1.1: Illustration of semisupervised multilabel learning for image annotation.

training images, which are labeled by human supervisors to learn a classifier for image annotation [5]. In [5], Cao et al. applied canonical correlation analysis (CCA) to web image annotation. Based on -norm regularization, a semisupervised algorithm was proposed in [6] for image annotation. A relevance model between image and word was proposed in [6] for automatic image annotation. Compared with user-generated tags, the labels of training images are clean and more reliable. However, a limitation of labeling is that much human labor may be required to annotate large-scale image repositories.

Usually, a single image may be associated with multiple labels, and the image annotation is a typical multilabel classification problem. A straightforward way to deal with this problem is to decompose it into several binary classification problems, with one for each label. However, the limitation is that this type of approach does not consider correlations among different class labels. Intuitively, such information is helpful for us to better understand image content. For example, the keyword “sea” may often be accompanied with the keyword “beach.” Such information is quite helpful to better understand the multimedia semantics .Thus, another method to reduce the required labor in image labeling is to utilize the label correlation for image annotation. In the field of machine learning and data mining, some researchers have also suggested that incorporating the information of label correlation into multilabel learning is beneficial for a reliable classification result. These research efforts have shown that utilizing class correlation information can improve the performance of multilabel classification in many domains.

In Fig. 1. In part A of the figure, many training images are labeled as “beach” and “sea.” Ideally, the system should learn a pattern that there is a strong relationship between “beach” and “sea.” Then, for the training image in part B, which is labeled as “beach” and “sunset,” the system additionally labels it as “sea.” The unlabeled training image in part C, which is visually similar to the image in part B, is then labeled as “beach,” “sunset,” and “sea.” In that way,both the label correlation and visual information are considered for image annotation during training.

A new framework is proposed for automatic web and personal image labeling by integrating shared structure learning (SSL) and graph-based learning into a joint framework. Compared with other existing algorithms, our algorithm simultaneously utilizes the information in the unlabeled data and the label correlation information.

The rest of this report is organized as follows:In chapter 2 a brief review of related works is presented. In chapter 3 details of proposed framework, whereas in chapter 4 theoretical discussion of the relationship between proposed framework and some other related approaches.In chapter 5 and 6 presents extensive experiments and conclusion respectively.

Chapter 2

RELATED WORK

In this section, some related algorithms are discussed.

2.1 Shared Structure Learning (SSL)

Real-world problems usually exhibit dual-heterogeneity, i.e. every task in the problem has features from multiple views, and multiple tasks are related with each other through one or more shared views. To solve these multi-task problems with multiple views, a shared structure learning framework is proposed, which can learn shared predictive structures on common views from multiple related tasks, and use the consistency among different views to improve the performance. This paper suggests a method for multiclass learning with many classes by simultaneously learning shared characteristics common to the classes, and predictors for the classes in terms of these characteristics. This problem is casted as a convex optimization problem, using trace-norm regularization and study gradient-based optimization both for the linear case and the kernelized setting.

The challenge of accurate classification of an instance into one of a large number of target classes surfaces in many domains, such as object recognition, face identification, textual topic classification, and phoneme recognition. In many of these domains it is natural to assume that even though there are a large number of classes (e.g. different people in a face recognition task), classes are related and build on some underlying common characteristics. For example, many different mammals share characteristics such as a striped texture or an elongated snout, and people's

faces can be identified based on underlying characteristics such as gender, being Caucasian, or having red hair. Recovering the true underlying characteristics of a domain can significantly reduce the effective complexity of the multiclass problem, therefore transferring knowledge between related classes.

Simultaneously learning the underlying structure between the classes and the class models is a challenging optimization task. Many of the heuristic approaches explored in the past aim at extracting powerful non-linear hidden characteristics. However, this goal often entails non-convex optimization tasks, prone to local minima problems. In contrast, modeling the shared characteristics, as linear transformations of the input space. Thus, this model will postulate a linear mapping of shared features, followed by a multiclass linear classifier.

In [7], suggests a method for multiclass learning with many classes by simultaneously learning shared characteristics common to the classes, and predictors for the classes in terms of these characteristics. To address this as a convex optimization problem, using trace-norm regularization and study gradient-based optimization both for the linear case and the kernelized setting.

Multi-task learning (MTL) aims to improve generalization performance by learning multiple related tasks simultaneously. In [8], considers the problem of learning shared structures from multiple related tasks. An improved formulation (iASO) is used for multi-task learning based on the non-convex alternating structure optimization (ASO) algorithm, in which all tasks are related by a shared feature representation. iASO is converted, a non-convex formulation, into a relaxed convex one, which is, however, not scalable to large data sets due to its complex constraints. An alternating optimization (cASO) algorithm is proposed which solves the convex relaxation efficiently, and further show that cASO converges to a global optimum. In addition, a theoretical condition is proposed, under which cASO can find a globally optimal solution to iASO.

Tagged Web images provide an abundance of labeled training examples for visual concept learning. However, the performance of automatic training data selection is susceptible to highly inaccurate tags and typical images. Consequently,

manually curated training data sets are still a preferred choice for many image annotation systems. This paper introduces ARTEMIS—a scheme to enhance automatic selection of training images using an instance-weighted mixture modeling framework. An optimization algorithm is derived to learn instance-weights in addition to mixture parameter estimation, essentially adapting to the noise associated with each example. The mechanism of hypothetical local mapping is evoked so that data in diverse mathematical forms or modalities can be cohesively treated as the system maintains tractability in optimization. Finally, training examples are selected from top ranked images of a likelihood-based image ranking.

Experiments indicate that ARTEMIS exhibits higher resilience to noise than several baselines for large training data collection. The performance of ARTEMIS-trained image annotation system is comparable with usage of manually curated data sets. In recent years, easy access to loosely labeled Web images has greatly simplified training data selection. Search engines retrieve potential training examples by comparing concept names with image labels (user-assigned tags or surrounding text keywords). In this context, a concept is illustrated by all images labeled with the concept name and an image with multiple labels exemplifies co-occurring concepts. The retrieved images could be directly used to train annotation systems, except that they are often irrelevant from a machine learning perspective. Even user-assigned tags are highly subjective and about 50

Tags appear in no particular order of relevance and the most relevant tag occurs in top position in less than 10 percentage of the images . Consequently, several strategies have been proposed to refine retrieved collections. Our approach is based on the observation that the distribution of relevant images has a more regular form compared to noise, thereby resulting in a higher signal to noise ratio at the modes of the distribution as opposed to its boundaries. In that case, the precision of training data selection may be enhanced by tapping the high-likelihood region of the distribution. This in turn evokes a causality dilemma because the distribution parameters cannot be robustly determined without suppressing the effect of outliers and outliers cannot be suppressed without a good reference distribution.

A new instance-weighted mixture-modeling scheme that simultaneously estimates mixture parameters and instance weights. It is named ARTEMIS after Au-

tomatic Recognition of Training Examples for Modeling Image Semantics. In this parametric scheme, the reference model for each concept is a mixture model of visual and textual features computed from images tagged with the target concept. Similar to K-Means, the ARTEMIS initialization stage assigns equal weights to all data instances. However, it then deviates by systematically learning unequal weights to curb the contribution of noisy images in iterative reference model learning. Training data is selected by ranking images in the decreasing order of mixture likelihood.

2.2 Semisupervised Inductive Learning

Although transductive classification is comparatively effective for image annotation, it is not suitable for large-scale image databases whose size grows dynamically. On the one hand, manually annotating many training data is expensive and time consuming. On the other hand, insufficient labeled training data may induce overfitting. To relieve the tedious work in supervised learning, some researchers suggest improving the learning performance by leveraging unlabeled data, e.g., [10] and [11]. Compared with traditional supervised learning algorithms, such as linear discriminant analysis (LDA), this type of algorithm is able to reduce the required number of labeled data during the training stage. Compared with transductive learning, the inductive algorithm is able to predict the labels of unseen data, which are outside the training set. It is therefore more suitable to apply the algorithm to dynamic image database annotation. However, in most of existing semisupervised learning algorithms such as [10] and [11], a linear constraint is imposed on the image labels, whereas data distribution of multimedia data is demonstrated to be more of a nonlinear manifold . It is beneficial to make the classifier more flexible [12]. Moreover, the correlations among different labels are not considered either.

Although it has been studied for years by the computer vision and machine learning communities, image annotation is still far from practical. In this paper, we propose a novel attempt at model-free image annotation, which is a data-driven approach that annotates images by mining their search results. Some 2.4 million images with their surrounding text are collected from a few photo forums to support this approach. The entire process is formulated in a divide-and-conquer framework where a query keyword is provided along with the uncaptioned image to improve

both the effectiveness and efficiency. This is helpful when the collected data set is not dense everywhere. In this sense, our approach contains three steps:

1. the search process to discover visually and semantically similar search results
2. the mining process to identify salient terms from textual descriptions of the search results
3. the annotation rejection process to filter out noisy terms yielded by Step 2.

To ensure real-time annotation, two key techniques are leveraged—one is to map the high-dimensional image visual features into hash codes, the other is to implement it as a distributed system, of which the search and mining processes are provided as Web services. As a typical result, the entire process finishes in less than 1 second. Since no training data set is required, our approach enables annotating with unlimited vocabulary and is highly scalable and robust to outliers.

Recently, some researchers began to leverage Web-scale data for image understanding. Fundamentally, the aim of image auto-annotation is to find a group of keywords that maximizes the conditional distributions .

2.3 Supervised Learning Of Semantic Classes

A probabilistic formulation for semantic image annotation and retrieval is proposed. Annotation and retrieval are posed as classification problems where each class is defined as the group of database images labeled with a common semantic label. It is shown that, by establishing this one-to-one correspondence between semantic labels and semantic classes, a minimum probability of error annotation and retrieval are feasible with algorithms that are

1. conceptually simple
2. computationally efficient
3. do not require prior semantic segmentation of training images.

In particular, images are represented as bags of localized feature vectors, a mixture density estimated for each image, and the mixtures associated with all images

annotated with a common semantic label pooled into a density estimate for the corresponding semantic class. This pooling is justified by a multiple instance learning argument and performed efficiently with a hierarchical extension of expectation-maximization. The benefits of the supervised formulation over the more complex, and currently popular, joint modeling of semantic label and visual feature distributions are illustrated through theoretical arguments and extensive experiments. The supervised formulation is shown to achieve higher accuracy than various previously published methods at a fraction of their computational cost. Finally, the proposed method is shown to be fairly robust to parameter tuning.

2.4 Feature Selection For Multimedia Analysis

While much progress has been made to multi-task classification and subspace learning, multi-task feature selection has long been largely unaddressed. In this paper, we propose a new multi-task feature selection algorithm and apply it to multimedia (e.g., video and image) analysis. Instead of evaluating the importance of each feature individually, our algorithm selects features in a batch mode, by which the feature correlation is considered. While feature selection has received much research attention, less effort has been made on improving the performance of feature selection by leveraging the shared knowledge from multiple related tasks. Our algorithm builds upon the assumption that different related tasks have common structures. Multiple feature selection functions of different tasks are simultaneously learned in a joint framework, which enables our algorithm to utilize the common knowledge of multiple tasks as supplementary information to facilitate decision making. An efficient iterative algorithm is proposed to optimize it, whose convergence is guaranteed. Experiments on different databases have demonstrated the effectiveness of the proposed algorithm.

Chapter 3

PROPOSED FRAMEWORK

To exploit label correlations for image annotation, it is reasonable to assume that different image labels are related and built on some underlying common structures. For example, different photos taken at the beach share common characteristics, including sea, sky, and sand. Inspired by the recent work of SSL , it is assumed that there is a common subspace shared by multiple image labels. The final label of each image is predicted by its vector representation in the original feature space, together with the embedding in the shared subspace. Motivated by the recent success of semisupervised learning [10], [11] we construct a graph model according to image visual features to exploit the unlabeled data and assume that the distribution of image labels is consistent with it. In that way, we integrate SSL and graph-based transductive classification into a joint framework to learn a reliable multilabel classifier. Note that our framework is inductive, making it applicable to large-scale dynamic image databases, which grow dynamically.

3.1 Formulation of Proposed Framework

Define the class indicator matrix of the training images as

$$Y = Y_l^T, Y_u^T \varepsilon \{0, 1\} \tag{3.1}$$

Let y_{il} be the l^{th} element of y_i.Denote x_i as the visual feature of the i_{th} image in the training set.If x_i belongs to the l^{th} class,y_{il}=1 ; otherwise,y_{il}=0 .The prediction

function f_l of the l_{th} label is then defined as

$$f_l(x) = v_l^T x + p_l \Theta^T x = w_l^T x \tag{3.2}$$

where $v_l \varepsilon \Re^{dx1}$ and $p_l \varepsilon \Re^{rx1}$ are weight vectors,$\Theta \varepsilon \Re^{dxr}$ is a transformation matrix of shared low-dimensional subspace,$w_l = v - l + \Theta_{pl}$, and is the dimension of the shared subspace.

A visual graph A is constructed according to image visual features, whose element $A_i j$ reflects the visual similarity between the two images x_i and x_j. Practically,A can be defined as follows to reduce the number of parameters:

$A_{ij} = 1$, if x_i and x_j are k nearest neighbors and 0 otherwise.

3.2 Optimization

Although the proposed framework is nonconvex, the global optimum can be obtained by performing generalized eigendecomposition.In summary, the training process of the proposed LMGE algorithm is listed here.

1. Perform principal component analysis to reduce the dimension of X, in which all the eigenvectors corresponding to the nonzero eigenvalues of the covariance matrix is preserved.

2. Compute C
$$C = I - \beta(XX^T + (\alpha + \beta)I - \mu X B^{-1} X^T) \tag{3.3}$$

3. Compute D
$$D = N^{-1} X B^{-1} U Y Y^T U B^{-1} X^T N^{-1} \tag{3.4}$$

4. Compute W
$$W = (M - \beta\theta\theta^T)^{-1} XF \tag{3.5}$$

3.3 Implementation Issues

3.3.1 Nonsingularity Issue

In the proposed framework need to compute the inverse of several matrices.A proof of theorem is given to show that all matrices to be inversed during the training stage

are invertible.

3.3.2 Computation Complexity

At the training stage, compute the Laplacian matrix L , of which the complexity is dxn^2 . To optimize the objective function, we need to compute the inverse of several matrices and perform eigen-decomposition. The complexity of these operations is $\max(d^3, n^3)$. Note that is usually much larger than d . Thus, the complexity of training process is about n^3. In the experiment , the image annotation performance is satisfactory when the size of training images is 10 000. Once W is obtained, perform $cxdxv$ times multiplications to annotate testing images. Therefore, the image annotation complexity of this framework is approximately linear with respect to v, making it applicable to large-scale image databases. The experiment also demonstrated that proposed algorithm can be applied to large-scale image annotation.

Chapter 4

DISCUSSIONS

This section includes the possible extensions of LMGE proposed in this report and the relationship between proposed framework and some other related works.

4.1 Extensions Of LMGE

In this framework, the least-squares loss is used. Some other loss function, such as hinge loss and logistic loss, could also be used in the framework. However, the optimization for these loss functions is much more complicated. The least-squares loss function achieves comparable performance to other complicated loss functions, provided that appropriate regularization is added. Therefore, employed the least-squares loss as the loss function for its simplicity.

Another possible extension of LMGE is to generalize it to a nonlinear algorithm by utilizing kernel tricks. To this end, first map the image data into a Hilbert space and assume that there is a transformation function that assigns each image datum in one or multiple image label(s).Kernelized principal component analysis is performed as preprocessing. Therefore, framework can be easily extended to a nonlinear one.

4.2 Relationship To Dimension Reduction Algorithms

During recent years, many dimension reduction algorithms have been applied to image classification and retrieval. For example, SDA and applied it to image classification and retrieval,CCA to image annotation . In the following, the relationship between our framework and some other dimension reduction algorithms:

If the orthogonal constraint is removed in and the data are centered, turns to the objective function of SDA . It is easy to see that LDA is a special case of SDA. In addition, as proved in, CCA reduces to LDA in multiclass learning problems. Therefore, the optimal in framework coincides with the optimal projections of SDA, CCA, and LDA.

4.3 Relationship To SSL Algorithms

In this algorithm of SSL has been proposed for multitask learning, whose objective function is to uncover shared structures. The problem is nonconvex, and thus, an iterative approach, i.e., alternating structure optimization (ASO), is proposed to obtain the local optima . More recently, SSL has been applied to multilabel learning.

4.4 Relationship To Transductive Classification

Compared with the transductive classification algorithms proposed in related papers and in this framework is able to annotate the images out of the training set. Moreover, the correlations among different image labels are also exploited in framework, resulting in more accurate labeling results.

4.5 Relationship To Traditional Graph Regularization

In many traditional inductive semisupervised learning algorithms, such as SDA and Laplacian regularized least squares (LRLS), has been frequently employed as graph regularization to exploit the unlabeled data . If analyze this regularization term under this framework, that a linear constraint is imposed on the graph embedded label prediction matrix . In the framework, however, such linear constraint is relaxed by simultaneously minimizing the linear constraint terms .This property, i.e., relaxed linear constraint, makes the framework more flexible and intrinsically different from most of the existing inductive semisupervised learning algorithms.

Chapter 5

EXPERIMENTAL EVALUATION

This section include extensive experiments to test the performance of the proposed framework in terms of web image and personal album labeling.

5.1 Data Set Description

In experiments, two web image data sets are used, i.e., NUS-WIDE [13] and MSRA-MM 2.0 [14], and one personal image collection, i.e., the Kodak Consumer Video Benchmark Data set [15], to test the labeling performance of the framework. In all of the three image databases, an image may be associated with more than one label.

The NUS-WIDE image database collected by Lab for Media Search in the National University of Singapore consists of 269 000 real-world web images crawled from Flickr [13]. Downloaded all the 269 000 images from http://lms.comp.nus.edu.sg/research /NUS-WIDE.htm. Among the 269 000 images, there are 59 653 unannotated images and 209 347 images, which are annotated with ground-truth labels from 81 concepts. The unannotated images have removed and used all of the remaining 209 347 images, along with the ground-truth labels in the experiments.

The MSRA-MM 2.0 image database is collected by Microsoft Research Asia [14]. In MSRA-MM 2.0, there are around 1 million web images acquired by Live Image Search using different predefined queries. The queries are manually classified into eight categories, i.e., "Animal," "Cartoon," "Event," "Object," "Scene," "PeopleRelated," "NamedPerson," and "Misc" [14]. Among the 1 million images, there are 42 266 images that are manually annotated with ground-truth labels. 3

Each image from the MSRA-MM 2.0 database is labeled as positive or negative with respect to each concept. There are 100 concepts in total. Detailed information of this database can be found in [14]. All of the 42 266 annotated images are used from this data set in the experiment.

The Kodak Consumer Video Benchmark Data set is collected by the Eastman Kodak Company from about 100 consumers over a period of one year. There are 5166 images (keyframes) extracted from 1358 consumer video clips. Of these images, 3590 from this database are annotated by students from Columbia University, who are asked to assign binary labels (presence or absence) for each concept. There are 22 visual concepts in total. In the experiment, have used all of the 3590 annotated images.

5.2 Compared Schemes

In the experiment, the proposed LMGE framework is compared with baseline and a number of related state-of-the-art algorithms.

Employed rigid regression (RR) as the baseline algorithm in the experiment.The proposed framework is compared with two dimension reduction algorithms, i.e., CCA , which has been applied to image annotation in recent work [16], and SDA [17], which has been applied to image classification and retrieval in [17]. For CCA and SDA, the two algorithms are first applied to reducing the dimension of the input visual feature vector, and then, RR is performed as a classifier. In addition, we compare our algorithm with the shared structure multitask learning algorithm, i.e., ASO, .Also compared the framework with two recently proposed multilabel classification algorithms, including dimensionality reduction with multilabel classification and SSL . To compare the framework with existing graph-based inductive classification algorithms, additionally compared LMGE with the well-known semisupervised learning algorithm LRLS and the Multilabel Learning by Solving a Sylvester Equation (denoted as SYLVE in this section). Because the experiment setting is inductive and transductive algorithms, such as LGC , are not appropriate for dynamic image databases, so did not compare proposed algorithm with them.

5.3 Experiment Setup and Evaluation Metrics

In the experiment, the three types of visual features are downloaded, including 144-dimension normalized color correlogram, 128-dimension normalized wavelet texture, and 73-dimension normalized edge direction histogram provided by the National University of Singapore [13] to represent the images in the NUS-WIDE data set. Also downloaded the three types of visual features, including 144-dimension normalized color correlogram, 128-dimension normalized wavelet texture, and 75-dimension normalized edge direction histogram provided by Microsoft Research Asia [14] to represent the images in MSRA-MM 2.0 data set. Similarly as in [13], have used the 144-dimension normalized color correlogram, 128-dimension normalized wavelet texture, and 73-dimension normalized edge direction histogram to represent the images in the Kodak data set. The three visual features are concatenated in the experiment to represent the images from the three data sets. The visual feature of each image is centered by subtracting the mean of the visual features of all the training data.

In the experiment, randomly sample n images as training data, of which m images are labeled. The remaining images are used as testing images.For NUS-WIDE and MSRA-MM 2.0 image databases $n = 10000$ are used. Because the Kodak database is much smaller, set $n = 2000$ for this database. Denote c as the number of total labels/concepts.Also set m as $5 \times c$,$10 \times c$, and$15 \times c$ respectively, and report all of the results. Following the convention of image annotation, during the sampling process, each label is guaranteed to appear in at least one image. The experiments are independently repeated ten times to generate different training and testing images, and we report the average results. Statistical significance test is also performed, with a significance level of 0.05.

As proposed both linear and kernel versions of the proposed LMGE algorithm, in this experiment, only the image annotation results from the linear method for the following two reasons: First, all the related algorithms to be compared are linear ones. Second, it is usually a nontrivial task to select proper kernels for different data sets.

5.4 Performance Evaluation

Fig.5.1 shows several image annotation results of the three databases, where the top four keywords are used to annotate the images. It can see from this figure that the proposed algorithm generally works well for the web image and personal image databases. Also observes from Fig. 5.1 that there are some wrong annotation results. For example, the keyword "ski" is wrongly annotated to two images in the third row. One possible reason could be that the visual features are not capable enough to represent the image semantics. An interesting observation is that, for both cases, "sports" is tagged to the images as well. Thus, another possible reason might be that the model overfits the pattern that there is a strong relationship between "sports" and "ski." Therefore, how to deal with overfitting could be a potential research direction in multilabel learning. Next, gives the numerical evaluation of the proposed algorithm.

As before,c denotes the number of labels/concepts in this section. Tables I –III show the mean MicroAUC and mean MacroAUC of ten times independent experiments with standard deviation of different algorithms when $5 \times c$,$10 \times c$, and $15 \times c$ images are labeled, respectively. A significance test (t-test) is performed as well. Results that are significantly better than others are indicated in boldface.

First, Tables $I - III$ shows that proposed framework outperforms all of the other algorithms in terms of mean MicroAUC and MacroAUC. According to the t-test, the algorithm significantly outperforms all of the other algorithms in all of the 18 cases. This observation indicates that the proposed LMGE can effectively learn from both the label correlations and visual similarities of images. It can also be found in the three tables that, as the number of labeled images increases, the performance of all the algorithms gradually improves. For example, when 405 images are labeled, the mean MicroAUC of proposed framework for the NUS-Wide image data set is 0.8835. When the number of labeled images increases to 810 , the mean MicroAUC of proposed framework for the NUS-Wide image data set is 0.8976. If the number of labeled images increases, it can be seen that the MicroAUC reaches 0.9035 when 1215 images are labeled. Note that there are 199 347 testing images in total. Compared with nearly 200 000 testing images, the number of labeled images

is quite small, but the performance is good. The experiment indicates that it is possible to annotate large-scale real-world images by only labeling a comparatively small amount of training images.

Figure 5.1: Image annotation examples. The wrongly annotated keywords are indicated by red color.

Table 5.1: PERFORMANCE COMPARISON (MEAN MICROAUC ± STANDARD DEVIATION AND MEAN MACROAUC ± STANDARD DEVIATION) WHEN $5 \times c$ IMAGES ARE LABELED.

Table 5.2: PERFORMANCE COMPARISON (MEAN MICROAUC ± STANDARD DEVIATION AND MEAN MACROAUC ± STANDARD DEVIATION) WHEN $10 \times c$ IMAGES ARE LABELED.

Table 5.3: PERFORMANCE COMPARISON (MEAN MICROAUC ± STANDARD DEVIATION AND MEAN MACROAUC ± STANDARD DEVIATION) WHEN $15 \times c$ IMAGES ARE LABELED.

5.5 Parameter Sensitivity

There are several parameters to be tuned in the framework. In experiment, it is observed that the performance is not sensitive to the dimension of the shared subspace. When μ is smaller than 0.01, the performance of LMGE is good for the three image databases, yet it might be lager for some other data sets. In addition, also observed that the performance is less sensitive to α and β when they are in the range of $[10^{-4}, 10^{2}]$.Next, we conduct experiments to test the performance sensitivity w.r.t. parameter κ.

Fig. 5.2 shows the performance variance with respect to parameter κ , which specifies the number of κ nearest neighbors to compute the Laplacian matrix. The number of nearest neighbors should not be too large. Thus, κ is set as 10, 15, 20, 25 in this experiment.Fig. 5.2 shows that the performance of LMGE proposed varies slightly, provided that is not large.

Figure 5.2: Precision–recall curve of image annotation results using different k's to compute the Laplacian matrix, when $15 \times c$ images are labeled. (a) NUS-WIDE. (b) NSRA-MM 2.0. (c) Kodak.

5.6 Annotation Time

Lastly, tested the efficiency of the proposed framework for large-scale image annotation. All experiments are run on a server with 2.67-GHz central processing unit (CPU). 5 The algorithm is implemented by Matlab. Note that, once the classifier, i.e.,W in (12), is trained, can use it to compute the label prediction vector of each testing image according to (8) for image annotation. As discussed in Section III, the time complexity of this process is approximately linear with respect to the total number of testing images. Table IV shows the elapsed time (in seconds) to compute the label prediction vectors of all the testing images after is obtained. The proposed algorithm is very efficient to compute label prediction vectors for large-scale image data sets. More specifically, the framework only takes 0.3483 s to compute label prediction vectors for nearly 200 000 images from NUS-WIDE image data sets. Therefore, proposed framework can be applied to annotating large-scale dynamic image data sets.

Chapter 6

CONCLUSIONS

A new framework for web and personal image annotation is proposed to simultaneously mine label correlations and visual similarities by integrating SSL and relaxed visual graph embedding into a joint framework. Different from previous image annotation algorithms, which usually learn the classifiers by minimizing the regularized empirical error, proposed system minimize the prediction error with respect to a graph embedded label prediction matrix. Compared with traditional graph-based inductive classification algorithms, the linear constraint on the label prediction matrix is relaxed, making it more flexible. Moreover, the label correlation has been learned by uncovering the shared structure of different labels.It is shown that, although the problem is nonconvex, a global optimal solution can be obtained by performing generalized eigen-decomposition.the connection between our algorithm and other related works. It is also proved that the algorithm generalizes several well-known dimension reduction algorithms and classification algorithms. According to the discussion, proposed framework has provided a new perspective to analyze traditional graph regularization, which has been frequently employed in previous semisupervised learning algorithms. Extensive experiments are conducted in two realworld web image data sets and one personal image data set. In these experiments, proposed framework has several advantages, such as efficiency and accuracy.

BIBILOGRAPHY

[1] L. Cao, J. Luo, H. S. Kautz, and T. S. Huang, "Image annotation within the context of personal photo collections using hierarchical event and scene models," IEEE Trans. Multimedia, vol. 11, no. 2, pp. 208–219, Feb. 2009.

[2] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma, "Annotating images by mining image search results," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 11, pp. 1919–1932, Nov. 2008.

[3] Y. Yang, Y. Zhuang, D. Xu, Y. Pan, D. Tao, and S. J. Maybank, "Retrieval based interactive cartoon synthesis via unsupervised bi-distance metric learning," in Proc. ACM Multimedia, 2009, pp. 311–320.

[4] L.Wu, S. Hoi, R. Jin, J. Zhu, and N. Yu, "Distance metric learning from uncertain side information with application to automated photo tagging," ACM Trans. Intell. Syst. Technol., vol. 2, no. 2, pp. 13:1–13:28, Feb. 2011.

[5] L. Cao, J. Yu, J. Luo, and T. S. Huang, "Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression," in Proc. ACM Multimedia, 2009, pp. 125–134.

[6] Z. Ma, Y. Yang, F. Nie, J. Uijlings, and N. Sebe, "Exploiting the entire feature space with sparsity for automatic image annotation," in Proc. ACM Multimedia, 2011.

[7] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in Proc. ICML, 2007, pp. 17–24.

[8] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning shared structures from multiple tasks," in Proc. ICML, 2009, pp. 137–144.

[9] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, and Y. Pan, "A multimedia retrieval framework based on semi-supervised ranking and relevance feedback," IEEE Trans. Pattern Anal. Mach. Intell., 2011, to be published.

[10] D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in Proc. ICCV, 2007, pp. 1–7.

[11] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," J. Mach. Learn. Res., vol. 7, pp. 2399–2434, Dec. 2006.

[12]F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," IEEE Trans. Image Process., vol. 19, no. 7, pp. 1921–1932, Jul. 2010.

[13]T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUSWIDE: A real-world web image database from national university of singapore," in Proc. CIVR, 2009, pp. 1–9.

[14]H. Li, M. Wang, and X.-S. Hua, "MSRA-MM 2.0: A large-scale web multimedia dataset," in Proc. ICDMW, 2009, pp. 164–169.

[15]A. Loui, J. Luo, S.-F. Chang, D. Ellis, W. Jiang, L. Kennedy, K. Lee, and A. Yanagawa, "Kodak's consumer video benchmark data set: Concept definition and annotation," in Proc. MIR, 2007, pp. 245–254.

[16]L. Cao, J. Yu, J. Luo, and T. S. Huang, "Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression," in Proc. ACM Multimedia, 2009, pp. 125–134.

[17]D. Cai, X. He, and J. Han, "Semi-supervised discriminant analysis," in Proc. ICCV, 2007, pp. 1–7.

.

Prepared using LaTeX 2ε