TAG EXPANSION METHODS FOR PHOTO-SHARING WEBSITES

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE INSTITUTE OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Sare Gül Sevil

July, 2010

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Pınar Duygulu Şahin(Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Dr. Fazlı Can

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Aydın Alatan

Approved for the Institute of Engineering and Science:

Prof. Dr. Levent Onural
Director of the Institute

ABSTRACT

TAG EXPANSION METHODS FOR PHOTO-SHARING WEBSITES

Sare Gül Sevil
M.S. in Computer Engineering
Supervisor: Asst. Prof. Dr. Pınar Duygulu Şahin
July, 2010

Due to the fast development of affordable digital cameras and the new trend of sharing media through the web, large amounts of images have become available on the Internet. Thus, at a time when a single site alone hosts over 4 billion photos, the necessity of managing these massive numbers of photos for efficient and effective browsing/searching operations has increased.

To properly organize large amounts of data, systems have been using collaborative tagging methods by assigning descriptive words, *tags*, to data and performing text-based search and retrieval operations on these words. Unfortunately, due to various reasons, both the amount and quality of these tags assigned by users are low.

In this work, we present and analyze two applications of tag expansion methods on photo-sharing websites. The purpose of these methods is to assist users for proper tagging at upload time. The goal of the approaches is not to give users a complete set of tags that could be directly used but to give a list of, possibly, incomplete set of tags that would help or guide the users to tag in accordance with the image content. With this assistance, problems such as incorrect tagging and insufficient tagging are expected to be solved.

Keywords: photo tagging, visual similarity, photo-sharing.

ÖZET

RESİM PAYLAŞIM SİTELERİ İÇİN ETİKET ÖNERİM YÖNTEMLERİ

Sare Gül Sevil
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Yrd. Doç. Dr. Pınar Duygulu Şahin
Temmuz, 2010

Son yıllarda dijital kameraların hızlı gelişimi ve sosyal paylaşım sitelerinin kullanımının artması ile büyük oranda resim internet ortamında paylaşılır hale geldi. Tek bir paylaşım sitesinin bile 4 milyardan fazla resim sağladığı bir dönemde, internet ortamındaki bu resimlerin etkin bir şekilde düzenlenmesi ihtiyacı giderek artmaktadır.

Büyük miktarda veriyi rahat bir şekilde yönetmek için, ortak etiketleme yöntemleri sistemler tarafından kullanılmaktadır. Bu yöntemler, verilere kullanıcılar tarafından atanmış, içerik bilgisi barındıran kelimeleri kullanarak düzenleme işlemlerini gerçekleştirmektedirler. Ancak çeşitli sebeplerden dolayı, kullanıcılar tarafından sağlanan bu etiketler, hem içerik hem de miktar olarak yetersiz kalmaktadırlar.

Bu çalışmada, yetersiz ve yanlış etiketleme sorununu çözebilmek amacıyla kullanılacak iki etiket önerim sisteminin incelemeleri sunulmaktadır. Bu sistemlerin amacı, resim yükleme sırasında kullanıcılara, resmin içeriği ile ilgili etiketler önermek ve etiketleme işlemini kolaylaştırmaktır. Yöntemlerin amacı kullanıcılara tamamlanmış bir etiket listesi sunmak değil, onlara yardımcı olacak, resmin içeriği hakkında bilgi veren etiketler önermektir.

Anahtar sözcükler: resim etiketleme, görsel benzerlik, resim paylaşımı.

Acknowledgement

I would like to thank my advisor Pınar DUYGULU ŞAHİN for her contributions to this work.

I am thankful to my jury members Aydın ALATAN and Fazlı CAN for their suggestions on improving this work. I would like to especially express my thanks to Prof. Fazlı CAN; he has always shown concern for our work and shared his valuable time. I will not forget his helpful and positive attitude as he asked me how he could make my life easier one friday evening.

I would also like to thank Selim AKSOY for his direct and indirect guidance throughout the past 2 years. He has been a great role model.

I would like to thank my family - Yasemin, Salih and Bengi - for bearing with me all this time, eventhough I was mostly sharing the stressful parts.

I would like to thank Onur(KÜÇÜKTUNÇ) for his 'maddi and manevi' support to this work. Working with him was a pleasure on my side even when his critisms were exceedingly harsh.

Thank you Daniya! Eventhough you went far far away, you were still here for me.

I would like to thank my cousins İrem and Orhun, my dearest friends Hande(ERKAN), Aslı(KALE), Fırat(KALAYCILAR), Özlem(GÜR), and Bahadır(ÖZDEMİR) for their love and support.

I also thank my friends at Bilkent; Retina Vision and Learning Group, the 3D-Phone Group, Bilkent Dexterous Robotics and Locomotion Group and all my other friends who are not part of any groups.

To keep a promise, I would like to thank my true supporters, Retina 11, 12 and 13. Even at times when I was using them beyond their own capacities, they have not let me down. Thanks guys!

And finally I would like to thank my unit manager Dr. Kıvanç DİNÇER, my project manager Kadriye ÖZBAŞ ÇAĞLAYAN and my team leader Çağlar GÜNEL at TÜBİTAK UEKAE G222, for their respect, patience and support to my efforts during the last year of this research.

This thesis is dedicated to:
my mother, for her guidance and endless love;
my father, for being my example of endless accomplishments;
my sister, to set up an example so that she does better;
and myself, for finally achieving yet another goal . . .

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As media-sharing websites get more and more popular everyday, large amounts of user-generated data become available on the web. Due to the fast development of affordable digital cameras within the last decade, people started to share photos more than any other media type. As of October 2009, Flickr alone hosts more then 4 billion photos[1]. Thus, the necessity of managing these massive numbers of photos for efficient and effective browsing/searching operations has increased.

To properly organize large amounts of data, systems have been using collaborative tagging methods by assigning descriptive words, *tags*, to data and performing text-based search and retrieval operations on these words. Most photo-sharing websites, including Flickr, use this approach by providing their users the option of tagging. Thus with complete reliance on tags given by the users, these systems organize and search for images with certain visual contents using textual descriptions.

Unfortunately there are several issues affecting this approach. First of all, photo tagging is a very subjective task. Different people can assign different tags to the same image. This can be a result of their focus on the different aspects of the image, or simply because different descriptions can be given to the same part of the image; while for some, a tag like 'Joe' might describe a man in an image,

[1] http://blog.flickr.net/en/2009/10/12/4000000000/

others not knowing the man can describe him with words such as 'man with red hat'. When performing search through a large set of images, image specific tags that can only be known by the photographer (him/her)self, would not be useful. Thus subjectivity is an important issue.

Since having highly descriptive tags is required, the number of tags assigned to an image also becomes important. The more descriptive tags there are, the better the system works. However, another issue affecting the approach is that tagging is a very time-consuming task. Unfortunately, an average user tends to tag in a quick and lazy manner. Studies on Flickr have shown that, although some photos contain more than 50 tags, photos with one to three tags cover more than 60% of all photos, [32].

Moreover, there is the issue of improper tagging. While it is never enough to emphasize the importance of the quality of the tags, users continue to improperly tag images in different ways. A common example of improper tagging is misspelling of words. In addition, there are users that do not have correct or sufficient knowledge about the places/objects within an image so they use incorrect words. And last but not least, some users that tag only with the intention of having higher image viewing rates; these users pick frequently searched words (that are not related to the image) as tags so that their image gets viewed more.

Since fully automatic tagging methods have issues of their own and having user opinions in the process of tagging is important, due to the subjectivity and range of the tagging process, a tag suggestion system integrated to photo-sharing websites may be beneficial in encouraging users to add sufficient number of eligible content-related tags for their photos.

In this work, we present an analysis of the application of tag expansion methods on photo-sharing websites, specifically an application on Flickr. Flickr was chosen because, being one of the most widely used photo-sharing websites, it has enough amount of data to display web-scale characteristics. Here we first present several automatic tag expansion methods that suggest additional tags for a given photo with one-three initial tags. We then give a detailed explanation of the difficulties and outcomes of applying these methods on web-scaled data. We than suggest an improved method that is applied on a large dataset of approximately 300,000 images.

The presented algorithms are designed for photo sharing websites where they can be used to assist users in proper tagging at upload time. The methods utilize the visual content of the photo to be tagged, as well as the textual information obtained from the provided initial tags to form and weight a set of candidate tags. The goal of the approaches is not to give users a complete set of tags that could be directly used, but to give a list of, possibly, incomplete set of tags that would help or guide the users to tag in accordance with the image content. The most important characteristic of these suggested tags is their generality, in the sense that they do not carry information that would only be known by the photographer. Suggested tags aim to describe photos by observable concepts.

This thesis offers the following contributions: two semi-supervised systems for tagging photos in photo-sharing websites. The first method is a simple provided for small-scaled experimentation that performs tag expansion using the textual information, provided by an initial tag set, together with visual information available. The second method is a grouping based algorithm designed for large scale sets that uses both textual and visual information during the tagging process. Moreover, a detailed analysis on the nature and problems of web-scale data and the effects of visual content diversity is provided.

Chapter 2

Background

Due to the simplicity and speed of text-based categorization, websites started providing the option of tagging in order to create folksonomies. Since the greatest benefits of tagging are known as recall and search support [21], the increase in amount of labeled data elevated the dependency of provided services on user-defined tags. Of course the quality of the tags directly affect the results so a way of tag management is needed. This challenge is addressed in several recent studies on non-visual media resources such as del.icio.us[1] and last.fm[2]. [11, 4, 24, 40].

Flickr, among the most widely used social media sites, is a rich source for visual data. Due to the limitations of Content Based Image Retrieval (CBIR) methods [33, 30], other than some recent efforts for web-scale retrieval [13], accessing image content through tags is still the most commonly used approach. However, since tagging is a subjective and laborious task, fully or partially automatic generation of tags may be beneficial.

Amongst various automatic tagging and automatic tag recommending systems, one preferred approach is to use group profiling. As extensively analysed in [21], Flickr has some differences compared to other social media resources and

[1] http://del.icio.us

[2] http://last.fm

therefore requires different strategies for automatic tagging. Since personal images are shared in Flickr, images are most commonly tagged only by their owners; as a result tags usually represent personal aspects, and come in limited numbers. Therefore, mining the tag usage from group profiles is not applicable, and usage of a single user's profile is not sufficient to describe the visual content required due to subjectivity in the vocabulary.

Recent studies in automatic annotation of visual data is promising [1, 2, 3, 5, 7, 8, 12, 15, 16, 22, 25, 26, 27, 38, 39]. However, most of these methods usually work on small amount of data compared to the resources on the web, and usually only a few keywords are assigned to images which may not be sufficient to describe the rich visual content of the web images.

In some approaches proposed for web images, only text-based methods are used for generation of tags. In [32], co-occurrence information for tags, obtained from a large pool of Flickr images, are used to recommend additional tags to a photo with initial user-defined tags. Since only text cues are used for generation of new tags without considering visual information, photos having the same initial tags will always be tagged with the same set of new tags [17].

Recently, methods that combine textual and visual techniques have been proposed. In [18], at the first step, images are classified into a set of concepts by a multi-class SVM. At the next step, tags of visually similar photos are propagated. Since the method requires trained concepts, it can only work on specific catagories and is limited.

The work presented in [29] collects geo-tagged images from Flickr and clusters them using textual, visual and spatial cues. Using frequent itemset mining techniques, relevant word combinations are found for each cluster which are then also used to link the clusters to Wikipedia entries.

In some studies tagging by multiple users is simulated through finding similar images that can be considered as the same resource, and then learning the tagging behaviour from the tags associated with this set. The common approach is to find, over a large number of pre-collected pool of images, the neighbor images which are most visually similar to a target image, and then ranking the tags inside this set either using only frequency of tags or by incorporating visual similarities [17, 35, 36, 34].

In these approaches, visual similarities are considered as the first step to collect a set of candidate images from which the tags are used. However, this process requires the construction of a large pool of fixed images (usually in the order of millions), and computation of image similarities in these large collections for obtaining a smaller candidate tag sets.

This cost can be overcome with the help of minimal number of textual cues provided during the query time. A few initial tags allow to collect a set of candidate images which are likely to be relevant to the target image. Visual similarities can then be further considered in order to prune this set.

In [20], given co-occurrence information of tags, for each pair including the user-defined-tag a classifier is trained. The degree of memberships of the images for a list of classifiers are then used to find the relevance of tags for recomendation. This method is very costly since a separate classifier must be built for each pair in a pre-collected co-occurence list.

In [37], starting with a textual query, among semantically similar images, content based image search is used to find images which are also visually similar. This set is clustered, and related words for each cluster are extracted to be used as recommended tags. Since the images are pruned, and then a cluster-based assignment is performed, there is a possibility to loose some important tags.

Compared to these approaches, our proposed method has several differences. In comparison to works such as [32], we are use the visual information in addition to textual information. Unlike methods [18] and [20], our methods do not require a training process. While [29] is working on a restricted set of geo-tagged images, we are working on an unrestricted set of images obtained from photo-sharing websites. Moreover, different from works of [17, 35, 36, 34] that use neighbor voting, and [37] that uses a query based initial step, we suggest to use a grouping process.

Chapter 3

Tag Expansion on Controlled Set

In this chapter we discuss the initial version of our work which has been experimented on a small scale dataset. The description of the method and a brief explanation of the observed results are in the following sections. More detailed explanations of the work can be found in [14] and [31].

In the earlier phase of the work, we have implemented Tag Suggestr, an automatic tag expansion method for photo-sharing websites. Tag Suggestr is a system designed to serve as an interface between users and photo-sharing websites during photo upload. It is applicable to all photo-sharing websites with tagging capability; Tag Suggestr was applied to and experimented on Flickr.

The method used can be summarized as follows: when a user is uploading a photo, the method requires the user to provide 2-3 initial tags that generally describe the image. These initial tags are used to retrieve a set of *relevant photos*; that are potentially content-related to the target photo. Recommended tags are chosen among the distinct tags that come along with the set of relevant photos. While recommending tags, visual similarity between each related photo and the photo to be uploaded is taken into account.

3.1 Method

Figure 3.1 visually describes the Tag Suggestr which can be summarized with the following steps:

Step 0: Obtain target photo and corresponding initial tags from user. Let I_t be the target photo to be uploaded, and $T_{init} = \{t_{init1}, t_{init2}\}$ be the initial tags for this photo.

Step 1: Connect to Flickr server and fetch the first m relevant photos $I_R = \{I_1, ..., I_m\}$ and their corresponding tags $T(I_i)$. Each *relevant* photo must contain the given initial tags T_{init} as a subset of $T(I_i)$.

$$\forall I_i \in I_R,\ T_{init} \subset T(I_i) \tag{3.1}$$

Step 2: Let $T_{unique} = \{t_1, t_2, ...\}$ be the unique set of tags of all relevant photos. By subtracting the set of stopwords $T_{stopwords}$ from T_{unique}, get the candidate tag list $T_{candidate}$, which contains n distinct candidate tags.

$$T_{candidate} = T_{unique} - T_{stopwords} \tag{3.2}$$

Step 3: Extract visual features f_{I_t} for the target photo, and f_{I_i} for all relevant photos. Then, find the weight ω_i representing the visual similarity between the target photo and the i$^{\text{th}}$ relevant photo I_i.

$$\omega_i = \frac{1}{dist(f_{I_t}, f_{I_i})},\ i \in \{1, ..., m\} \tag{3.3}$$

$$\boldsymbol{\omega} = \begin{bmatrix}\omega_1\ \omega_2\ \cdots\ \omega_m\end{bmatrix} \tag{3.4}$$

Generate a binary $m \times n$ matrix C, (where n is number of candidate tags, m is the number of relevant photos). Set C_{ij}, if photo I_i contains tag t_j.

$$C_{ij} = 1 \Leftrightarrow t_j \in T(I_i) \tag{3.5}$$

Multiply each row i with the visual similarity ω_i, sum the columns to get a $1 \times n$ matrix W of tag weights as follows:

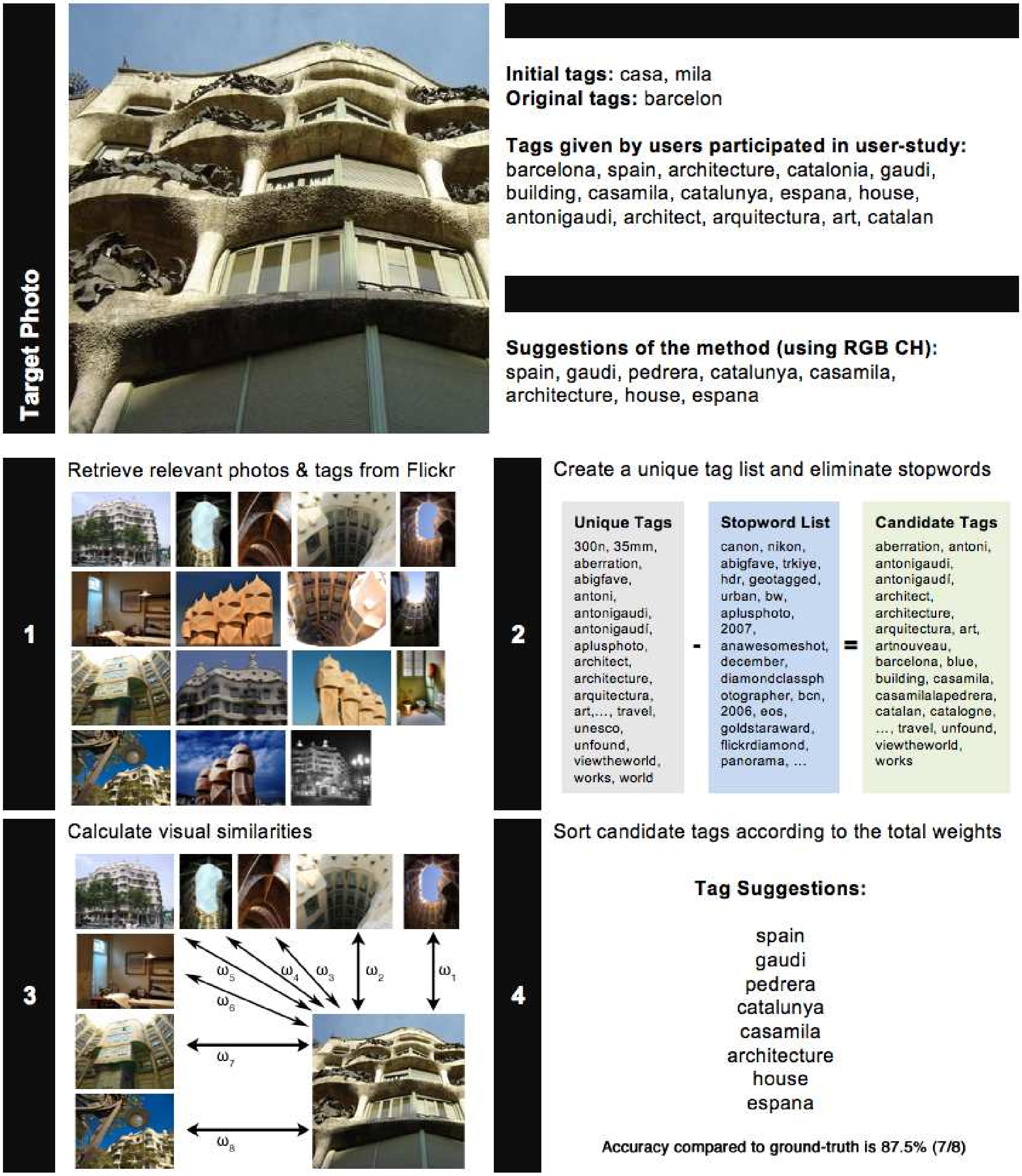

Target Photo
Initial tags: casa, mila
Original tags: barcelon
Tags given by users participated in user-study:
barcelona, spain, architecture, catalonia, gaudi, building, casamila, catalunya, espana, house, antonigaudi, architect, arquitectura, art, catalan
Suggestions of the method (using RGB CH):
spain, gaudi, pedrera, catalunya, casamila, architecture, house, espana
1
Retrieve relevant photos & tags from Flickr
2
Create a unique tag list and eliminate stopwords
Unique Tags
300n, 35mm, aberration, abigfave, antoni, antonigaudi, antonigaudí, aplusphoto, architect, architecture, arquitectura, art,..., travel, unesco, unfound, viewtheworld, works, world
-
Stopword List
canon, nikon, abigfave, trkiye, hdr, geotagged, urban, bw, aplusphoto, 2007, anawesomeshot, december, diamondclassphotographer, bcn, 2006, eos, goldstaraward, flickrdiamond, panorama, ...
=
Candidate Tags
aberration, antoni, antonigaudi, antonigaudí, architect, architecture, arquitectura, art, artnouveau, barcelona, blue, building, casamila, casamilalapedrera, catalan, catalogne, ..., travel, unfound, viewtheworld, works
3
Calculate visual similarities
ω1
ω2
ω3
ω4
ω5
ω6
ω7
ω8
4
Sort candidate tags according to the total weights
Tag Suggestions:
spain
gaudi
pedrera
catalunya
casamila
architecture
house
espana
Accuracy compared to ground-truth is 87.5% (7/8)

Figure 3.1: Overview of Tag Suggestr. For a given target photo and set of initial tags, the method retrieves relevant photos and their corresponding tags from Flickr (Step 1); forms a candidate tag list by eliminating stopwords from the unique tag list of all relevant photo tags (Step 2); computes visual similarities between target photo and relevant photos, then assigns weights to candidate tags using these similarities (Step 3); finally suggests tags according to their weights (Step 4).

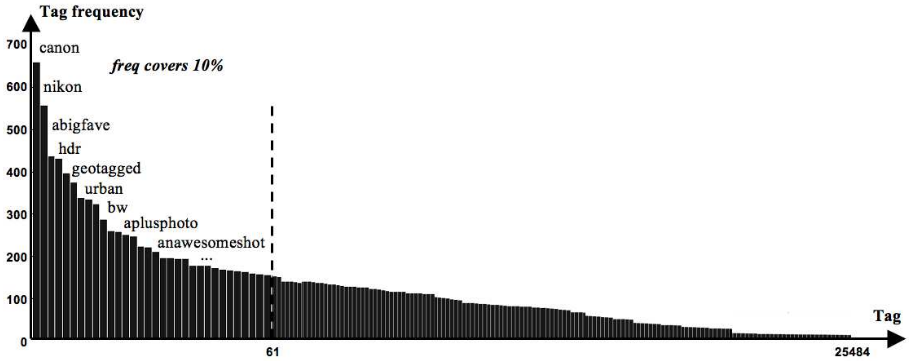


Figure 3.2: Tag frequency graph of all candidate tags in the dataset used for Tag Suggestr. Top 10% of the most frequent tags are considered as *stopwords*.

$$W = \boldsymbol{\omega} * C = \begin{bmatrix} \omega_1 \; \omega_2 \; \cdots \; \omega_m \end{bmatrix} * \begin{array}{c c} \begin{array}{ccccc} t_1 & t_2 & t_3 & \cdots & t_n \end{array} & \\ \begin{bmatrix} 1 & 0 & 1 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} & \begin{array}{c} I_1 \\ I_2 \\ \vdots \\ I_m \end{array} \end{array} \tag{3.6}$$

Step 4: Suggest tags in $T_{candidate}$ according to their total weights W.

3.2 Stopword List

In the method explained in the previous section, there is a step of stopword removal. Tag Suggestr recommends tags by giving weights to candidate tags that are formed from the tag sets of relevant photos. However, these candidate tag sets usually contain many tags that are not related to the image content. Users tend to 'flag' their photos with various kinds of tags such as tags describing camera characteristics. Moreover, there are tags that are specific to the owner of the photograph and therefore subjective. Thus it was found appropriate to remove such tags by forming a *stopword list*, since the objective is to suggest photo-content related tags.

Figure 3.2 shows tag frequency distribution of all 25,484 candidate tags used throughout the experiments of Tag Suggestr. From this distribution, it was observed that the most frequent tags which appear in the top 10% of the entire data cover most of the non-image-content-related and these tags can be grouped as a stopword list.

In contrast to the stopword lists used in document retrieval applications, the most frequently used tags in this set do not include conjunction words. As there are many conjunction words in documents, datasets including these words have higher frequency-cut off percentages. However, high ratios are not realistic for photo tags. Therefore a suitable set of most frequent 61 tags, that cover 10% of all tags, were chosen.

In addition to the most frequent tags, tags related to camera brands (*fuji, olympus, sony, panasonic, lumix*), camera models (*powershot, coolpix*), lens brands (*sigma, nikkor, tamron*), and photo editing software (*photoshop, photomatix, iphotoedited*), and tags with numeric values, such as years (*2008, 2009, etc.*), lens properties (*50mm, 70-300usm, etc.*), camera models (*400d, d200, f30, etc.*), and geotag information (*geo:lat=41363...*) were also eliminated at the stopword elimination step of the method.

3.3 Visual Features

Six visual features, to be used in visual similarity computation, were selected for Tag Suggestr. Descriptions of the features and similarity metrics are as follows.

RGB Color Histogram

RGB color space is quantized into 27 equal subspaces; three bins per band. Visual similarity is defined as one minus the Euclidean distance between two normalized color histograms.

SIFT Descriptors

The SIFT operator [19] is used to extract interest points. From these interest points, similarities between image pairs are calculated by using the matching algorithm provided by Lowe [19]. Then, *number of matching points* between two images is used as a similarity measure.

MPEG 7 Features

MPEG-7 is an ISO/IEC standard that provides a set of multimedia content descriptors [23]. It was designed to functionally represent information about multimedia data for efficient searching in various applications.A set of visual features defined in MPEG-7 standards is selected for calculating the image similarities:

Color Layout Descriptor (CLD) captures the layout information of color feature. Because of its high retrieval efficiency and small computational costs, CLD is preferred in image and sequence matching and sketch queries.

Color Structure Descriptor (CSD) holds both the color content (like a color histogram) and also the structure of this content. CSD provides a better retrieval performance on natural images compared to ordinary color histograms.

Homogenous Texture Descriptor (HTD) provides a precise quantitative description of a texture that can be used for accurate search and retrieval. The computation of this descriptor is based on filtering using scale and orientation selective kernels.

Edge Histogram Descriptor (EHD) represents the spatial distribution of four directional edges and one non-directional edge. It provides better performances on image matching with non-uniform edge distribution.

An MPEG-7 feature extraction library adapted from MPEG-7 XM (eXperimentation Model) reference software is used for extracting MPEG-7 visual features [10]. l_1-norm is used as the similarity measure.

3.4 Experimented Set

For the experiments, 150 arbitrary photos were chosen from Flickr. For evaluation, these *target* photos were retrieved with their corresponding tags, and initial tags were chosen amongst the corresponding tags of each photo. In the process of tag expansion, 100 relevant photos were retrieved per target photo. As a result over 15,000 photos have been processed throughout the evaluations.

Some of the target photos were chosen with respect to their number of corresponding tags. These photos were random in content and had at least 10 tags assigned to them. Most of these randomly chosen photos were unique or rare in visual content, which made them more difficult to tag.

Tag Suggestr is a software realization of how a real person would tag an image with no background knowledge on the image content. When tagging an image with objects, people or locations unfamiliar to the tagger, the tagger would look for similar images with descriptions of the image content and try to identify which textual descriptions can be given to the original image. This is a very instinctive and effective approach; and it is exactly what is performed by Tag Suggestr. Therefore, in order to truely reflect the effects of the system, using rarely found photos is not the best approach.

In order to have a better experimentation process, the remaining target photos were chosen in a different manner; instead of considering the number of tags of images, images with easily describable and popular contents were chosen. These chosen images were taken in the most visited touristic places of five important cities around the world; Barcelona, Florence, Istanbul, London, and Rome.

3.5 Evaluation Process

Judging correctness of suggestted tags is a difficult process for many reasons. First of all, describing visual content requires large number of desciptive tags. But studies on Flickr have shown that, although some photos contain more than

Table 3.1: Minimum, maximum, and average number of original and ground-truth tags

	Min	**Max**	**Avg**
Original tags	0	71	10
Tags given by users	8	40	25

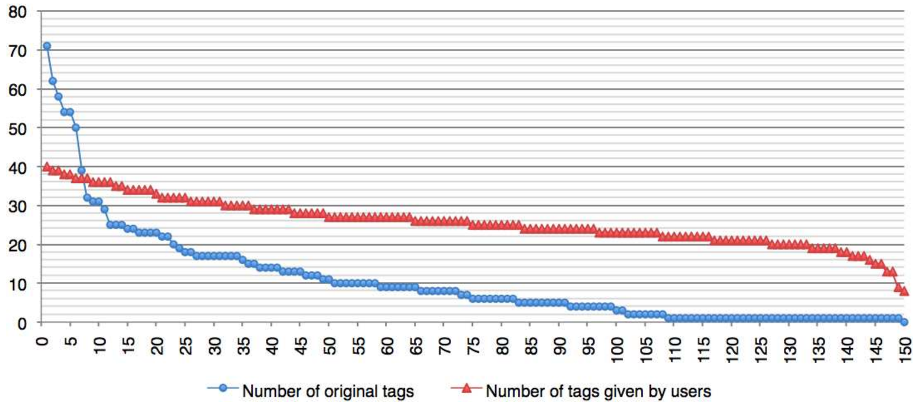


Figure 3.3: Number of original tags vs. number of ground-truth tags as the result of user-study for each photo in target photo set

50 tags, photos with one to three tags cover more than 60% of all photos, [32]. For our target photo set, the average number of tags given to a photo by the photo owners was 10 and there were not many photos with this many tags (See Table 3.1). As shown on Figure 3.3, number of tags per photo is a lot higher and better distributed in the user-study results.

Moreover, the correctness of a tag cannot be judged only by comparing to tags given by a single person. Different people can give different tags relevant to the same image. Thus, just because a tag does not appear in the ground-truth tag list doesn't necessarily mean it is irrelevant to the image content. The best evaluation can be done manually; but even then, subjectivity will be an issue when the tag vocabulary is not limited. In order to have the healthiest result analysis, one needs to analyze images that have been tagged by a number of independent users.

Table 3.2: Average precision of each visual feature for P@8, P@20, and P@25 using user-study tags as ground-truth.

	Baseline **Freq**	**RGB CH**	**SIFT**	**CLD**	**CSD**	**EHD**	**HTD**
P8	60.50	62.17	60.92	62.67	63.00	62.58	55.58
P20	44.10	45.50	44.77	46.07	45.93	45.63	42.60
P25	40.13	41.41	41.20	42.21	41.95	42.13	39.41

For experimental evaluation, two alternative ground-truth tag sets were analized. The first set was formed using original tags. Since the target photos were chosen from Flickr, the tags given by photo owners could be easily obtained and assumed to be valid. However, despite the practicality of obtaining this set, the statistical and observational analysis on the tagging behaviour of Flickr users show its deficiencies. So, to form a more reliable second set, we conducted a user-study in which users chose tags for the target photos from a provided list of tags. These provided tag lists were the candidate tag lists used within the tag weighting step of the suggested method. Results of various comparisons between the two sets have indeed shown that using user-study results as ground-truth was a much more reliable approach.

Table 3.2 shows the average precision values of all 150 photos for the different visual features. Three different suggestion sizes were used in the experiments: suggest top 8, 20 and 25. These values were chosen based on the average number of tags used in the selected ground-truth set. From Table 3.2 it can be seen that using visual similarity has improved the suggestion performance when frequency based suggestion is taken as the baseline approach. For a detailed analysis on these results, please refer to our work in [31].

Chapter 4

Tag Expansion on Large-Scale Set

In this chapter we explain the method that was built upon the knowledge we have gained through the implementation of Tag Suggestr, described in Chapter 3. Tag Suggestr was, by design, applicable to small scale experimentations. In this work our aim is to find and analyze a method that can handle the nature of Web-Scale data in a fully automated fashion. Here we are not trying to find the best algorithm that would solve the problem of automatic tagging of Web images; but rather presenting a detailed analysis of an extended version of our previously published method, which is applicable to web-scale data.

4.1 Improved Automatic Tag Expansion Algorithm for a Larger Dataset

With a fixed dataset of large number of images, the objective is still the same: given an image from the set, what is need is to be able to suggest most relevant tags to that image. In order to do this, a divide and conquer approach was performed. The steps of the method are as follows:

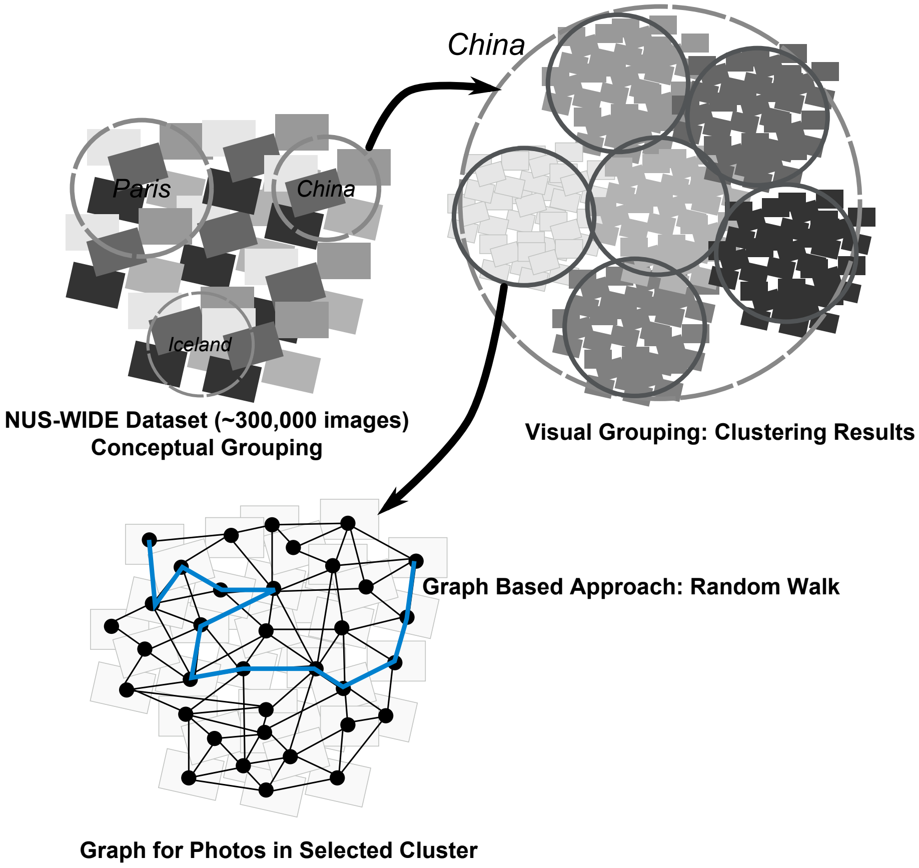

China
Paris
China
Iceland
NUS-WIDE Dataset (~300,000 images)
Conceptual Grouping
Visual Grouping: Clustering Results
Graph Based Approach: Random Walk
Graph for Photos in Selected Cluster

Figure 4.1: Overview of Extended Method

- Divide the dataset into groups of conceptually similar groups using globally used tags.
- Then divide each individual group into subgroups of visually similar images using visual features.
- And finally take the cluster of images which the image to be taged belongs to and apply a graph based approach for tag suggestion(4.1).

The first step is grouping images under textual *concepts*. The *concepts*, which are identified by certain frequently used tags, have a very similar definition to initial tags of Tag Suggestr. While initial tags are used for retrieving relevant images, concepts are used for grouping images under certain categories. Here the concepts are represented by what we call *target tags*; these target tags are chosen such that they cover (i.e. appear in lists of) a large group of images that are conceptually similar.

Once the groups are formed, the next step is to divide the group into clusters of visually similar images. Visual features extracted from all images are used for comparison and a chosen clustering algorithm is used to cluster each group obtained from the previous step.

After the clustering step, a managable size of conceptually and visually similar sets of images are formed. As the final step, the similarity matrices of the images within the cluster are computed and we apply the Random Walk with Restart algorithm([28]) on the cluster for tag suggestion. For the RWR step, the image to be tagged, target image, is used as the starting point. The candidate tag lists are obtained from the tag sets of the other images within the cluster. According to the results of this final step, candidate tags are weighted and tags with highest weights are suggested to the target image.

Different from the previous method, a separate stopword elimination step is not necessary due to the characteristics of the dataset used. The dataset will be explained in further detailed in Section 4.2.

4.1.1 Conceptually Similar Groups: Selection of Target Tags

Working with a set of images with size in the order of hundred thousands for tagging a single image is both not feasible and unnecessary. Knowing that these images are personal photos of web-site users, even a group in the order of hundreds will have considerable diversity. Thus, it is important to be able to categorize the photos.

Tags provide a good starting point for categorization. In order to choose a categorizing tag two things are needed: the meaning of the tag should be general enough to specify certain images but not any image; and the tag should appear in significant number of images. Thus what is required is frequently used and categorizing tags.

Both manual and automatic target tag selection is possible. The common step in both selections is the computation of tag frequencies and choosing among the tags with higher frequencies. In this work, the target tags were chosen manually; selected target tags consisted of names of places or specific concepts. But it is also possible to do this selection automatically with the help of lexical databases such as WordNet.

The selection of target tags will be explained in Chapter 5.

4.1.2 Visually Similar Groups: Clustering

Having reduced image set size from order of hundreads of thousands to a couple of thousands, the second step is to divide the individual categorized groups to visually related sub-groups. Using the visual features extracted from images, a clustering algorithm is used for the division process. Although it is important to obtain good clusters from the clustering algorithm, the aim is not to obtain the best results. Choosing a good clustering algorithm is a difficult task and it requires knowledge of the dataset. Since we are working with a dataset obtained from the web and our main goal is to find an automated solution for web data; thus a general clustering algorithm needs to be chosen and several results can be analyzed for optimal result selection. In this work K-means clustering algorithm was chosen for the clustering step.

In order to obtain optimal results for an unknown set of data, several K values were evaluated to be able to choose the ones with better results. Since it is not possible to know the exact number of real visual clusters of the set of images, the only sources of comparison are tag and image size distributions of clusters. Once the clustering results for several different K values are obtained, these distributions of each cluster are analyzed to find which clustering is giving optimal results. The decision algorithms will be discussed in the following chapter.

4.1.3 Graph Algorithm: Random Walk with Restart

Once a relatively smaller set of visually similar images is obtained, a graph based algorithm, namely Random Walk with Restart(RWR) is used in order to identify which images within the set are more similar to the photo to be tagged. The Random Walk with Restart algorithm is a graph based algorithm widely used in applications on textual data analysis.

To apply a graph based algorithm, a similarity matrix is formed for each cluster, using all images within that cluster. This similarity matrix is then normalized such that row sum equals to 1. RWR algorithm takes this normalized matrix as input.

The RWR process is defined as follows:

$$V_i = (1 - c) * A * V_{i-1} + c * V_{i-1} \tag{4.1}$$

where A is the normalized similarity matrix; c is the restart probability that determines the locality or globality of the random walk process; and $_i$ is the output of ith iteration: a probability vector that carries the similarity information between the starting point image(i.e. image to be tagged).

This process is an iterative process and requires an initial state of the V vector. All elements of the V vector corresponding to an image within the considered cluster; at the initial step, all images except the target image have a value of 0 and target image has the value of 1.

4.1.4 Tag Suggestion

The tag suggestion step of this approach is just like the one described in Chapter 3. A binary matrix is formed where rows represent images and columns represent candidate tags. The values in the matrix indicate whether or not a given tag is in the tag set of an image.

The output of the RWR step, the probability vector, is multiplied by this binary matrix. The weights of each candidate tag is obtained by taking the column wise sum of the matrix. Tags with the highest weights are suggested by the system. The number of tags was chosen by taking the tag set size of the target photos into consideration. The details of the analysis are given in Chapter 5.

4.1.5 Visual Features and Distance Measures

4.1.5.1 Visual Features

For this work we have used the six low-level visual features provided by the NUS-WIDE dataset [6]. Among the six, four are global features containing color, texture and edge information; one is a grid-based feature also giving color information, and one is a bag of visual words obtained from SIFT descriptors [19].

We have also performed the Principal Component Analysis(PCA) for these features to obtain their reduced versions with 0.99 variance. At the feature selection step, results of both versions of each feature was used. The descriptions of the different features are given in Section 4.2.

4.1.5.2 Distance Measures

The graph of used in the graph algorithm step is represented by a similarity matrix. As we have a variety of visual features and different visual features work better with different distance measures; we have used 3 different distance measures for the extraction of the similarity matrix: rectilinear(L1) distance, euclidean(L2) distance and chi-square distance.

4.2 Dataset: NUS-WIDE

To expand our method, we have decided to work with a larger dataset prepared by Chua et. al., [6]. This section gives the details of the dataset. The dataset of

Chua et. al. was formed by crawling 269,648 images and their associated tag sets from Flickr. The dataset includes 6 different low-level features extracted from each image.

4.2.1 Visual Features of NUS-WIDE

The descriptions of the different features that are provided by NUS-WIDE are as follows.

64-D Color Histogram (LAB)

This feature is an adaptation of the LAB(L for *lightness*, A and B for color component dimensions) color space for modeling the color image. Each component of the space is quantized into four bins resulting a feature of 64 dimensions(4x4x4).

144-D Color Auto-Correlogram (HSV)

While color histograms capture the color distribution in images, this feature characterizes these distributions and describes the global distribution of local spatial correlation of colors([9]). For this feature the HSV color components where quantized into 36 bins and the distance intervals were set to four odd intervals generating the feature to be of dimensions 144(36x4).

73-D Edge Direction Histogram

Encoding the distribution of edge directions, the 73-D feature captures the count of edges with directions (quantized at five degree intervals) in the first 72 bins and the last bin captures the pixels that do not correspond to an edge. For edge detection, Canny filter is used; and Sobel operator is used by the gradient of each point for direction computation.

128-D Wavelet Texture

For effectively characterizing different scales of textures, wavelet transforms are used. Using a recursive filtering ans sub-sampling approach, images are decomposed into four frequency sub-bands. The feature vectors are constructed using mean and standard deviation of energy distributions of the sub-bands at different leves. Both Pyramid-Structured Wavelet and Tree-Structured Wavelet Transforms are used. The 128 component features are formed by using 24 component PWT features and 104 component TWT features.

225-D Block-Wise Color Moments (LAB)

For capturing the color distributions of images, three color moments are used. While the first and second order moments give the mean and the variance, the third order moment, skewness, gives a measure of asymmetry degree for the distribution. For the three color moments, three components are used. Due to the compactness of the feature, the moments are extracted for 5x5 grid partitions resulting a feature dimension of 225.

500-D Bag of Visual Words

The visual words are formed in three steps: first scales and key-points are detected by applying Difference of Gaussian filter on the gray scale images; then SIFT descriptors are extracted from the local regions defined by the key-points and scales; and visual vocabulary is constructed using the vector quantization of the SIFT region descriptors. 500 clusters were generated resulting in a bag of 500 visual words.

4.2.2 Tag Analysis

Figure 4.2 shows the tag frequency distribution of Nus-Wide. Initially there were a total of 9,325 unique tags but a noise reduction technique was applied on the tag

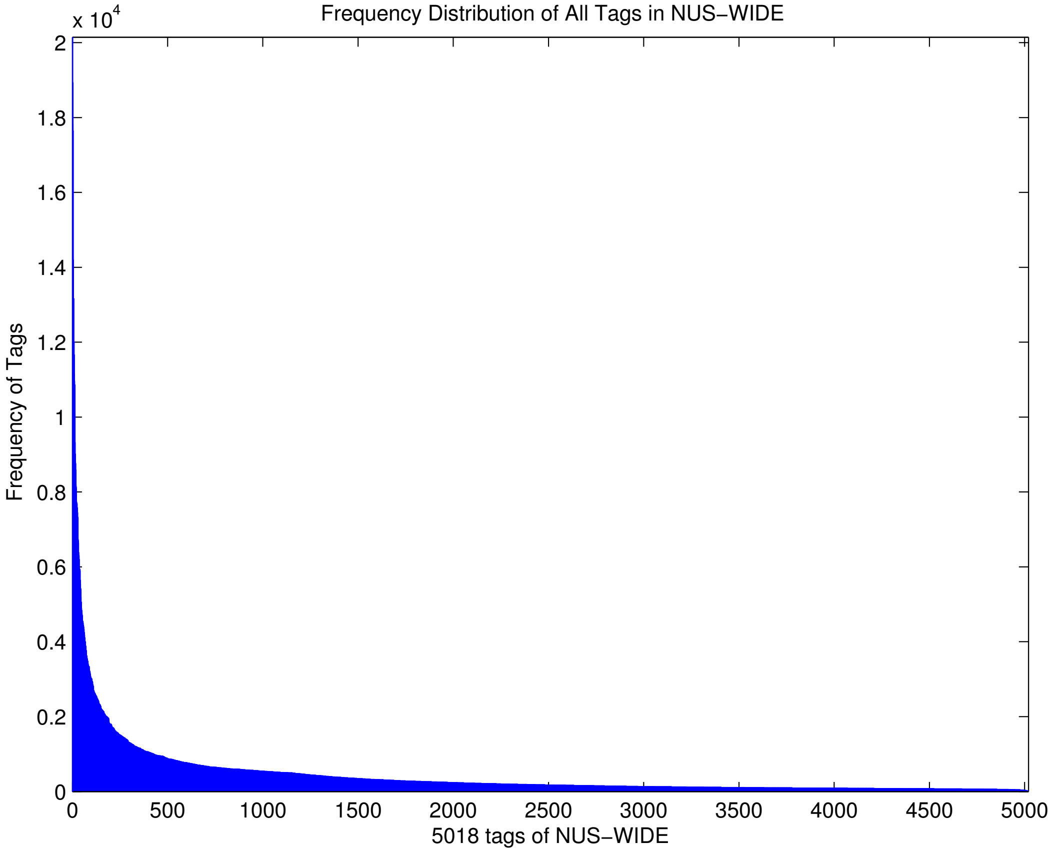


Figure 4.2: Tag Frequency distribution of NUS-WIDE dataset.

Table 4.1: Table of most and least frequently used tags of NUS-WIDE dataset.

Tags	Frequencies	Tags	Frequencies
nature:	20142	rollers:	26
sky:	18935	decadent:	25
blue:	17822	gale:	21
water:	17646	glide:	21
clouds:	14201	beliefs:	20
red:	13172	marvelous:	19
green:	13169	religions:	19
bravo:	12003	portfolios:	14
landscape:	11667	pilgrimages:	12
explore:	11144	famed:	10

sets using WordNet[1] as a source for comparison. The tags that did not occur in WordNet were removed. A final total of 5,018 unique tags were left. In Table 4.1, twenty tags with highest and lowest frequencies have been listed.

4.2.3 Images to be Tagged: Target Images

For each target group, 100 randomly chosen target images were used as input to the system. As a result, a total of 5000 tags have been tagged by the system throughout the experiments.

Appendix A has subset of the selected target images of some of the target groups. By looking at even a small subset of the images, the variety of different images and the diversity of their contents can be noticed.

[1] http://wordnet.princeton.edu/

Chapter 5

Experimental Work

This chapter is organized according to the steps of the method. It includes experimental results, data/result analysis and discussions.

5.1 Target Tags

As explained in the previous chapter, target tags are used in the first step of the algorithm: grouping conceptually similar images. In this work, we have manually selected the target tags. In order to do this, we have computed the frequency of occurrence of each tag and selected our target tags from the top 500 tags which cover approximately 10% of all tags. This percentage was chosen with respect to the tag distribution of images.

For the selection, we focussed on two types of tags: tags specifying location, namely city or country names; and *other* tags, such as underwater or architecture. Of course it is harder to find tags of concepts that can be treated as target tags so the majority of our target tags are specifying location.

There are a total of 50 target tags. The list of the target tags can be seen in Table 5.1.

Table 5.1: List of target tags used. 39 tags specifying places and 11 other tags; totaling 50 tags. (Tags with identifying same location, such as brasil and brazil, where merged during the process; data obtained from both of the tags were used for conceptual grouping.)

Places		**Other**
africa	america	abandoned
amsterdam	arizona	abstract
australia	brasil/brazil	aircraft
britain/england	canada	architecture
chicago	china	death
colorado	florida	industrial
france	germany/deutschland	military
greece	hawaii	photography
iceland	ireland	town
italy/italia	japan	twilight
london	mexico	underwater
michigan	ontario	-
oregon	portugal	-
quebec	scotland	-
seattle	spain	-
switzerland	sydney	-
texas	thailand	-
tokyo	toronto	-
turkey	vancouver	-
washington	-	-

5.2 Target Tag Groups: Target Groups

A *Target Group* is a group of images that include a given target tag within their tag sets. For each target tag, there is a corresponding target group in the system. There are a total of 50 target groups.

The sizes of the Target Groups change between 874(amsterdam) images to 8329(architecture) images. Figure 5.1 shows the image sizes of all groups. Approximately 118,000 images were included by these Target Groups, which means more than 40% of the NUS-WIDE dataset was covered throughout the experiments.

Figure 5.2 shows distribution of the total number of tags used per Target Group. As it can be seen from the figure, the number of tags used are not directly proportional to the number of images in a given group. While with small groups we see tags twice the number of images, for larger groups we see that tags per image size ratio decreases.

From Figure 5.2 it can be seen that as the number of images in target groups increases, after a certain threshold, the tag sizes stablize. Of all the target groups, the group with highest number of tags is ‘photography’. Even though its actual image size, 4290, is almost half as much as the image size of architecture, 8329, which is the largest group, it has significantly higher tag size. This gives a good example on how the broadness of the target tag affects the obtained data. Since photography is a too general topic, the variety of tags increases.

Figure 5.3 shows the tagging statistics for all Target Groups. As it can be seen, the average number of tags assigned to an image is not proportional to the number of images in a group. This shows that the number of tags given to an image is not related to the topic or concept of the image. The maximum number of tags given to photos are close as well. Image with highest number of tags has 178 tags.

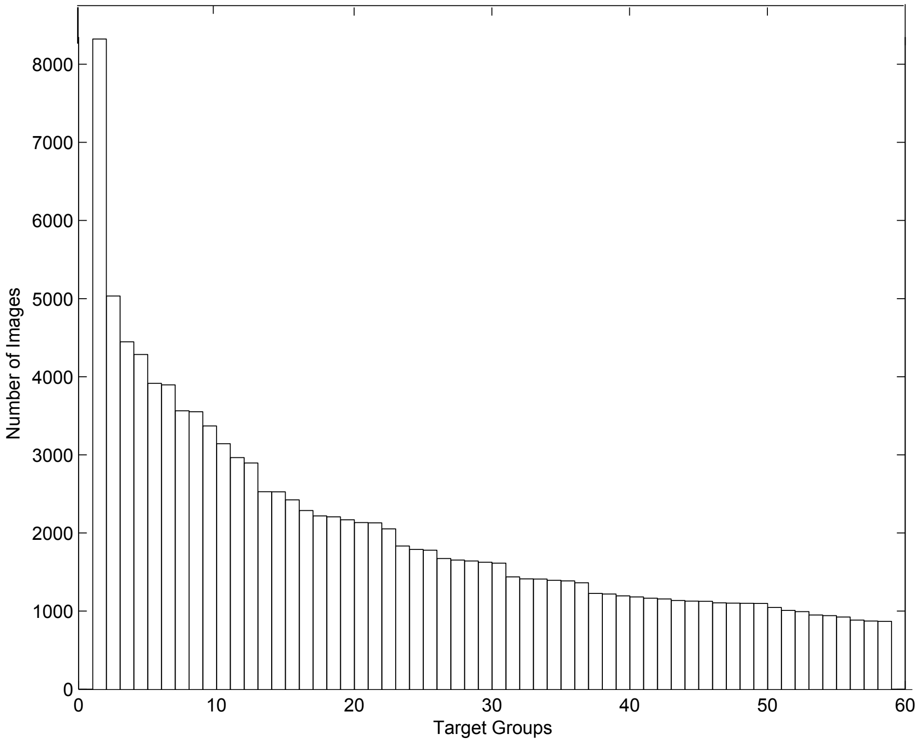


Figure 5.1: Image Sizes of Groups: distribution of number of images per Target Groups used in the experiments.

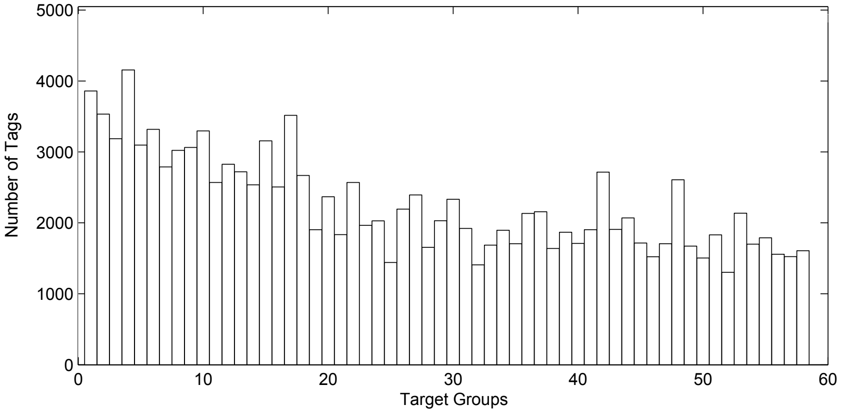


Figure 5.2: Group Tag Sizes: distribution of number of Tags per Target Groups used in the experiments. The Groups are in descending order with respect to their image sizes.

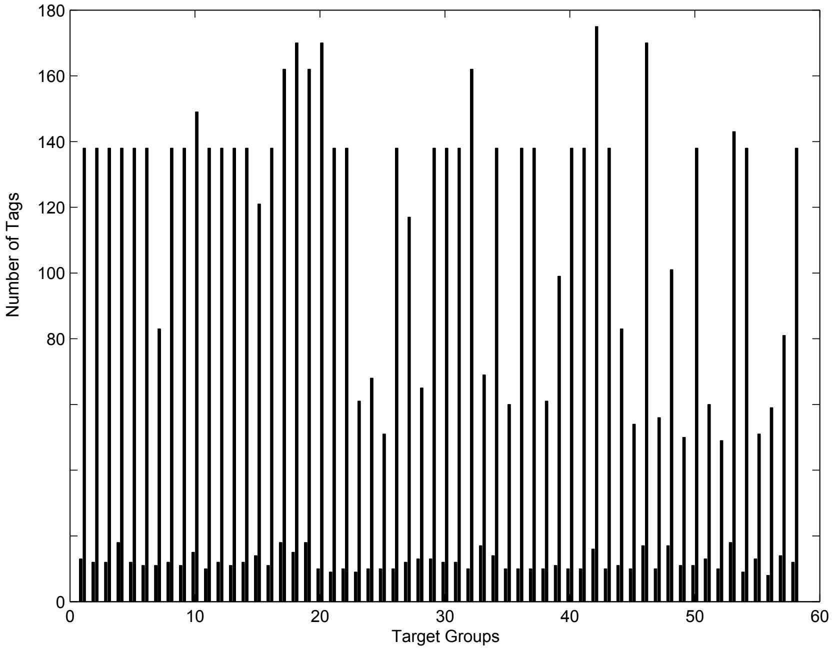


Figure 5.3: Tagging Statistics: the maximum and average number of tags assigned to an image in all Target Groups. The Groups are in descending order with respect to their image sizes.

5.3 Clustering

The visual clustering step aims to narrow down the dataset, so what is expected from the results is having clusters that actually have visually similar images. However, the data used is very diverse by nature and it is not possible both to know and to find the best visual clusters. But a good analysis on a variety of alternative results can help the process of finding optimal clusters.

As explained in the previous chapter, for forming visually similar groups, we have used K-means clustering algorithm. In our experiments, we have extracted the clustering results for 16 different K values. These values were 5, 6, 7, 8, 9, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90, and 100.

To have a better analyis, we have applied the K-means clustering algorithm on a total of 12 visual features. Six of these features were taken from the Nus-Wide dataset directly and additional 6 features have been obtained from applying PCA with variance of 0.99 on the same features. Thus, 192 different clustering outputs were analyzed to select the optimal clustering solution.

Due to the massive size of the dataset and the results themselves, manual selection is a not feasible choice. Furthermore, a manual selection is undesirable due to subjectivity issues similar to the ones described earlier. The only reliable descriptions present on the content of images come from the tags. And the only other information that can be used in the selection process is the image size distributions of clusters.

In this section we first give the analysis of the tag distributions and image size distributions obtained from different clustering results, then we describe our selection process.

5.3.1 Cluster Tag Distribution

The aim of the clustering step is to find clusters with visually similar images. Since it is not possible to manually evaluate the visual correctness of clustering results, one needs to use the descriptive tags as a base of evaluation. With a well distributed tag set, clusters with visually similar images would be expected to have certain dominating tags that only appear in those clusters. One way of doing the tag distribution analysis is as follows:

1. Given tag lists of all images, form Target Group dictionaries and individual cluster dictionaries.

2. While forming these dictionaries, also save the frequency of occurence information.

3. For each cluster obtained, using individual cluster dictionaries, form tag distribution histograms.

4. Using entropy as a diversity measure, compute the entropy values from the cluster tag distribution histograms.

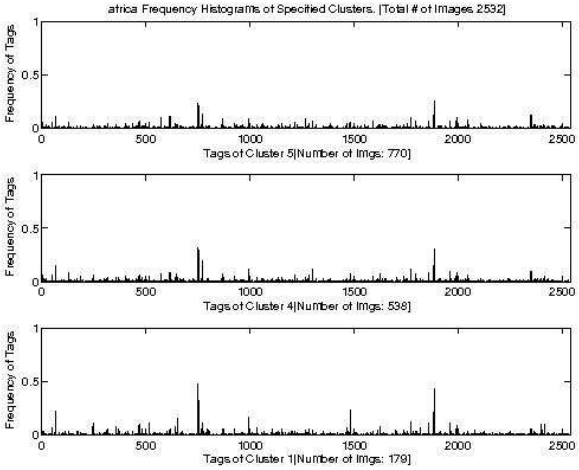


Figure 5.4: Tag frequency distributions for target tag ‘africa’, using K value of 5 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

What is expected is to see clusters with higher diversities amongst their tags as better clusters; diveristy should indicate the presence of ‘cluster identifying tags’. With this information, one could choose the optimal value K value by computing the average of entropies of all clusters of each K. The K value that gives the highest average entropy (i.e. highest tag diversity) could thus be chosen.

Unfortunately, this measure is not applicable to this type of data because the data does not have such ideal characteristics. Figures 5.4, 5.5, 5.6, and 5.7 show the tag frequency distributions of various clustering results. These frequencies have been normalized with respect to the number of images that are within each cluster. The values are in 0-1 range where 1 means the tag occurs in all of the images within the cluster.

The first thing that can be noticed from the figures is that the majority of clusters do not have tags with normalized frequencies above the value 0.5. One can only see high values on very small clusters with less than 20 images. Which means the tags that occur most within a cluster generally do not appear in the set of more than half of the images.

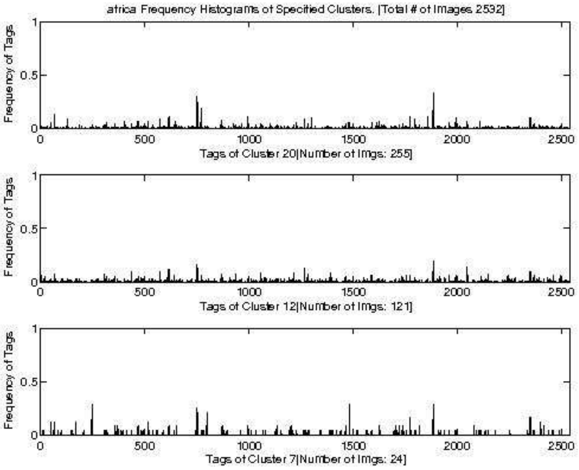


Figure 5.5: Tag frequency distributions for target tag 'africa', using K value of 20 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

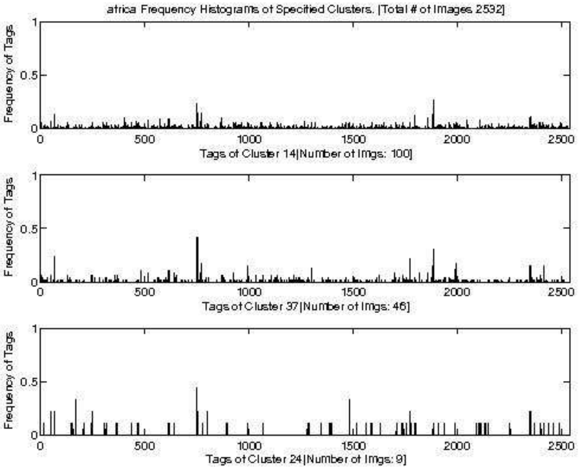


Figure 5.6: Tag frequency distributions for target tag 'africa', using K value of 50 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

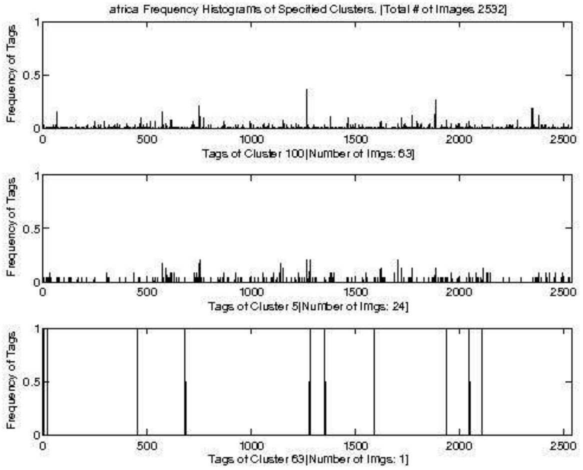


Figure 5.7: Tag frequency distributions for target tag ‘africa’, using K value of 100 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

Another important property of these histograms is that in most cases, the tags with the highest values are not high only within the cluster but also within the other selected clusters. This means that these tags have high occurences in both the clusters and the Target Group itself. Of course not all of the tags with relatively high values are the same within all selected clusters but it is clear that in most cases tags with significantly higher frequencies are not cluster identifying tags.

In order to have a better view of the truely cluster identifying tags, we have computed the TF-IDF values as follows, where T is the tag set of the group and m is the size of the tag set T.

$$TF - IDF(T_i) = \frac{Occurence\ in\ given\ cluster}{Occurence\ in\ all\ clusters},\ i \in \{1, ..., m\} \tag{5.1}$$

In Figures 5.8, 5.9, 5.10, and 5.11, the TF-IDF results for the same clustering results are shown. In these graphs, the TF-IDF values are in the 0-1 unit range where 1 means the tag has only occured in that cluster. In the graphs, the chaotic data in the middle reagon show that there are a lot of tags in clusters that occur

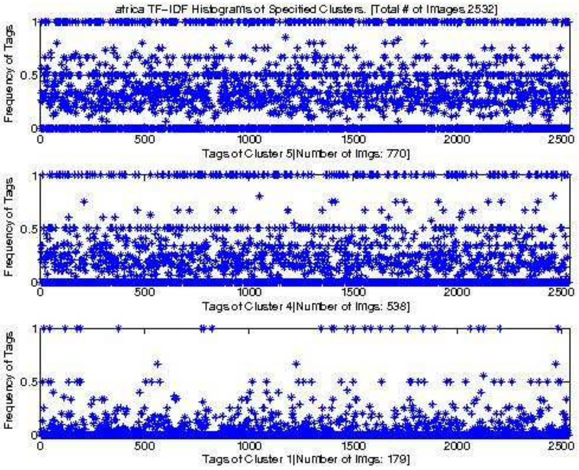


Figure 5.8: TF-IDF distributions for target tag ‘africa’, using K value of 5 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

only in few clusters. It is also clear that there are a lot of tags with TF-IDF values equal to one, which are tags that appear only appear in those clusters. It can be seen that both the chaotic reagons of the graphs and the number of tags with TF-IDF values of 1 reduce as the cluster size decreases.

In clusters with very high image sizes, which occur mostly in small K values, the chaotic reagon is distributed over the middle reagon while with more reasonably sized clusters this reagon is closer to the x-axis. Since the chaotic reagons represent tags that are potentially cluster identifying tags, this shows that smaller clusters have less potentially cluster identifying tags.

If only the tag frequency or the TF-IDF results are going to be used for measuring the diversity of tags in clustering results, clusters with higher sizes would be shown to be good clustering results. However, these are clearly incorrectly divided clusters because their sizes are so large that they start showing the same characteristics of the Target Group itself. Clustering results such as these, completely eliminate the purpose of visual grouping step of the method. Moreover, although the TF-IDF results show significant number of tags that only occur in certain clusters, from the tag frequency histograms we have already seen these

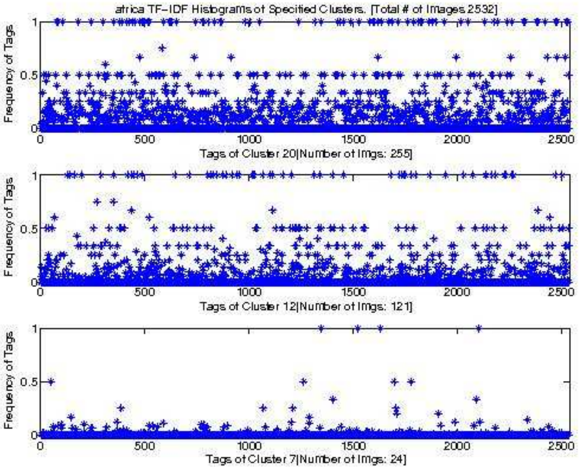


Figure 5.9: TF-IDF distributions for target tag 'africa', using K value of 20 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

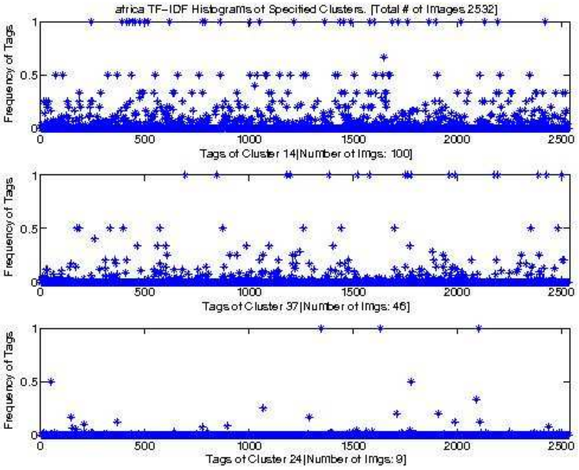


Figure 5.10: TF-IDF distributions for target tag 'africa', using K value of 50 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

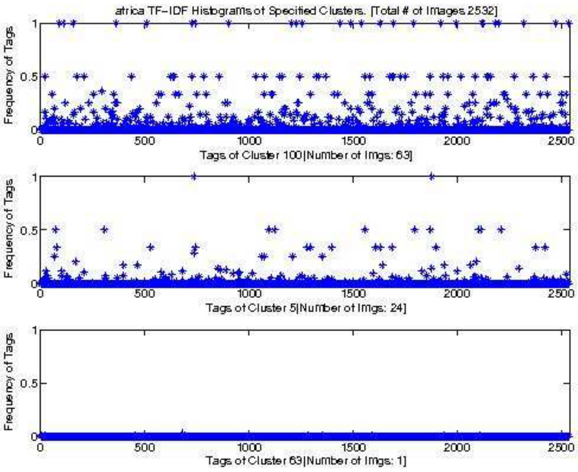


Figure 5.11: TF-IDF distributions for target tag ‘africa’, using K value of 100 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

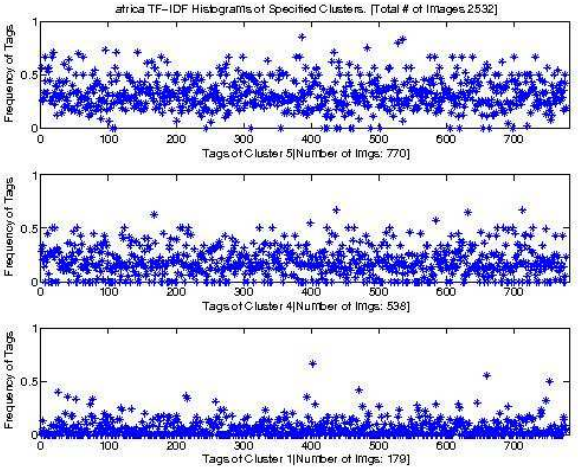


Figure 5.12: TF-IDF distributions after least frequent tag elimination; for target tag ‘africa’, using K value of 5 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

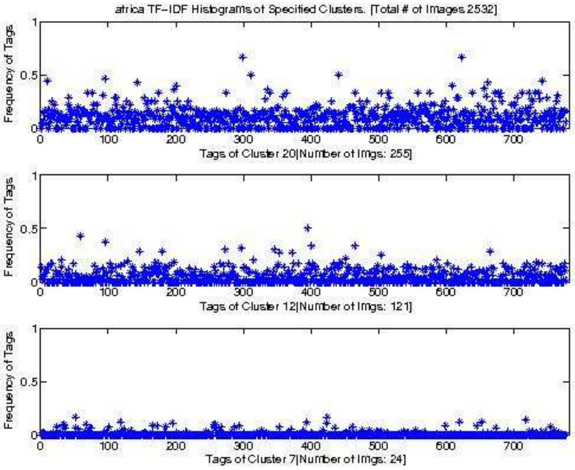


Figure 5.13: TF-IDF distributions after least frequent tag elimination; for target tag ‘africa’, using K value of 20 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

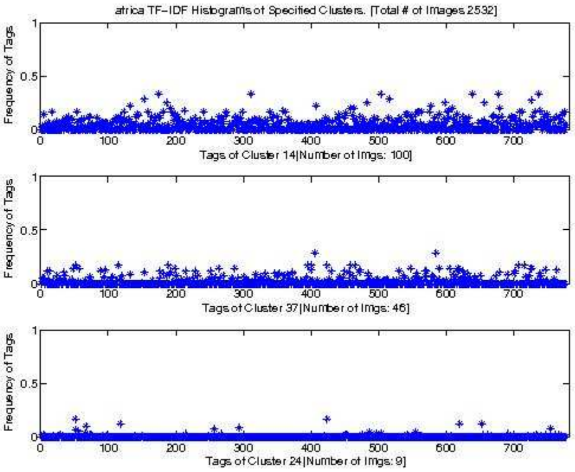


Figure 5.14: TF-IDF distributions after least frequent tag elimination; for target tag ‘africa’, using K value of 50 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

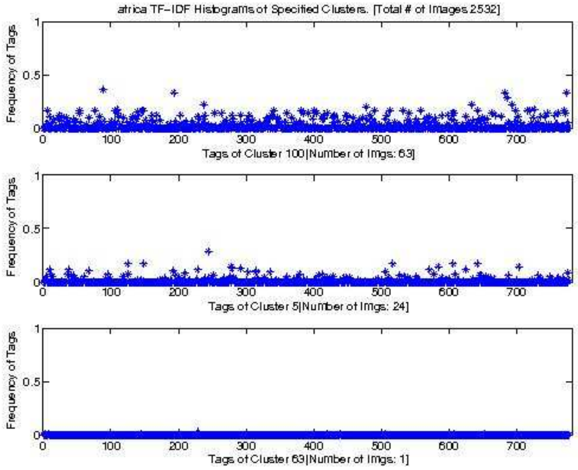


Figure 5.15: TF-IDF distributions after least frequent tag elimination; for target tag 'africa', using K value of 100 and visual feature of PCA reduced Wavelet Texture. Three distributions are given for three selected clusters: cluster with maximum, minimum and median sizes. Cluster IDs and total number images in clusters are given under each graph.

tags can not be cluster identifying tags because most of them are rare tags that have appeared a within a few images of a cluster.

To varify this fact, we made an elimination of tags with low frequencies and re-graphed the TF-IDF graphs. As it can be seen from Figures 5.12, 5.13, 5.14, and 5.15, both the tags with TF-IDF of value 1's and the chaotic reagons have disappeared after the elimination of low frequency tags. This prooves that no matter what features, Target Groups or K values are used, it is not possible get ideal results.

5.3.2 Cluster Size Distribution

Another property that can be analyzed for clustering result selection is the size image distribution of clusters. Using target groups with sizes varying between approximately 800-8000 images, the size distributions of clustering results show common characteristics.

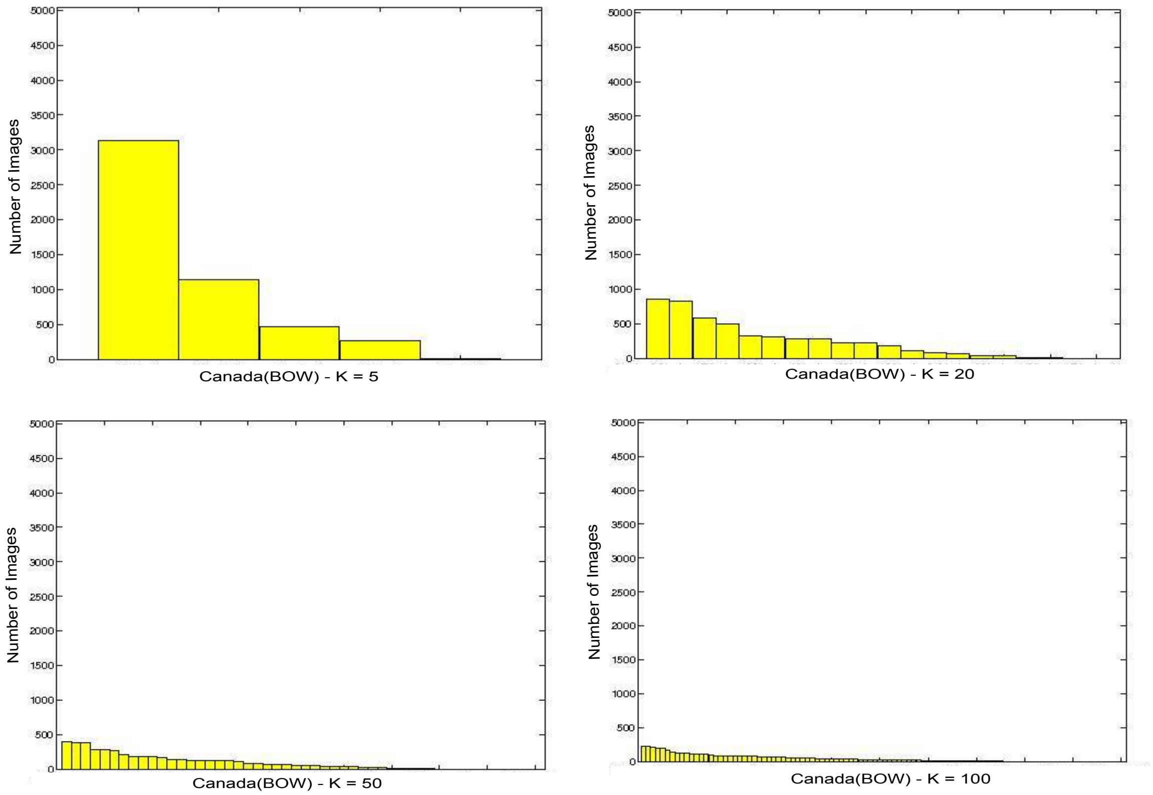


Figure 5.16: Cluster Size Distribution Example: an example for a bad cluster size distribution. Example belongs to the target group Canada. Distributions for 4 different K values are presented. The visual feature used for the clustering results is 500-D Bag of Visual Words(BOW).

While with small K values, the gap between the largest and the smallest clusters are wide, with larger K values, images are more evenly distributed amongst clusters. Although most of the clustering results show similar size distributions, some features show worse results than others.

Figure 5.16 shows an example for these badly distributed cluster sizes. This figure shows the clustering size distributions of the target group ‘canada’ for four different K values, 5, 20, 50 and 100 using the visual featureBag of Words. The ‘canada’ target group has a total of 5039 images.

As it can be seen from the figure, for K value 5, more than half of the images are gathered in one cluster; this cluster is larger than most of the target groups used within the experiments. On the other hand, for K values 50 and 100, there are a lot of clusters with sizes less then 10 images.

From the analysis of tag distributions it was noted that clusters do not have a good visual separation. So working with small cluster sizes is too risky for suggesting proper tags. Conversely, working with significantly large cluster sizes is still not desirable since these clusters are more like target groups than clusters. Thus an optimal solution needs to be selected.

5.3.3 Feature and K Value Selection

Due to the size of the dataset, experimentation using all K values and features is not feasible. So a selection of a single feature and K value for each target group is a requirement. This section explains the selection process of the visual feature and the K values used.

5.3.3.1 Feature Selection

As explained in the previous chapter, we have evaluated 12 features in the clustering process, 6 unique features with both original and PCA reduced sizes. Due to the similarity of cluster image size distributions and the difficulty of choosing respect to tagging statistics, we have decided to perform the feature selection before selecting K values.

In the selection process, features were first selected for each K value of each target group. For a given target group, and the clustering results of a given K value, the standard deviation of cluster size distributions were compared for each visual feature. Because a smaller standard deviation indicates a smaller difference between the sizes of individual clusters, features with the smallest standard deviations were selected.

For final feature selection step, the number of times a feature was selected for a target group-K value pair was counted. Amongst 800 feature selection results, the feature that was selected by most paires was *PCA reduced Normalized Wavelet Texture* feature with 400 selection count.

Throughout the experiments, this visual feature has been used for all target groups.

5.3.3.2 K Value Selection

Like the selection of the visual feature, selection of the K values was based on the cluster size distribution of clustering results. Different from the feature selection, a K values were selected independently for each target group. Since the sizes of the target groups are varying, a single chosen K value for all target groups would not be appropriate. Thus a K value was selected for each target group.

The system requires clusters that have sufficiently high number of images. Having too many images would eliminate the purpose of clustering and not having enough number of images would make the suggestion step ineffective. Taking these into consideration, the selection process was performed as follows:

- Clustering results that have a minimum image size less than 5% of group size or maximum image size greater than 25% of group size are eliminated.
- From the remaining clusters, select clustering result that has the highest min/max ratio.

Table 5.2 presents the K values selected for each target group. As it can be observed from the table, the selected K values are within a small range.

5.4 Tag Suggestion

In order to analyze various results during the experiments, in addition to the full method described in Chapter 4, a combination of the methods described in

Table 5.2: K values selected for target groups.

Target Group	**K Value**	**Target Group**	**K Value**
abandoned:	8	abstract:	7
africa:	8	aircraft:	9
america:	7	amsterdam:	8
architecture:	7	arizona:	7
australia:	7	brasil_brazil:	9
britain_england:	9	canada:	6
chicago:	8	china:	8
colorado:	5	death:	9
florida:	6	france:	6
germany_deutschland:	9	greece:	9
hawaii:	7	iceland:	6
industrial:	9	ireland:	8
italy_italia:	8	japan:	6
london:	6	mexico:	7
michigan:	8	military:	7
ontario:	7	oregon:	7
photography:	7	portugal:	8
quebec:	6	scotland:	8
seattle:	7	spain:	8
switzerland:	8	sydney:	6
texas:	8	thailand:	9
tokyo:	9	toronto:	6
town:	8	turkey:	7
twilight:	6	underwater:	6
vancouver:	6	washington:	8

Chapters 3 and 4 have been used. In this combined method, at the tag suggestion step, instead of using the RWR output, the weighting measure described in Chapter 3 was used.

For the selection of the number of suggested tags, original tag set size of target images were analyzed. From the over all analysis on all target groups, the target images had a minimum of 1, maximum of 138, and average of 11.9 tags in their original tag sets. The median value for tag sizes was 9. By looking at this data, we have decided to suggest 10 tags per image throughout the experiments.

5.5 Performance Evaluation

5.5.1 Performance Metrics

Performance of the method is evaluated using average precision for target groups. The average precision for a target group of m photos is defined as follows:

$$P_n(A_{avg}) = \frac{\sum_{i=1}^{m} P(I_i)}{m} \tag{5.2}$$

where $P_n(I_i)$ is the precision of n suggested tags for a given image I_i. $P_n(I_i)$ is defined as follows:

$$P_n(I_i) = \frac{|OT_i \cap ST_{n,i}|}{n} \tag{5.3}$$

where OT_i is the original tags of target image I_i, $ST_{n,i}$ is the top n tags of its weighted candidate tag list (i.e. suggested tags). As explained in the previous section, n value of 10 was used throughout the experiments.

Table 5.3: List of tags with top 10 frequencies within the NUS-WIDE dataset. These are the tags suggested for any image if a purely frequency based method is used for tag suggestion on the dataset.

nature	sky	blue	water
clouds	red	green	bravo
landscape	explore	-	-

5.5.2 Baseline: Frequency Based Suggestion

In order to have a baseline for comparison, an intuitive approach of suggesting tags with highest frequencies was chosen. This is a very common method relying only on textual information provided by the tags. It's generality and robustness makes it a good choice for performance evaluation.

Of course, for our suggested approach, suggesting tags with highest frequencies would give different results at different steps of the algorithm. In order to have a deeper analysis, the baseline method will also be used for observing the effects of the individual steps of the proposed method.

The frequency based suggestion can be applied at different stages of the algorithm. It can be applied before the first step of forming conceptually similar groups; but this will not be very meaningfull since any image within the dataset will be suggested the same top-10 tags (see Table 5.3). As it can be seen from the table, these tags are too general and diverse in their meanings.

The average precision of suggesting these tags to all target images is listed for each target group in Table 5.4. As expected, their performance are significantly low.

The frequency method can also be applied at two other steps of the approach: forming target groups and forming individual clusters obtained after visual clustering. Thus the results of the proposed approach are compared with the frequency tag suggestions for each conceptual group and with the frequency tag suggestions for the visual cluster that contains the image to be tagged.

Table 5.4: Average precision values obtained by suggesting top 10 frequency tags of Nus-Wide to target images. Precisions are listed for individual target groups.

Target Group	**Average Precision**	**Target Group**	**Average Precision**
abandoned:	0.045	abstract:	0.087
africa:	0.06	aircraft:	0.057
america:	0.05	amsterdam:	0.063
architecture:	0.064	arizona:	0.117
australia:	0.073	brasil_brazil:	0.063
britain_england:	0.08	canada:	0.084
chicago:	0.036	china:	0.049
colorado:	0.075	death:	0.058
florida:	0.069	france:	0.056
germany_deutschland:	0.074	greece:	0.075
hawaii:	0.065	iceland:	0.089
industrial:	0.05	ireland:	0.087
italy_italia:	0.098	japan:	0.054
london:	0.04	mexico:	0.053
michigan:	0.07	military:	0.039
ontario:	0.074	oregon:	0.107
photography:	0.086	portugal:	0.076
quebec:	0.079	scotland:	0.071
seattle:	0.061	spain:	0.055
switzerland:	0.082	sydney:	0.059
texas:	0.05	thailand:	0.067
tokyo:	0.041	toronto:	0.037
town:	0.07	turkey:	0.078
twilight:	0.177	underwater:	0.082
vancouver:	0.061	washington:	0.041

Table 5.5: List of tags with top 10 frequencies within the target group *underwater*.

scuba	diving	water	fish
ocean	sea	blue	coral
dive	nature	-	-

Table 5.6: List of tags with top 10 frequencies within the target group *texas*.

austin	houston	dallas	sky
clouds	blue	explore	night
red	architecture	-	-

Tables 5.5, 5.6, and 5.7 show the frequency tag suggestions for five of the conceptual groups. As it can be seen from the lists, the suggestions made for the individual groups are slightly better then the too broad list obtained in table 5.3. These tags give more information about the variety of images within the groups but are still general for specific images.

Table 5.8 shows the performance of suggesting top 10 frequency tags from target groups. As it can be seen, these results have relatively higher precision values as opposed to the ones presented in Table 5.4, but still low in general. There are only two groups that show exceptionally high performance: aircraft and underwater. These groups are exceptional because they have significantly small variety of visually different photos and have photos that are mostly tagged with widely used generally descriptive tags.

Table 5.9 shows the performance of suggesting top 10 frequency tags from

Table 5.7: List of tags with top 10 frequencies within the target group *canada*.

ontario	quebec	toronto	nature
alberta	water	sky	snow
landscape	winter	-	-

Table 5.8: Average precision values obtained by suggesting top 10 frequency tags from target groups.

Target Group	**Average Precision**	**Target Group**	**Average Precision**
abandoned:	0.095	abstract:	0.122
africa:	0.136	aircraft:	0.395
america:	0.084	amsterdam:	0.155
architecture:	0.124	arizona:	0.178
australia:	0.151	brasil_brazil:	0.19
britain_england:	0.109	canada:	0.131
chicago:	0.112	china:	0.092
colorado:	0.12	death:	0.118
florida:	0.114	france:	0.12
germany_deutschland:	0.142	greece:	0.141
hawaii:	0.164	iceland:	0.11
industrial:	0.13	ireland:	0.122
italy_italia:	0.175	japan:	0.117
london:	0.11	mexico:	0.064
michigan:	0.13	military:	0.119
ontario:	0.218	oregon:	0.146
photography:	0.115	portugal:	0.123
quebec:	0.215	scotland:	0.093
seattle:	0.121	spain:	0.099
switzerland:	0.2	sydney:	0.223
texas:	0.092	thailand:	0.2
tokyo:	0.214	toronto:	0.099
town:	0.161	turkey:	0.136
twilight:	0.278	underwater:	0.322
vancouver:	0.105	washington:	0.091

the clusters of individual target images. From the table it can be seen that for only one group precisions get close to 50% and are generally below that value. Knowing that there is no limit to the variety of tags that can be given to an image, when compared to a limited set of fixed ground-truth tags, these average precision values are exceptable.

Since the number of images of the clusters are significantly large and the original tag sets that used have significantly low sizes with a high percentage of widely used general descriptions , it is not possible for the base-line suggestions not to have low hit rates on the ground-truth; the exceptably good statistical performance of base-line results are understandable. However, these statistical results do not truely reflect the performance since the ground-truth set deliberately favors the frequency results.

5.5.3 Method Suggestions: Similarity Based Suggestion Results

Tables 5.10 and 5.11 show the performance results of suggesting tags using visual similarity between images during the tag weighting step. Performances using 3 different distance measures are presented.

Like the base-line performance results, the performance of similarity based suggestion are relatively good and generally pretty close to the base-line performance. Again the highest precision value is less than 50% and in groups with high variety of image content, such as 'hawai', the presicions are relatively lower.

For the evaluated distance measures, a small difference between the average precision values are observed. However, the differences change according to the content of the target groups and for groups such as architecture, industrial, japan and onorio, the three distance measures show up to 1% of difference.

Table 5.9: Average precision values obtained by suggesting top 10 frequency tags from clusters of target images.

Target Group	**Average Precision**	**Target Group**	**Average Precision**
abandoned:	0.198	abstract:	0.225
africa:	0.234	aircraft:	0.498
america:	0.199	amsterdam:	0.26
architecture:	0.218	arizona:	0.279
australia:	0.256	brasil_brazil:	0.304
britain_england:	0.209	canada:	0.226
chicago:	0.207	china:	0.198
colorado:	0.229	death:	0.24
florida:	0.232	france:	0.225
germany_deutschland:	0.24	greece:	0.232
hawaii:	0.262	iceland:	0.212
industrial:	0.283	ireland:	0.225
italy_italia:	0.279	japan:	0.225
london:	0.215	mexico:	0.177
michigan:	0.239	military:	0.244
ontario:	0.336	oregon:	0.242
photography:	0.212	portugal:	0.23
quebec:	0.36	scotland:	0.197
seattle:	0.236	spain:	0.205
switzerland:	0.311	sydney:	0.335
texas:	0.209	thailand:	0.29
tokyo:	0.336	toronto:	0.214
town:	0.271	turkey:	0.247
twilight:	0.405	underwater:	0.419
vancouver:	0.219	washington:	0.211

Table 5.10: Average precision values obtained by suggesting top 10 highest weighted tags using suggested method with similarity based weighting process. (Part 1)

Target Group	**Chi-Square**	**L1**	**L2**
abandoned:	0.176	0.176	0.177
abstract:	0.211	0.202	0.217
africa:	0.217	0.214	0.218
aircraft:	0.484	0.483	0.489
america:	0.183	0.176	0.186
amsterdam:	0.246	0.242	0.252
architecture:	0.212	0.216	0.216
arizona:	0.264	0.249	0.264
australia:	0.242	0.242	0.245
brasil_brazil:	0.284	0.283	0.289
britain_england:	0.199	0.196	0.203
canada:	0.202	0.207	0.212
chicago:	0.196	0.189	0.196
china:	0.179	0.17	0.185
colorado:	0.22	0.214	0.217
death:	0.218	0.215	0.217
florida:	0.231	0.227	0.23
france:	0.216	0.216	0.216
germany_deutschland:	0.232	0.231	0.235
greece:	0.222	0.218	0.226
hawaii:	0.252	0.253	0.256
iceland:	0.194	0.192	0.193
industrial:	0.242	0.243	0.257
ireland:	0.214	0.215	0.215
italy_italia:	0.276	0.27	0.278
japan:	0.215	0.212	0.215
london:	0.201	0.199	0.205
mexico:	0.161	0.161	0.164
michigan:	0.224	0.222	0.232
military:	0.233	0.231	0.238
ontario:	0.307	0.299	0.313
oregon:	0.217	0.218	0.219
photography:	0.198	0.201	0.196
portugal:	0.209	0.202	0.207
quebec:	0.339	0.339	0.342
scotland:	0.181	0.187	0.186

Table 5.11: Average precision values obtained by suggesting top 10 highest weighted tags using suggested method with similarity based weighting process. (Part 2)

Target Group	**Chi-Square**	**L1**	**L2**
seattle:	0.197	0.195	0.202
spain:	0.184	0.189	0.186
switzerland:	0.289	0.285	0.293
sydney:	0.305	0.297	0.309
texas:	0.197	0.197	0.202
thailand:	0.273	0.276	0.272
tokyo:	0.312	0.31	0.323
toronto:	0.204	0.204	0.205
town:	0.256	0.251	0.262
turkey:	0.224	0.221	0.229
twilight:	0.395	0.396	0.397
underwater:	0.414	0.414	0.421
vancouver:	0.194	0.198	0.195
washington:	0.181	0.181	0.195

5.5.4 Method Suggestions: RWR Suggestion Results

With performances of less than 44% percent, though used in various other applications using textual information, was not as effective for this method. The suggested method uses the visual similarity between the images for weighting the edges between the image graph but because these visual similarities are not strongly discrimitive, when they are normalized for the RWR process, they loose their significance and reduce the method's performance.

Tables 5.12 and 5.13 show the performance results of suggesting tags using the output of RWR during the tag weighting step. Performances using 3 different distance measures are presented. With generally lower precision percentages, the results of RWR method are more variant for different groups. There are some groups with relatively higher performances.

5.6 Discussion

As explained in the previous sections, the original tags that are used as ground-truth are not sufficient both in number and content. Although any other alternative is too costly, using these tags for performance evaluation gives statistical results that do not reflect the true performance.

In various examples, the similarity based suggestion method, which was proved to work better with proper ground-truth set for evaluation, has indeed suggested meaningful tags, much better than the ones suggested by the baseline method.

Figures 5.17 and 5.18 show several examples from the suggestion results of similarity based method. As it can be seen, most of the good suggestions made by the similarity based method are image specific and contain detailed descriptions of the content. Most suggestions made by the similarity method, on the other hand, contain general descriptions.

In groups underwater and twilight one can see that the suggestions that were

Table 5.12: Average precision values obtained by suggesting top 10 highest weighted tags using suggested method with RWR based weighting process. (Part 1)

Target Group	**Chi-Square**	**L1**	**L2**
abandoned:	0.103	0.103	0.103
abstract:	0.127	0.127	0.127
africa:	0.144	0.143	0.144
aircraft:	0.431	0.431	0.431
america:	0.108	0.107	0.108
amsterdam:	0.162	0.162	0.162
architecture:	0.131	0.131	0.131
arizona:	0.185	0.184	0.185
australia:	0.16	0.16	0.161
brasil_brazil:	0.221	0.221	0.218
britain_england:	0.116	0.116	0.116
canada:	0.135	0.135	0.135
chicago:	0.106	0.107	0.107
china:	0.097	0.097	0.097
colorado:	0.127	0.127	0.129
death:	0.135	0.135	0.134
florida:	0.141	0.142	0.142
france:	0.13	0.13	0.13
germany_deutschland:	0.148	0.147	0.147
greece:	0.134	0.133	0.134
hawaii:	0.167	0.167	0.167
iceland:	0.112	0.113	0.112
industrial:	0.172	0.173	0.171
ireland:	0.12	0.121	0.123
italy_italia:	0.186	0.186	0.187
japan:	0.131	0.131	0.132
london:	0.117	0.117	0.116
mexico:	0.073	0.073	0.073
michigan:	0.138	0.137	0.138
military:	0.153	0.153	0.153
ontario:	0.235	0.235	0.234
oregon:	0.149	0.149	0.15
photography:	0.119	0.119	0.119
portugal:	0.131	0.13	0.128
quebec:	0.261	0.261	0.258
scotland:	0.098	0.098	0.099

Table 5.13: Average precision values obtained by suggesting top 10 highest weighted tags using suggested method with RWR based weighting process. (Part 2)

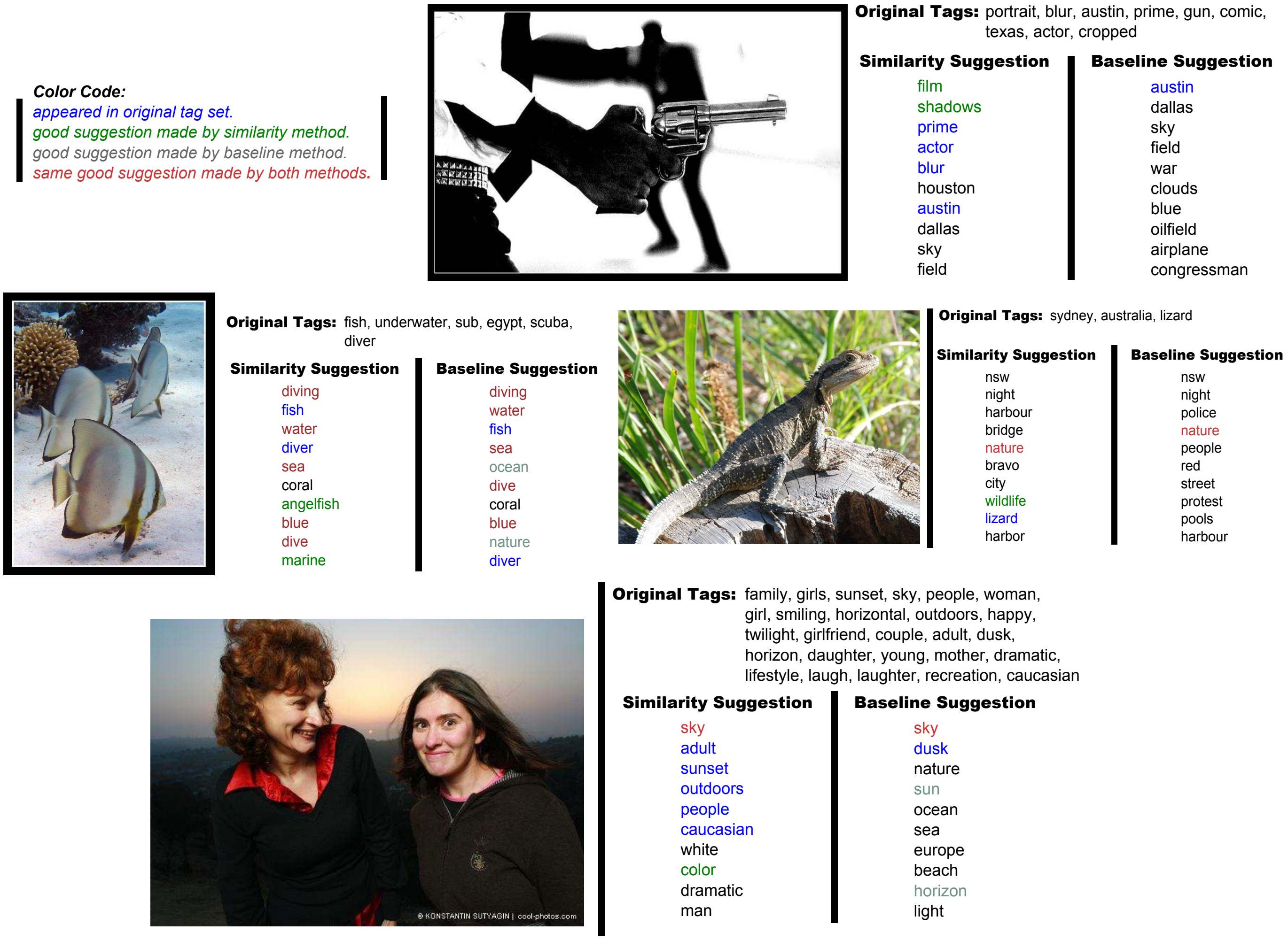
Target Group	**Chi-Square**	**L1**	**L2**	**Cluster Freq.**
seattle:	0.123	0.123	0.123	
spain:	0.109	0.11	0.11	
switzerland:	0.213	0.212	0.213	
sydney:	0.244	0.244	0.243	
texas:	0.114	0.112	0.114	
thailand:	0.194	0.194	0.193	
tokyo:	0.246	0.246	0.247	
toronto:	0.113	0.113	0.112	
town:	0.171	0.171	0.173	
turkey:	0.147	0.148	0.148	
twilight:	0.33	0.33	0.328	
underwater:	0.335	0.335	0.336	
vancouver:	0.116	0.116	0.117	
washington:	0.111	0.111	0.111	

only made by the baseline method are words that can apply to almost all the images in those target groups.

For the image on the top of Figure 5.18, only two of the ten suggestions are incorrect for the image. But it can be seen that only on of the two words, bicycles, is completely irrelevavnt. Since the image actually contains the colors of season 'autumn', this tag was suggested.

For the image at the bottom of Figure 5.18, on the other hand, similarity method suggestions could only suggest general tags. The baseline suggestion was able to suggested the word 'church'. Although this means that there are many images of churchs in that cluster, these images are not visually similar images to the target image, thus the similarity method could not catch this tag.

Figure 5.19 is a good example on how the visual 'similarity' can sometimes lead to incorrect tag suggestion. In this image, a flying bird was described as an airplane by the similarity suggestion method. As it can be seen, the suggested tags describe objects that are similar in color or shape to the content of the image.



Color Code:
appeared in original tag set.
good suggestion made by similarity method.
good suggestion made by baseline method.
same good suggestion made by both methods.

Original Tags: portrait, blur, austin, prime, gun, comic, texas, actor, cropped

Similarity Suggestion	**Baseline Suggestion**
film	austin
shadows	dallas
prime	sky
actor	field
blur	war
houston	clouds
austin	blue
dallas	oilfield
sky	airplane
field	congressman

Original Tags: fish, underwater, sub, egypt, scuba, diver

Similarity Suggestion	**Baseline Suggestion**
diving	diving
fish	water
water	fish
diver	sea
sea	ocean
coral	dive
angelfish	coral
blue	blue
dive	nature
marine	diver

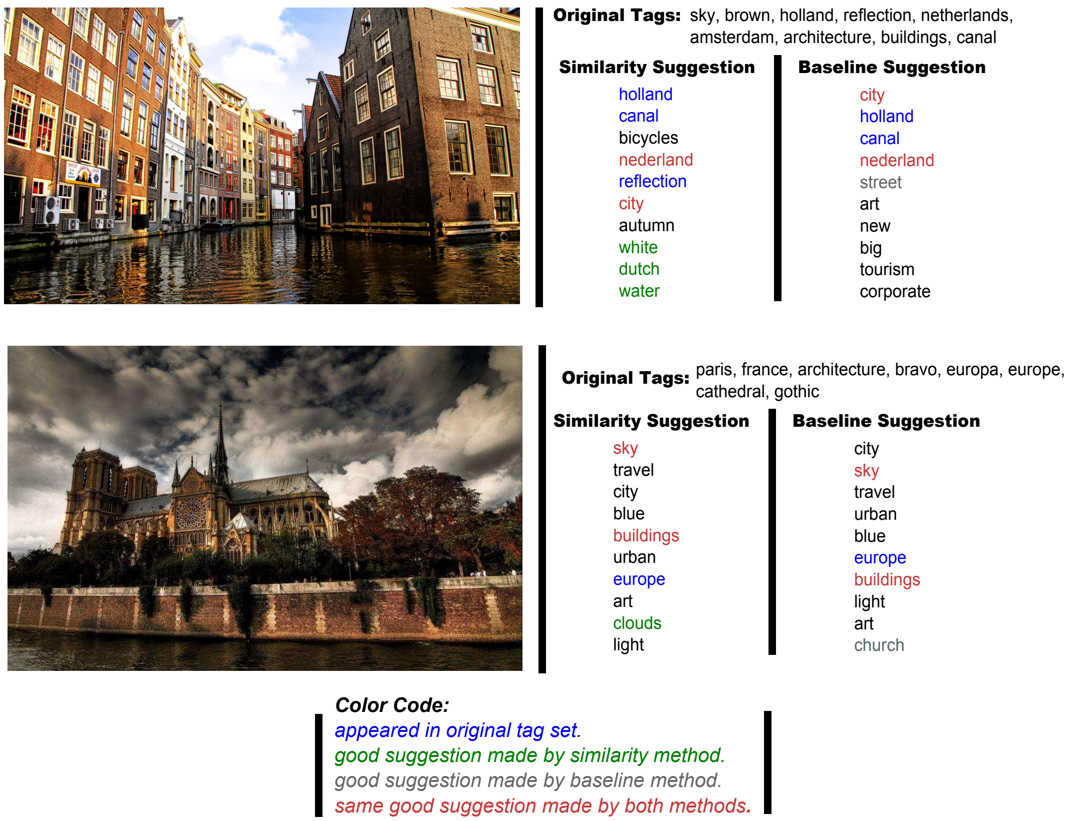
Original Tags: sydney, australia, lizard

Similarity Suggestion	**Baseline Suggestion**
nsw	nsw
night	night
harbour	police
bridge	nature
nature	people
bravo	red
city	street
wildlife	protest
lizard	pools
harbor	harbour

Original Tags: family, girls, sunset, sky, people, woman, girl, smiling, horizontal, outdoors, happy, twilight, girlfriend, couple, adult, dusk, horizon, daughter, young, mother, dramatic, lifestyle, laugh, laughter, recreation, caucasian

Similarity Suggestion	**Baseline Suggestion**
sky	sky
adult	dusk
sunset	nature
outdoors	sun
people	ocean
caucasian	sea
white	europe
color	beach
dramatic	horizon
man	light

Figure 5.17: Result comparison for sample images chosen from target groups Texas, underwater, Sydney, and twilight.



Original Tags: sky, brown, holland, reflection, netherlands, amsterdam, architecture, buildings, canal

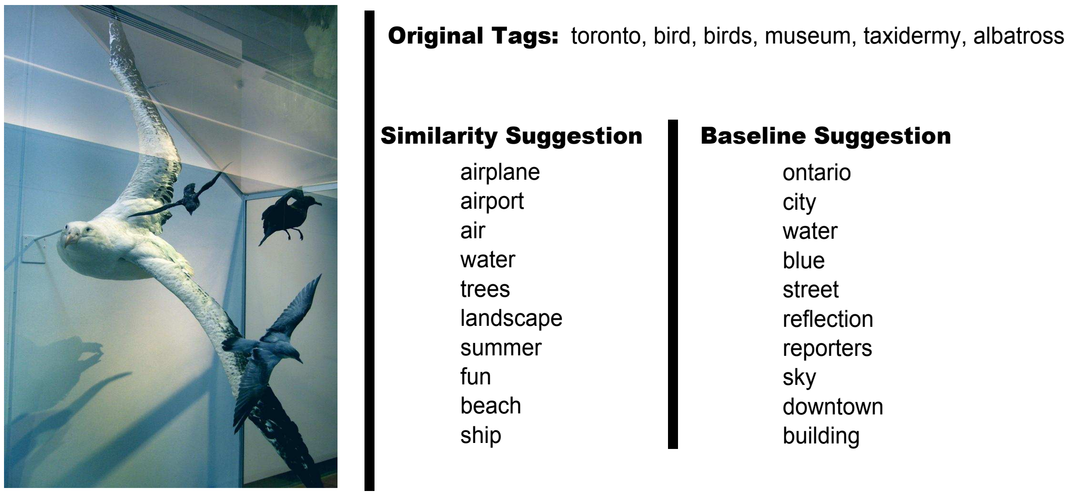
Similarity Suggestion	Baseline Suggestion
holland	city
canal	holland
bicycles	canal
nederland	nederland
reflection	street
city	art
autumn	new
white	big
dutch	tourism
water	corporate

Original Tags: paris, france, architecture, bravo, europa, europe, cathedral, gothic

Similarity Suggestion	Baseline Suggestion
sky	city
travel	sky
city	travel
blue	urban
buildings	blue
urban	europe
europe	buildings
art	light
clouds	art
light	church

Color Code:
appeared in original tag set.
good suggestion made by similarity method.
good suggestion made by baseline method.
same good suggestion made by both methods.

Figure 5.18: Result comparison for sample images chosen from target groups Amsterdam and architecture.



Original Tags: toronto, bird, birds, museum, taxidermy, albatross

Similarity Suggestion	Baseline Suggestion
airplane	ontario
airport	city
air	water
water	blue
trees	street
landscape	reflection
summer	reporters
fun	sky
beach	downtown
ship	building

Figure 5.19: Result comparison for a target image from target group Toronto.

Chapter 6

Conclusion

In this work, we have presented applications of two automatic tag suggestion methods on photo-sharing websites. The purpose of these methods were to guide and assist the tagging process of users in order to improve the management issues of large amounts of images that are available on the internet.

With the presented automatic tag suggestion methods, most problems of user generated tags can be solved. With the help of sufficient number of correctly suggested visual content describing tags, the tagging process is simplified for the users. Although the purpose of the methods is not to give users a complete set of tags for images, users will be inspired and encouraged by these suggestions for giving more and better tags.

Since these systems are suggestion systems and they do not restrict the user in the tagging process, not all issues can be solved with these applications. These applications cannot prevent users from tagging with content irrelevant words in order to get higher viewing rates.

From the experiments and discussions it can be seen that although statistical results do not reflect the truth due to the problems of the ground-truth set, the proposed methods do suggest image specific relevant tags that cannot be suggested by frequency based baseline suggestion method.

Bibliography

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:11071135, 2003.

[2] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. volume 2, page 408415. In Proceedings of the International Conference on Computer Vision, 2001.

[3] D. Blei and M. Jordan. Modeling annotated data. page 127134. In Proceedings of 26th Annual International ACM SIGIR Conference, July 2003.

[4] A. Byde, H. Wan, and S. Cayzer. Personalized tag recommendations via tagging and content-based similarity metrics. In Proceedings of the International Conference on Weblogs and Social Media, March 2007.

[5] G. Carneiro and N. Vasconcelos. Formulating semantic image annotation as a supervised learning problem. volume 2, page 163168. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[6] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. ACM International Conference on Image and Video Retrieval, July 2009.

[7] P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. volume 4, page 97112. In Proceedings of 7th European Conference on Computer Vision, May 2002.

[8] S. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. volume 2, page 10021009. In Proceedings of International Conference on Computer Vision and Pattern Recognition, 2004.

[9] J. Huang, S. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlogram. page 762768. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, June 1997.

[10] G. Institute for Integrated Circuits, Technische Universitat Munchen. Mpeg-7 xm software, June 2001.

[11] R. Jaschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme. Tag recommendations in social bookmarking systems. *AI Communications*, 21(4):231–247, December 2008.

[12] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. page 119126. In Proceedings of 26th Annual International ACM SIGIR Conference, July 2003.

[13] Y. Jing and S. Baluja. Visualrank: Applying pagerank to large-scale image search. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 30(11):1877–1890, November 2008.

[14] O. Kucuktunc, S. Sevil, A. Tosun, H. Zitouni, P. Duygulu, and F. Can. Tag suggestr: Automatic photo tag expansion using visual information for photo sharing websites. In *Lecture Notes in Computer Science*, volume 5392/2008 of *5*, pages 63–71, Koblenz, Germany, December 2008. In Proceedings of 3rd International Conference on Semantic and Digital Media Technologies (SAMT '08), Springer Verlag, Berlin Heidelberg.

[15] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. volume 16, page 553560. In Proceedings of 17th Annual Conference on Neural Information Processing Systems, 2003.

[16] J. Li and J. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):10751088, 2003.

[17] X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 2009.

[18] S. Lindstaedt, R. Mrzinger, R. Sorschag, V. Pammer, and G. Thallinger. Automatic image annotation using visual content and folksonomies. *Multimedia Tools and Applications*, 1:42, 2009.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 2004.

[20] M. Lux, O. Marques, and A. Pitman. Using visual features to improve tag suggestions in image sharing sites. In Proceedings of Knowledge Acquisition from the Social Web, 2008.

[21] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, toread. In Proceedings of the 17th Conference on Hypertext and Hypermedia, August 2006.

[22] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. page 341349. In Proceedings of the 15th International Conference on Machine Learning, 1998.

[23] J. M. Martinez. Overview of the mpeg-7 standard, 2001.

[24] G. Mishne. Autotag: A collaborative approach to automated tag assignment for weblog posts. In Proceedings of the 15th International Conference on World Wide Web (WWW '06), 2006.

[25] F. Monay and D. Gatica-Perez. Plsa-based image auto-annotation: Constraining the latent space. page 348351. In Proceedings of ACM International Conference on Multimedia, October 2004.

[26] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In Proceedings of 1st International Workshop on Multimedia Intelligent Storage and Retrieval Management, 1999.

[27] J. Pan, H. Yang, P. Duygulu, and C. Faloutsos. Automatic image captioning. volume 3, page 19871990. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo, June 2004.

[28] J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. Automatic multimedia cross-modal correlation discovery. Seattle, WA, August 2004. In Proceedings of the 10th ACM SIGKDD Conference.

[29] T. Quack, B. Leibe, and L. Gool. World-scale mining of objects and events from community photo collections. In Proceedings of ACM International Conference on Image and Video Retrieval, July 2008.

[30] Y. Rui, T. Huang, and S. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(4):3962, 1999.

[31] S. Sevil, O. Kucuktunc, P. Duygulu, and F. Can. Automatic tag expansion using visual similarity for photo sharing websites. *Multimedia Tools and Applications*, 2009.

[32] B. Sigurbjrnsson and R. V. Zwol. Flickr tag recommendation based on collective knowledge. pages 327–336. In Proceedings of the 17th International Conference on World Wide Web (WWW '08), April 2008.

[33] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[34] A. Torralba, R. Fergus, and W. Freeman. 80 million tiny images: A large dataset for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[35] C. Wang, F. Jing, L. Zhang, and H. Zhang. Scalable search-based image annotation. *Multimedia Systems*, 14(4):205–220, 2008.

[36] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In Proceedings of 19th International Conference on Pattern Recognition, 2009.

[37] X. Wang, L. Zhang, F. Jing, and W. Ma. Annosearch: Image auto-annotation by search. In Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR '06), June 2006.

[38] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field. Semiautomatic image annotation. pages 326–333. In Proceedings of the 8th IFIP TC.13 Conference on Human-Computer Interaction (INTERACT01), July 2001.

[39] L. Wenyin, Y. Sun, and H. Zhang. Mialbum - a system for home photo managemet using the semi-automatic image annotation approach. pages 479–480. In Proceedings of the 8th ACM International Conference on Multimedia (MULTIMEDIA00), 2000.

[40] Z. Xu, Y. Fu, J. Mao, and D. Su. Towards the semantic web: Collaborative tag suggestions. In Proceedings of Third International Conference on Internet and Web Applications and Services, 2008.