

Lehrstuhl für Verteilte Informationssysteme
Fakultät für Informatik und Mathematik
Universität Passau



Doctoral Thesis

Search-based Automatic Image Annotation Using Geotagged Community Photos

Conducted as cotutelle-de-thèse in cooperation with
Laboratoire LIRIS
INSA de Lyon
Lyon, France

M.Sc. Hatem Mousselly Sergieh

October 2014

Abstract

In the Web 2.0 era, platforms for sharing and collaboratively annotating images with keywords, called tags, became very popular. Tags are a powerful means for organizing and retrieving photos. However, manual tagging is time consuming. Recently, the sheer amount of user-tagged photos available on the Web encouraged researchers to explore new techniques for automatic image annotation. The idea is to annotate an unlabeled image by propagating the labels of community photos that are visually similar to it. Most recently, an ever increasing amount of community photos is also associated with location information, i.e., geotagged. In this thesis, we aim at exploiting the location context and propose an approach for automatically annotating geotagged photos. Our objective is to address the main limitations of state-of-the-art approaches in terms of the quality of the produced tags and the speed of the complete annotation process. To achieve these goals, we, first, deal with the problem of collecting images with the associated metadata from online repositories. Accordingly, we introduce a strategy for data crawling that takes advantage of location information and the social relationships among the contributors of the photos. To improve the quality of the collected user-tags, we present a method for resolving their ambiguity based on tag relatedness information. In this respect, we propose an approach for representing tags as probability distributions based on the algorithm of Laplacian score feature selection. Furthermore, we propose a new metric for calculating the distance between tag probability distributions by extending Jensen-Shannon Divergence to account for statistical fluctuations. To efficiently identify the visual neighbors, the thesis introduces two extensions to the state-of-the-art image matching algorithm, known as Speeded Up Robust Features (SURF). To speed up the matching, we present a solution for reducing the number of compared SURF descriptors based on classification techniques, while the accuracy of SURF is improved through an efficient method for iterative image matching. Furthermore, we propose a statistical model for ranking the mined annotations according to their relevance to the target image. This is achieved by combining multi-modal information in a statistical framework based on Bayes' rule. Finally, the effectiveness of each of mentioned contributions as well as the complete automatic annotation process are evaluated experimentally.

Keywords: Image Annotation, SURF, Tagging, Geotagging, Flickr, Folksonomy.

Résumé

La technologie Web 2.0 a donné lieu à un large éventail de plates-formes de partage de photos. Il est désormais possible d'annoter des images de manière collaborative, au moyen de mots-clés; ce qui permet une gestion et une recherche efficace de ces images. Toutefois, l'annotation manuelle est laborieuse et chronophage. Au cours des dernières années, le nombre grandissant de photos annotées accessibles sur le Web a permis d'expérimenter de nouvelles méthodes d'annotation automatique d'images. L'idée est d'identifier, dans le cas d'une photo non annotée, un ensemble d'images visuellement similaires et, a fortiori, leurs mots-clés, fournis par la communauté. Il existe actuellement un nombre considérable de photos associées à des informations de localisation, c'est-à-dire géo-localisées. Nous exploiterons, dans le cadre de cette thèse, ces informations et proposerons une nouvelle approche pour l'annotation automatique d'images géo-localisées. Notre objectif est de répondre aux principales limites des approches de l'état de l'art, particulièrement concernant la qualité des annotations produites ainsi que la rapidité du processus d'annotation. Tout d'abord, nous présenterons une méthode de collecte de données annotées à partir du Web, en se basant sur la localisation des photos et les liens sociaux entre leurs auteurs. Par la suite, nous proposerons une nouvelle approche afin de résoudre l'ambiguïté propre aux tags d'utilisateurs, le tout afin d'assurer la qualité des annotations. L'approche démontre l'efficacité de l'algorithme de recherche de caractéristiques discriminantes, dit de Laplace, dans le but d'améliorer la représentation de l'annotation. En outre, une nouvelle mesure de distance entre mots-clés sera présentée, qui étend la divergence de Jensen-Shannon en tenant compte des fluctuations statistiques. Dans le but d'identifier efficacement les images visuellement proches, la thèse étend sur deux points l'algorithme d'état de l'art en comparaison d'images, appelé SURF (Speeded-Up Robust Features). Premièrement, nous présenterons une solution pour filtrer les points-clés SURF les plus significatifs, au moyen de techniques de classification, ce qui accélère l'exécution de l'algorithme. Deuxièmement, la précision du SURF sera améliorée, grâce à une comparaison itérative des images. Nous proposerons un modèle statistique pour classer les annotations récupérées selon leur pertinence du point de vue de l'image-cible. Ce modèle combine différents critères, il est centré sur la règle de Bayes. Enfin, l'efficacité de l'approche d'annotation ainsi que celle des contributions individuelles sera démontrée expérimentalement.

Mots-clés: Annotation d'images, SURF, Tagging, Code Géographique, Flickr, Folksonomie.

Zusammenfassung

Seit der Einführung von Web 2.0 steigt die Popularität von Plattformen, auf denen Bilder geteilt und durch die Gemeinschaft mit Schlagwörtern, sogenannten Tags, annotiert werden. Mit Tags lassen sich Fotos leichter organisieren und auffinden. Manuelles Taggen ist allerdings sehr zeitintensiv. Animierte von der schier unermesslichen Menge an im Web zugänglichen, von Usern getaggen Fotos, erforschen Wissenschaftler derzeit neue Techniken der automatischen Bildannotation. Dahinter steht die Idee, ein noch nicht beschriftetes Bild auf der Grundlage visuell ähnlicher, bereits beschrifteter Community-Fotos zu annotieren. Unlängst wurde eine immer größere Menge an Community-Fotos mit geographischen Koordinaten versehen (*geotagged*). Die Arbeit macht sich diesen geographischen Kontext zunutze und präsentiert einen Ansatz zur automatischen Annotation geogetaggtter Fotos. Ziel ist es, die wesentlichen Grenzen der bisher bekannten Ansätze in Hinsicht auf die Qualität der produzierten Tags und die Geschwindigkeit des gesamten Annotationsprozesses aufzuzeigen. Um dieses Ziel zu erreichen, wurden zunächst Bilder mit entsprechenden Metadaten aus den Online-Quellen gesammelt. Darauf basierend, wird eine Strategie zur Datensammlung eingeführt, die sich sowohl der geographischen Informationen als auch der sozialen Verbindungen zwischen denjenigen, die die Fotos zur Verfügung stellen, bedient. Um die Qualität der gesammelten User-Tags zu verbessern, wird eine Methode zur Auflösung ihrer Ambiguität vorgestellt, die auf der Information der Tag-Ähnlichkeiten basiert. In diesem Zusammenhang wird ein Ansatz zur Darstellung von Tags als Wahrscheinlichkeitsverteilungen vorgeschlagen, der auf dem Algorithmus der sogenannten Laplacian Score (LS) aufbaut. Des Weiteren wird eine Erweiterung der Jensen-Shannon-Divergence (JSD) vorgestellt, die statistische Fluktuationen berücksichtigt. Zur effizienten Identifikation der visuellen Nachbarn werden in der Arbeit zwei Erweiterungen des Speeded Up Robust Features (SURF)-Algorithmus vorgestellt. Zur Beschleunigung des Abgleichs wird eine Lösung auf der Basis von Klassifikationstechniken präsentiert, die die Anzahl der miteinander verglichenen SURF-Deskriptoren minimiert, während die SURF-Genauigkeit durch eine effiziente Methode des schrittweisen Bildabgleichs verbessert wird. Des Weiteren wird ein statistisches Modell basierend auf der Baye'schen Regel vorgeschlagen, um die erlangten Annotationen entsprechend ihrer Relevanz in Bezug auf das Zielbild zu ranken. Schließlich wird die Effizienz jedes einzelnen, erwähnten Beitrags experimentell evaluiert. Darüber hinaus wird die Performanz des vorgeschlagenen automatischen Annotationsansatzes durch umfassende experimentelle Studien als Ganzes demonstriert.

Schlagwörter: Bildannotation, SURF, Tagging, Geokodierung, Flickr, Folksonomie.

Contents

Abstract	iii
Résumé	v
Zusammenfassung	vii
List of Figures	xiii
List of Tables	xvii
I Motivation and State-of-the-Art	1
1 Introduction	3
1.1 The Context	3
1.2 Automatic Image Annotation	6
1.3 The Challenges	7
1.4 The Solution	7
1.5 Structure of the Thesis	11
2 Related Work	13
2.1 Introduction	13
2.2 Content-based Image Retrieval	14
2.3 Model-based Automatic Image Annotation	16
2.4 Search-based Automatic Image Annotation	17
2.4.1 The World Wide Web as Annotation Resource	19
2.4.2 Community Photos as Annotation Resource	22
2.4.2.1 Classification-based Approaches	22
2.4.2.2 Semi-automatic Approaches	23
2.4.2.3 AIA using Spatial and Temporal Contexts	24
2.4.2.4 AIA using Social Context	27
2.5 Related Work around Search-based Image Annotation	29
2.5.1 Geo-based Photo Crawling	29
2.5.2 Resolving Tag Ambiguity	30

2.5.3	Improving Image Matching	34
2.6	Thesis Contributions Revisited	35
II Data Preparation and Tag Disambiguation		37
3	Geographical Crawling and Indexing of Community Photos	39
3.1	Introduction	39
3.2	Geo-based Data Crawling	40
3.3	Tag Cleaning	42
3.4	Indexing using Quad-tree	45
3.5	Qualitative Insight	47
3.6	Summary	49
4	Mining Tag Relatedness for Resolving Tag Ambiguity	51
4.1	Introduction	51
4.2	Tag Relatedness Approach	53
4.2.1	Approach Overview	53
4.2.2	Feature Selection for Tag Relatedness	54
4.2.3	Tag Probability Distribution	60
4.2.4	Distance Measure	61
4.3	Evaluation	65
4.3.1	Dataset	65
4.3.2	Qualitative Insight	65
4.3.3	Semantic Grounding using WordNet	67
4.3.4	Evaluation using Large Scale Co-occurrence Statistics	69
4.4	Resolving Tag Ambiguity	73
4.5	Summary	76
III Improving SURF-based Image Matching		77
5	Classification-based Keypoint Pruning	79
5.1	Introduction	79
5.2	Keypoint Pruning using Visual Attention Models	81
5.3	Keypoint Pruning as Classification Problem	82
5.3.1	Training Dataset	83
5.3.2	Classification Features	85
5.3.3	Classification Using Random Forest	87
5.4	Experimental Evaluation	88
5.4.1	Classification Performance	88
5.4.2	Effectiveness of Keypoint Pruning	89
5.4.3	Runtime Evaluation	92
5.5	Summary	96
6	SURF-based Iterative Image Matching	99
6.1	Introduction	99
6.2	Iterative Image Matching	100

6.3	Experimental Evaluation	102
6.3.1	Dataset	102
6.3.2	Evaluation Methodology	103
6.3.3	Results	104
6.4	Summary	105
IV	Tag Ranking and Global Evaluation	107
7	Probabilistic Model for Tag Ranking	109
7.1	Introduction	109
7.2	Problem Statement	109
7.3	Pseudo-generative Statistical Model	111
7.3.1	Estimating Image-to-Image Relevance: $P(I_j I_{in})$	113
7.3.2	Estimating Word-to-Image Relevance: $P(t_i I_j)$	114
7.4	Summary	116
8	Experimental Evaluation	117
8.1	Setup and Evaluation Procedure	117
8.2	Evaluation Metrics	118
8.3	Compared AIA Models	118
8.4	Results Discussion	119
8.4.1	Thesis Approach (TA) vs. Baseline (BL) Methods	120
8.4.2	Effectiveness of Tag Refinement	122
8.5	Summary	124
9	Conclusion and Future Work	125
9.1	Summary of Research	125
9.1.1	Two-Phase Automatic Image Annotation Approach	125
9.1.2	Geo-based Data Crawling and Indexing	126
9.1.3	Resolving Tag Ambiguity	126
9.1.4	Improving SURF-based Image Matching	127
9.1.5	Tag Ranking	128
9.2	Future Work	128
	Bibliography	131

List of Figures

1.1	Number of photos uploaded monthly to Flickr in the period from 1/1/2005 until 1/9/2013. We obtained the numbers using Flickr API	4
1.2	Sample photos from Flickr with the associated user-provided tags	5
1.3	Assigning location information to a photo taken in Damascus using Flickr’s interactive map	5
1.4	Number of <i>geotagged</i> photos uploaded monthly to Flickr in the period from 1/1/2005 until 1/9/2013. We obtained the numbers using Flickr API	6
1.5	The workflow of the proposed image annotation approach	8
2.1	The goal of automatic image annotation	13
2.2	Taxonomy of the search-based AIA approaches reviewed in this chapter	28
3.1	The geographical coordinates (latitude vs. longitude) of a sample of 300,000 images from our dataset	41
3.2	Photo density in the city of Paris according to our dataset	41
3.3	The number of images per city according to our dataset	42
3.4	Search results for the term ”newyork” according to Yahoo search engine with suggestions for related search terms	43
3.5	World map divided into tiles according to the photo density as given by our dataset. Dense tiles are further divided into sub-tiles	46
3.6	Quad-tree regions for our dataset. The quad-tree algorithm is applied on each tile separately to allow efficient computation	46
3.7	The number of photos per tile according to our dataset. The x-axis correspond to the tile identifier and the y-axis gives the total number of photos per tile	47
3.8	Geographical clusters at zooming level 5	48
3.9	Geographical clusters by zooming in to level 6	48
3.10	Sample images from our dataset which are located in Paris	49
3.11	Tags form our dataset corresponding to images taken in Paris	49
4.1	Tag proposals for a photo of London Eye	52
4.2	The workflow of the proposed tag relatedness approach	54
4.3	A sample tag co-occurrence matrix	58
4.4	Similarity graph for the data points corresponding to the rows of the matrix shown in Figure 4.3. The nodes corresponds to the tags with the edges weighted according to the cosine similarity	58
4.5	The similarity matrix S , the diagonal matrix D and the Laplacian matrix L as generated from the nearest neighbor graph of Figure 4.4	59

4.6	Empirical probability distributions of two tags "river" and "thames". Each distribution consists of several histogram channels corresponding to the elements of a feature set (x-axis). The value of a histogram channel is given by the normalized tag co-occurrence	61
4.7	Two histogram channels corresponding to the feature $f \in \mathcal{F}$ taken from the empirical probability distributions P and Q receptively. Each histogram channel is considered as a normally distributed random variable	63
4.8	Tag importance according to LS algorithm vs. tag frequency	67
4.9	Comparison between FRQ an LS feature selection methods	70
4.10	Comparison between tag relatedness distance measures	71
4.11	Histograms of the number of tag pairs identified by each of the distance measures AJSD, JSD and COS using the top 1,000 LS Features. Each channel of the histogram corresponds to the number of tag pairs in a predefined range of DISCO similarity	72
4.12	Workflow of the proposed tag proposals disambiguation procedure	73
5.1	Two images depicting the same scene from different perspectives. Key-point correspondences between the two images are connected using dotted lines	80
5.2	The model of saliency-based visual attention according to [Itti et al., 1998]	81
5.3	Keypoint pruning using saliency maps. a) The input image, b) the corresponding saliency map, c) the image with the identified SURF keypoints (without keypoint pruning), d) the image with the subset of SURF keypoints corresponding to salient regions	82
5.4	A sample of image groups, which were used to create the training dataset. The image groups were taken from the Object Recognition dataset [Nister and Stewenius, 2006]	84
5.5	Recall of different keypoint pruning approaches as a function of the applied decision/saliency thresholds. The test dataset includes 200 groups from the Object Recognition dataset [Nister and Stewenius, 2006]	91
5.6	The drop in the matching recall (relative to the recall achieved by a full matching) as a function of the keypoint reduction ratio achieved by different keypoint pruning methods. The results are obtained from 200 groups of the Object Recognition dataset	92
5.7	A sample of the image groups of our manually created dataset. The images have a high resolution (3264×1840 pixels) and an average volume of 1 MB	94
5.8	The drop in the matching recall (relative to the recall achieved by a full matching) as a function of the keypoint reduction ratio achieved by different keypoint pruning methods. The results are obtained using a manually created test dataset (Figure 5.7)	95
5.9	Comparison between the theoretical and the actual (practical) runtime ratio under different keypoint pruning configurations. The advantage of image matching using keypoint pruning is measured by how much it is faster than the full matching. The x-axis corresponds to the analyzed keypoint configurations. The y-axis represents the ratio between the runtime (the theoretical as well as the practical one) required by keypoint pruning-based image matching and that which is required by the full matching	96

6.1	Example: image matching using SURF	100
6.2	(a) A tree representation for a collection of images which are similar to a given input image (the root). The images are obtained through an iterative application of the image matching algorithm on each image at each level of the tree. (b) To reduce the computation cost, images at each level in the tree are first clustered and a representative image of each cluster is used to identify further possible matching images	101
6.3	A sample from the image dataset which we used to evaluate the iterative image matching approach. Each row corresponds to a subset of visually similar images	103
6.4	The average recall achieved by the naive as well as the clustering-based approaches at different matching iterations	104
6.5	Clustering-based vs. naive iterative matching: Average matching runtime required by each approach at different matching iterations	105
7.1	A simplified model showing the input for the tag ranking phase of our image annotation approach	110
7.2	(a) A Graph representation of a folksonomy corresponding to annotated images found in the geographical proximity of an input image I_{in} . (b) The graph extended by an additional vertex corresponding to the input image. The dashed edges connecting the input image to the geographically close images correspond to the visual similarity	111
7.3	A Bayesian model for tag ranking: I_{in} represents the input image, $I_j \in \mathcal{I}_{geo}$ are the geographical neighbors and $t_i \in \mathcal{T}_{geo}$ is the set of the associated tags	112
8.1	The performance of our AIA approach against the baseline models	121
8.2	Sample of the annotated test images with the ground tags and the top tags which were predicted according to our approach TA (TF-IDF) and the baseline method BL (geo-voting)	122
8.3	The performance of the proposed AIA with tag refinement	123
8.4	Sample of test images which were annotated using the approach: TA (TF-IDF) without tag refinement, compared to the same approach under tag refinement: TA+TR(TF-IDF). Tag proposals in bold correspond to refined tags and those in italic results from extending the annotation list to reach $k = 10$	124

List of Tables

3.1	Sample user-tags acquired from Flickr (first column) which have been automatically corrected using the presented tag cleaning algorithm (second column)	44
4.1	The feature vectors ordered according to their importance (Laplacian score) from most to least important	60
4.2	Sample tags with the corresponding most related tags	66
4.3	Configurations of the evaluated tag relatedness approaches	68
4.4	Similar tag pairs which are identified by the proposed tag relatedness approach. The listed tags do not have corresponding entries in WordNet. Tag pairs shown in italic are identified as related by DISCO. The description of each tag pair were obtained from Wikipedia	71
4.5	Sample of tag pairs extracted from the folksonomy presented in Section 4.3.1 with the corresponding semantic relationships and their disambiguation terms	75
5.1	Average runtime of the pre-matching (offline) phases	94

Part I

Motivation and State-of-the-Art

Chapter 1

Introduction

1.1 The Context

The Internet is a ubiquitous medium that facilitates communication among people. The way we use the Internet is evolving. With the emergence of Web 2.0 [Graham, 2005, O'Reilly, 2005], we entered an era in which every internet user became a prosumer, i.e., a consumer as well as a producer of Internet content. Another characteristic of today's Internet is the fact that it turned into a platform of services supporting different kinds of social interactions. With over 1 billion users, Facebook [Facebook, 2014] is one of the most successful examples of the social Web.

Photos represent one of the most common content types which are contributed and shared among the users of the Internet. This can be explained according to the availability of digital photography devices which provide an easy and a cheap medium for producing photos. At the same time, the bandwidth of the current Internet connections allows fast upload of photos. There are also several social aspects that make photos that popular. Photos are not only a documentary or reminders; they are also an emotional journal. Moreover, photos are a rich type of content that *"is worth a thousand words"* [Brisbane, 1911], they capture our moods and feelings and provide a proof that we have been there. Additionally, photos represent a subtle means of social communication. People post their photos as a statement of positive affirmation regarding the way they live, what they do and what they achieved.

Social networks and specialized photo sharing websites like Flickr [Flickr, 2014a] are witnessing immense amounts of contributed images. To get an impression of the scale of digital photographs shared on the Web, we gathered statistics about photos hosted by Flickr over a period of eight years using the provided API [Flickr, 2014b]. Figure 1.1

shows that the number of photos is increasing from month to month and from year to year¹. In 2013, Flickr announced that they reached 8 billion photos [Robertson, 2013].

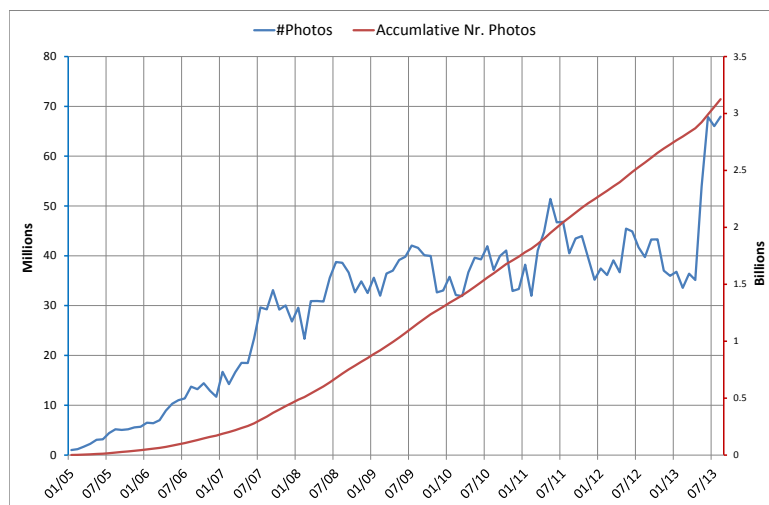


FIGURE 1.1: Number of photos uploaded monthly to Flickr in the period from 1/1/2005 until 1/9/2013. We obtained the numbers using Flickr API

With the explosively growing amounts of online photos, there is an indispensable need for efficient solutions for photo management. Currently, image sharing websites allow users to annotate their photos with keywords called tags (Figure 1.2). Users can also collaborate with each other in an activity called *social tagging* in order to enhance the quality of the provided tags. Tagging has become a pervasive component of the Web and the aggregation of the tags has even got its own name: **folksonomy** [Vanderwal, 2010]. Tags help to bridge the gap between the digital representation and the semantic of the photos. Consequently, better management and retrieval can be achieved. Thanks to tagging, photos can be retrieved using keywords in the same manner as text documents.

Recently, community photos have seen an additional improvement through the technique of geotagging. Geotagging is the process of adding location information (i.e. the longitude and the latitude) to photo metadata. Geotags can be automatically inserted into the EXIF descriptor [Technical Standardization Committee on AV & IT Storage Systems and Equipment, 2002] of the image through built-in GPS receivers of modern cameras or smart phones. It is also possible to assign location information manually using an interactive map as provided by Flickr (Figure 1.3). Geotags provide another dimension for retrieving and organizing photos based on their location of capture. Currently, there are considerable amounts of geotagged photos shared on the Web. For

¹Note that Flickr API only retrieves the number of photos which declared as public by their owners. Therefore, the presented statistics may not conform to the numbers published by the authorities of Flickr which also consider private photos.

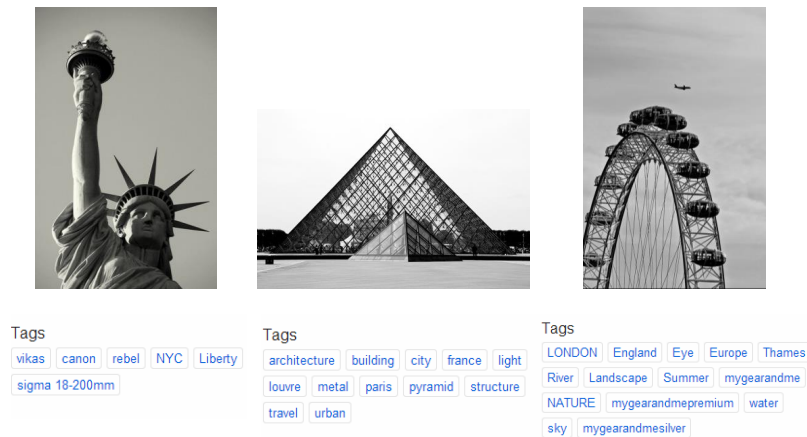


FIGURE 1.2: Sample photos from Flickr with the associated user-provided tags

instance, the rate of geotagged photos uploaded to Flickr shows an increasing trend over the years (Figure 1.4). At the time of writing this thesis, we counted around 214 million of publicly accessible geotagged photos hosted by Flickr.

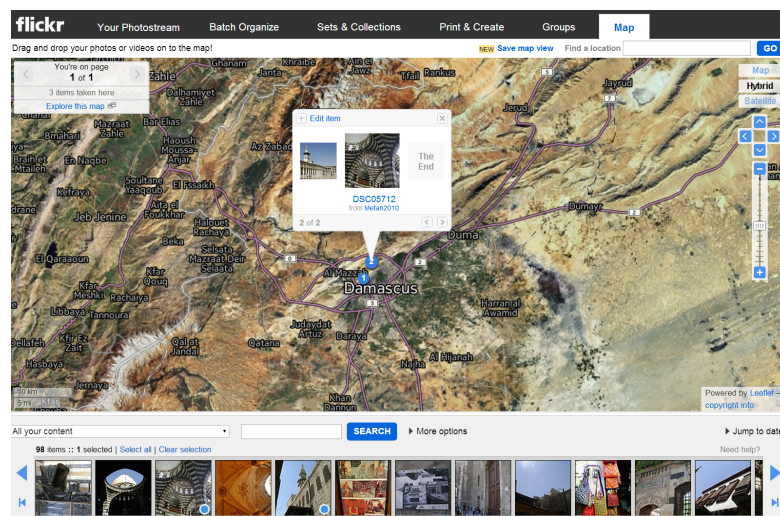


FIGURE 1.3: Assigning location information to a photo taken in Damascus using Flickr’s interactive map

In spite of the mentioned advantages of tagging and geotagging, managing the huge amount of personal as well as community photos is still far from satisfactory. On one hand, manual tagging remains a laborious task, thus, it is usually ignored. Furthermore, user-tags are noisy (i.e. incorrect and incomplete) since they are created in a free-style and uncontrolled manner. On the other hand, it is true that geotagging provides an effortless and simultaneously efficient way for photo organization and retrieval. However,

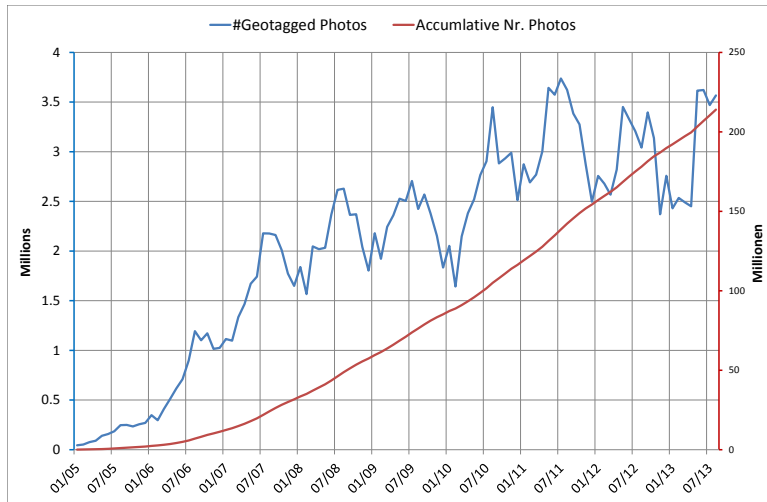


FIGURE 1.4: Number of *geotagged* photos uploaded monthly to Flickr in the period from 1/1/2005 until 1/9/2013. We obtained the numbers using Flickr API

is limited to one organizational aspect only, i.e., the location. Hence, geotagging alone cannot meet the diverse user needs for organizing and retrieving photos.

1.2 Automatic Image Annotation

To address the limitations of manual tagging, research on automatic image annotation has received a considerable attention. Automatic image annotation aims at associating unlabeled images with keywords that describe their contents. Early research on automatic annotation techniques focused on using machine learning techniques. The idea is to use a dataset of already labeled images in order to train models for predicting labels for un-annotated images. However, creating good training datasets is a challenging and time consuming task. Indeed, most available datasets are limited to images corresponding to small set of predefined concepts. Therefore, the annotations generated by such approaches are also limited and they cannot meet the diverse ways in which people describe and search for images.

Most recently, the sheer amounts of user-tagged photos available on the Web encouraged researchers to explore techniques for leveraging this important resource in the automatic annotation process. The idea is to annotate an unlabeled image by propagating the labels of community photos that are visually similar to the input image. For this purpose, content-based image retrieval (CBIR) techniques are applied to identify the visual neighbors of the input image. Consequently, the labels of the visual neighbors are analyzed

and transferred to the un-annotated image. Due to the implied retrieval step, such approaches are described as search-based. Compared to traditional approaches, there is no limitation on the annotations produced by search-based automatic image annotation approaches. That is because the produced annotations are derived from user-supplied tags which are diverse in nature.

1.3 The Challenges

For all its promising edge, search-based image annotation has to deal with several challenges. The first challenge is posed by community tags as a main resource from which annotations (for unlabeled images) are extracted. User-tags are created in an uncontrolled and free-style manner, thus, they are inherently noisy. Humans use inconsistent terms to describe the same thing or use the same term to express different meanings [Furnas et al., 1987]. In other words, polysemy and homonymy – two fundamental problems in information retrieval – are also present in user-provided tags.

Second, as mentioned before, identifying images similar to the un-annotated image is a core component of the automatic annotation process. Accordingly, automatic image annotation has also to deal with two main challenges of CBIR techniques, namely the accuracy and the speed of the applied technique. Generally, the accuracy of CBIR is ruled by the low level image representation that is used, i.e., image features. In turn, the complexity of extracting image features, representing them as descriptor vectors and comparing the descriptors are major factors that influence the retrieval speed. Therefore, in order to ensure the efficiency of automatic image annotation, solutions for improving the accuracy and boosting the performance of the applied CBIR process have to be investigated.

Third, automatic image annotation has to address the issue of estimating the relevance/importance between candidate annotations and the target image. Therefore, there is a need for robust models that are able to combine different relevance clues to rank candidate annotations.

1.4 The Solution

Due to the current technological developments, we expect that geotagged photos will dominate the Web in the near future. Geotagging is developing into a fully automatic task. On one hand, there is a rapid popularization of GPS-enabled digital cameras which can generate location information automatically. On the other hand, research for

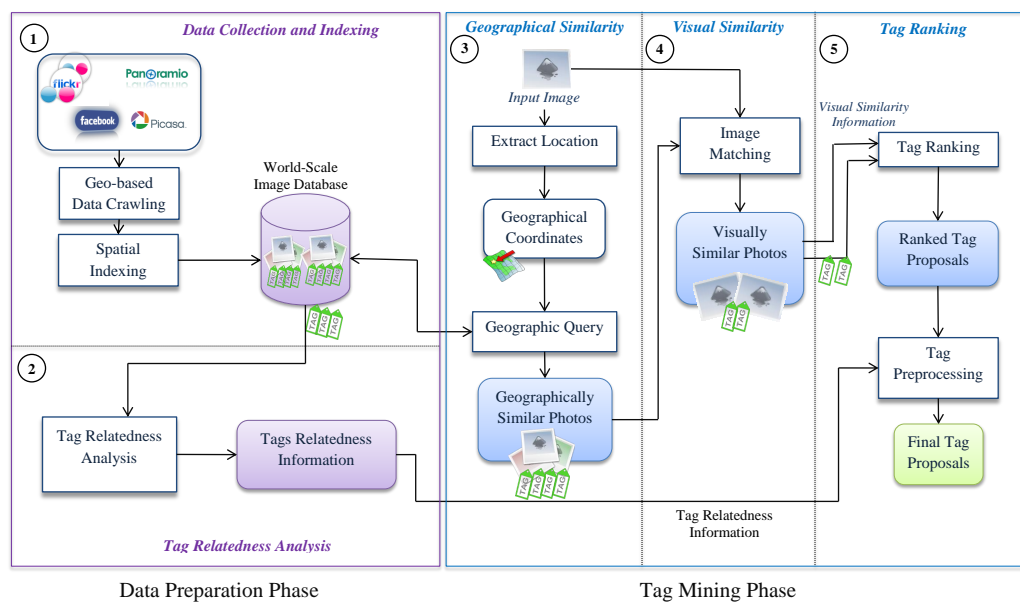


FIGURE 1.5: The workflow of the proposed image annotation approach

automatically identifying the location of non-geotagged images is witnessing more and more success.

Based on this assumption, this thesis proposes an approach for automatically annotating geotagged images. In contrast to current solutions, our approach leverages the ever increasing location information as a valuable resource in the process of mining candidate annotations for unlabeled images. In addition to addressing the mentioned challenges, such as efficient identification of similar images and tag ranking, the thesis also deals with the problem of collecting image data from online image sharing websites and indexing them to enable efficient retrieval. Furthermore, the thesis aims at addressing the problem of tag noisiness in order to improve the quality of the produced tag proposals.

The workflow of our automatic annotation approach is illustrated in Figure 1.5. The annotation process is divided into two main phases: a data preparation phase and a tag mining phase. In the data preparation phase, a location-based crawling strategy is applied to collect image data from community photo websites. The collected images are then indexed spatially based on the associated location information. Subsequently, the user-provided tags are analyzed in order to extract relatedness information which will be used later on to improve the quality of tag proposals. The tag mining phase receives an unlabeled image, which is geotagged, as input, and produces tag proposals as output.

In the following, the individual processing steps of the proposed approach and the corresponding thesis contributions are described.

① Data Collection and Indexing

Annotations for a target image are mined from tags of community photos taken in the same location. For this purpose, online sharing websites, such as Flickr or Panoramio [Panoramio, 2014] have to be queried. This is facilitated through specialized APIs which allow access to different kinds of metadata about the stored images. However, the time required to obtain the data is high. In general, search-based automatic annotation approaches demand an intensive data traffic between the online service and the annotation system which leads to a serious performance bottleneck. To address this problem, the initial phase of the proposed annotation approach provides efficient methods for collecting and indexing huge amounts of geotagged community photos. In order to annotate a wide-range of images taken in different places around the world, we created and spatially indexed a world-scale dataset of geotagged images with the associated metadata.

Thesis Contribution: The thesis proposes a strategy for crawling data from Flickr. Our method benefits from the idea of small-world phenomenon [Milgram, 1967] and exploits Flickr friendship’s graph to generate a representative dataset on a world-wide scale. More specifically, the collected data cover the whole world and the density of the photos for a given place reflect its popularity among photographers. To allow efficient retrieval, we introduced a method to index the data spatially based on the quad-tree data structure [Finkel and Bentley, 1974]. Using the proposed method, we were able to create and index a dataset of more than 14 million images.

② Tag Relatedness Analysis

The second step of the data preparation phase deals with the problem of ambiguous and redundant user-tags. In general, resolving tag ambiguity is done by creating a context for each tag based on its relatedness to other tags in a given folksonomy. Subsequently, the context is used to assist the process of identifying the correct meaning of the tag.

Thesis Contribution: The thesis presents a novel tag relatedness approach for resolving tag ambiguity. Our solution uses statistical means to model tags as probability distributions based on their co-occurrence patterns in folksonomies. We deal with two main issues that influence the quality of tag relatedness measures. First, we analyze the effect of tag representation on the quality of the tag relatedness metric. Accordingly, we propose a feature selection approach based on the technique of Laplacian score [He et al., 2005] and use the identified features to construct tag probability distributions. Second, since the relatedness between two tags is determined according to the distance between the corresponding probability distributions, the applied distance metric plays

a crucial role in this regard. Accordingly, we propose a distance metric based on the idea of Jensen-Shannon Divergence [Manning and Schütze, 1999]. The main advantage of the new metric, called Adapted Jensen-Shannon Divergence (AJSD), is its ability to deal with statistical fluctuations which are inherent in probability distributions generated from samples. Finally, we propose a simple technique for exploiting the extracted relatedness information in order to improve the quality of the mined tag proposals.

③ Geographical Similarity

The first step of the tag mining process implies searching for images taken in the same location as the input image. To achieve this, the geographical coordinates of the input image are extracted and used to query the image dataset created in the first phase.

Thesis Contribution: here, we provide an efficient processing of geographic queries based on the spatial index created in the first phase.

④ Visual Similarity

In this phase, the set of geographically similar images is investigated to determine images depicting the same or similar scenes as the input image. To achieve this goal, a CBIR technique for image matching is applied. The tags of the visual neighbors represent the set of candidate annotations for the target image.

Thesis Contribution: Although image matching techniques have witnessed great improvement in the last few years, they are still computationally expensive. In this thesis, we perform image matching based on an improved version of a state-of-the-art algorithm called SURF [Bay et al., 2008]. In this respect, we propose a method based on classification techniques aiming to speed up the matching by reducing the number of compared SURF descriptors. Furthermore, we introduce an efficient algorithm for iterative image matching in order to improve the matching accuracy.

⑤ Tag Ranking

In this final phase, the knowledge gained from the previous phases is fed into a tag ranking algorithm. Tag proposals are then further refined using the information provided by the phase of tag relatedness analysis.

Thesis Contribution: We propose a statistical model for tag ranking based on Bayes' rule. The model exploits three main information resources to determine the importance

of candidate tags. These include information about tag usage pattern, the similarity between the target image and the visual neighbors as well as user consensus regarding the importance of the tags. We provide different alternatives to quantify this information and analyze their effect on the quality of the final annotations.

The proposed automatic annotation approach was evaluated experimentally. We conducted several studies in order to assess the efficiency of each of the presented contributions on an individual basis as well as within the frame of the global automatic image annotation approach. The quality of the produced annotations was evaluated based on a ground truth obtained from the Flickr Getty Image Collection [Flickr, 2014c] which contains images annotated by experts.

1.5 Structure of the Thesis

The rest of the thesis is structured as follows:

- **Chapter 2** introduces a survey of research effort on automatic annotation and its associated challenges.
- **Chapter 3** describes an approach for crawling and indexing geotagged images from community photos websites. This chapter is based on the works published in [Mousselly Sergieh et al., 2014a,b].
- **Chapter 4** presents an approach for mining tag relatedness information for resolving tag ambiguity. This chapter is based on the works published in [Mousselly Sergieh et al., 2013, 2014c].
- **Chapter 5** presents an approach for accelerating the process of image matching based on SURF features. This chapter is based on the work published in [Mousselly Sergieh et al., 2012a].
- **Chapter 6** investigates improving the accuracy of SURF-based image matching through an iterative approach. This chapter is based on the work published in [Mousselly Sergieh et al., 2012b].
- **Chapter 7** describes an approach for tag ranking by combining different contextual clues in order to determine tag-to-image relevance. This chapter is based on the work published in [Mousselly Sergieh et al., 2012b].
- **Chapter 8** presents the results of the experiments which were conducted in order to evaluate the performance of the proposed automatic annotation approach.

- **Chapter 9** concludes the contributions of the thesis and gives an outlook for possible future work.

Chapter 2

Related Work

The aim of this chapter is to review state-of-the-art on automatic image annotation with a special focus on search-based methods. Furthermore, we will review related works around the automatic annotation approach. We will especially investigate research efforts related to the creation of representative datasets, the resolution of tag ambiguity as well as boosting the efficiency of image matching

2.1 Introduction

As discussed before, the aim of automatic image annotation (AIA) is to generate descriptive keywords (tags) for unlabeled images without (or with only a little) human interference (Figure 2.1).

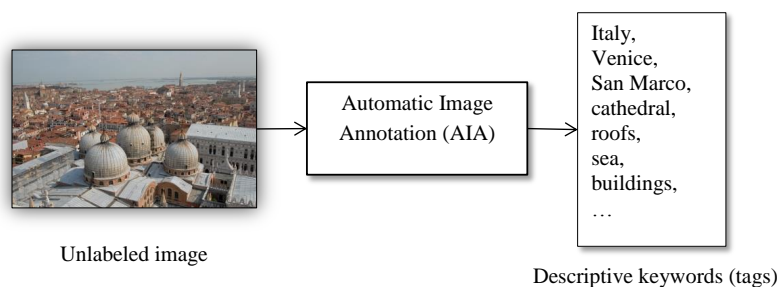


FIGURE 2.1: The goal of automatic image annotation

By digging into the literature on AIA, we can observe two broad categories of approaches. Before the emergence of Web 2.0, most AIA approaches aimed to solve the problem using machine learning techniques or statistical modeling. Researchers describe approaches of this category as *model-based* [Li et al., 2009a, Wang et al., 2012], *image*

categorization-based [Mensink et al., 2010a] or *statistical modeling-based* [Ballan et al., 2013]. The second category of AIA attempts to tackle the problem by exploiting the diverse contextual information and metadata of the huge amounts of images available on the Web of today. In the literature, such methods are known as *model-free* [Li et al., 2009a], *tag propagation-based* [Mensink et al., 2010a], *data-driven* [Ballan et al., 2013] or *search-based* [Wang et al., 2012]. In the course of this thesis we refer to AIA approaches of the first category as *model-based AIA* and we use the term *search-based AIA* to denote approaches of the latter category.

Before we discuss the literature on AIA, we roughly describe the process of identifying similar images based on visual contents. In fact, this process is an essential component for the majority of AIA approaches. Next, we provide a brief description of model-based approaches while search-based approaches are discussed in detail since the approach of this thesis fall into this category. Finally, we explore works which are relevant to particular processing phases of the AIA approach proposed in this thesis. Specifically, we discuss works about crawling image data, resolving the ambiguity of user-supplied tags and improving the performance of image matching.

2.2 Content-based Image Retrieval

The aim of content-based image retrieval (CBIR) is to enable image search based on the visual contents (colors, shapes, etc.) [Smeulders et al., 2000]. Users have access to different choices to conduct the search. In this thesis, we are interested in what is called *target search*, a.k.a. *query-by-example paradigm* [Cox et al., 2000], which retrieves images that are visually similar to a given query image. Such a task is useful for applications like search-based image annotation (refer to Section 2.4) which requires identifying duplicates and images depicting the same objects or scenes as the query image.

Roughly speaking, CBIR systems perform target search as follows. Given a database of images, low-level features are first extracted and represented as numerical vectors called descriptors. Subsequently, in order to find images similar to a query image the same features are extracted from the latter and the corresponding descriptors are compared to those of the images within the database using a distance function.

Image Features

The type and the quality of the extracted features are crucial factors for CBIR systems to succeed. Therefore, the question of identifying “good” image features has received a

great deal of interest by the computer vision community. In general, two kinds of image features can be recognized: *global* and *local* features.

Global features are generated on the basis of the entire image. Most color, texture and shape descriptors fall into this category [Rui et al., 1999]. Global feature can be efficiently computed and they provide a compact representation of image content. However, they are sensitive to clutter and occlusion.

In recent years, local image features showed promising results for many applications such as near-duplicates identification and object detection and tracking [Prasad, 2012]. Among the known algorithms for local features is the pioneering work SIFT (Scale-Invariant Feature Transform) [Lowe, 2004] and its variations, such as PCA-SIFT [Ke and Sukthankar, 2004] and SURF (Speeded-up Robust Features) [Bay et al., 2008]. Identifying similar images based on local features is a three-step process. First, a detection algorithm is applied to identify distinctive regions in the image called keypoints (or interest points). Thereby, common approaches for blob or edge detection can be applied. In the next step, a feature descriptor is created for each keypoint based on the characteristic of the surrounding image patch. Finally, the similarity between two images is defined in terms of keypoint correspondences. Compared to global features, local features have high retrieval accuracy, however, they require high computations. Furthermore, in contrast to global features, where the similarity between two images implies comparing only two image descriptors, image similarity based on local features requires comparing large number of local descriptors.

Image Matching

As mentioned before, the final step of CBIR implies comparing (matching) image descriptors in order to identify similar images. Generally, the matching process employs the notion of distance between two descriptors which correspond to two points in high dimensional space. Euclidean distance and Manhattan distance are among the most used metrics.

To be able to scale to large collection of images, CBIR has to deal with the problem of high dimensionality. Hence, several research efforts have been made to provide efficient dimension reduction and indexing algorithms, in order to speed up the matching process while preserving accuracy. A widely used indexing algorithm is the one proposed by [Ferhatosmanoglu et al., 2001] which is based on K-means clustering. Currently, a state-of-the-art solution is introduced with the technique of *bag of visual words* (BoW). This technique is widely applied with local features. The idea is to consider each keypoint descriptor as a visual word. After that, a clustering algorithm (e.g. K-means) is applied

and the centers of the learned clusters are used as codewords. Next, the codewords are combined in a codebook and each image is represented as a histogram over the words of the codebook. Finally, the similarity between two images is determined based on the distance between their histograms which can be computed using conventional metrics like the cosine measure or the Euclidean distance.

2.3 Model-based Automatic Image Annotation

Model-based approaches for automatic image annotation are based on the idea of finding a mapping between low-level image features and semantic concepts (e.g. sky, car, sea). This is achieved by analyzing a set of already labeled images, called the training set, and creating a corresponding prediction model. Model-based approaches can be classified into two categories: *probabilistic modeling* methods and *classification-based* methods [Liu et al., 2009].

AIA approaches, which apply probabilistic modeling, are generative models which aim to learn the joint probability distribution between image features and keywords. A widely cited work of this category is the machine translation (MT) model proposed by [Duygulu et al., 2002]. MT uses a training dataset of labeled images to learn a mapping between image regions (blobs) and a predefined set of keywords. In contrast to TM, which assumes one to one relationship between image regions and keywords, [Jeon et al., 2003] proposed the cross-media relevance model (CMRM) which learns the joint probability distribution by considering many to many correlations between image blobs and keywords. [Lavrenko et al., 2003] introduced an extension of CMRM, called continuous-space relevance model (CRM), which is able to deal with highly dimensional continuous features. [Feng et al., 2004] also built upon CMRM and proposed the multiple Bernoulli relevance model (MBRM) which uses kernel density estimates to model the images while the keywords are modeled using a multiple Bernoulli process.

Classification-based AIA approaches are discriminative statistical models which treat the problem of automatic image annotation as a classification problem. For this purpose, each keyword is considered as an independent class and a classifier is learned to predict the right class(es) of test images. A widely used method to construct the classifier is the technique of support vector machines (SVM) (e.g. [Chapelle et al., 1999, Cusano et al., 2003]). For detailed discussion about model-based AIA approaches, we invite interested readers to refer to [Wang et al., 2012, Datta et al., 2008, Zhang et al., 2012, Wang, 2011].

Limitations of model-based AIA

A well-labeled training dataset is essential for model-based AIA approaches to succeed. However, creating a training dataset is a laborious manual task. Therefore, available datasets are usually limited to images labeled with a small number of predefined visual concepts. Therefore, model-based approaches have low scalability. Indeed, to scale up, the learning algorithm must be applied periodically every time new concepts or images are added, which demands high computations. Furthermore, the accuracy of classification models drops significantly when number of the to-be-learned concepts (classes) increases [Deng et al., 2010].

With respect to the quality of the produced annotations, [Enser et al., 2005] defined two main limitations. The first problem, named *visibility limitation*, arises since the produced tags are usually limited to objects appearing in the image. Hence, some conceptual material or abstract concepts (e.g. happiness, anguish) cannot be predicted because they are non-visible in nature, i.e., there are no corresponding salient features that can capture them. The second problem, called *generic object limitation*, indicates that model-based AIA are usually limited to generic/perceptual vocabulary (e.g. "sky", "beach", "horse", etc.). Such vocabulary is not adequate since users aim to retrieve images identified by specific or unique concepts, e.g., proper names.

2.4 Search-based Automatic Image Annotation

The ubiquity of photo sharing services motivated researchers to explore methods for leveraging the highly available image metadata and the associated contextual information in order to automatically annotate unlabeled images. The advantage of the new research line, i.e., search-based AIA, is the fact that it is able to address the mentioned limitations of model-based AIA approaches. First, search-based AIA do not require a training dataset of already labeled images. Second, search-based AIA can resolve the problems of *visibility* and *generic object* limitations since it leverages user tags, which are diverse in nature, as a resource for annotating unlabeled image. Therefore, there is no restriction on the characteristics of the produced tags.

In general, for an unlabeled photo, search-based AIA retrieves a set of *similar* images from a large scale database of already labeled images, such as the web or specialized photo sharing platforms, e.g., Flickr. Subsequently, tags/keywords of similar images are analyzed and propagated to the target image. More specifically, to identify similar images, a two-phase search process is applied:

1. **Semantic/Contextual Search:** this phase aims at identifying a set of related images based on metadata or contextual information associated with the target image. In the early years, the metadata have been provided either manually in the form of initial keywords, or in the case of web images, they were extracted from the associated titles, surrounding texts, URLs, etc. Currently, additional kinds of contextual information can be obtained, such as timestamps, location information or the social context of the user who created the photo.
2. **Search by Image Contents:** the goal of this phase to refine the results of the previous search phase to the subset of images which are visually similar to the target image. To achieve this, content-based image retrieval (CBIR) techniques are applied.

Note that although most search-based AIA approaches follow this two-phase process, some approaches use only one of the mentioned search phases.

In the next step, the keywords/tags of the similar photos, i.e., the candidate annotations are propagated to the target image. Thereby, a ranking mechanism is applied before the final annotations are delivered. Generally, different kinds of relevance relationships are combined to rank candidate annotations:

- **Image-to-Image Relevance (IIR):** this kind of relevance is usually defined according to the visual similarity between two images as determined by the applied CBIR algorithm.
- **Word-to-Image Relevance (WIR):** it refers to the conditional probability of generating an image given a keyword/tag.
- **Word-to-Word Relevance (WWR):** it refers to the semantic similarity/relatedness between two keywords.
- **Word Importance (WI):** indicates how popular a keyword is. This can be estimated using contextual information, such as the popularity of the word for a given group of users or using specialized linguistic resources like a thesaurus or a dictionary.

In the following the literature on search-based AIA is discussed in detail. We will demonstrate how different works estimate each on the mentioned relevance scores and how they combine them. We refer with *target image* to the photo which we want to annotate automatically. The set of images obtained from database of labeled images through a semantic/contextual search is called the set of *semantic neighbors*. Moreover,

we refer to the subset of images which are visually similar to the target image as the *visual neighbors*. A taxonomic representation of the works discussed in the following sections is provided in Figure 2.2.

2.4.1 The World Wide Web as Annotation Resource

In the early years, search-based AIA aimed at leveraging the metadata associated with Web photos, such as, URLs, captions, surrounding text, etc. to annotate unlabeled images. There are two common observations that can be inferred from these works. First, a large part of them assume that the target image is associated with initial keywords. They are either manually provided or if the target is a Web image, then initial keywords can be extracted from the associated metadata. Second, the majority of these approaches adopt content-based image retrieval techniques based on global image features (e.g. color histogram, color correlogram, texture, etc.).

AnnoSearch is one of the earliest systems of this category [Wang et al., 2006a]. It assumes that the target image is associated with at least one keyword. The initial keywords are used to query the Web for semantically similar images. After that, CBIR is applied to identify the visual neighbors of the target image. For this purpose, 36-bin color correlogram image feature is used and indexed using 32-dimension hash codes [Huang et al., 1997]. In order to generate a list of candidate keywords, the authors apply the *search result clustering* (SRC) algorithm proposed by [Zeng et al., 2004] to cluster the visual neighbors based on the associated textual metadata. Subsequently, a name is generated for each cluster and used as a candidate keyword. To rank the candidate keywords, the authors propose two methods: the *maximum cluster size criterion* which ranks a keyword (cluster name) according to the number of images found in the corresponding cluster. The second method, called *average member image score criterion*, ranks a keyword based on the average visual distance between the images found in the corresponding cluster and the target image.

[Li et al., 2006] extended *AnnoSearch* to address the problem of requiring initial keywords to start the annotation process. For this purpose, they propose an algorithm, called *Multi-Index*, for indexing high dimensional image features. Consequently, the semantic search phase is suppressed and the CBIR process is directly applied to determine the visual neighbors.

[Wang et al., 2006b] proposed the scalable search-based image annotation (SBIA) approach. Thereby, the visual neighbors are identified from web images by applying CBIR based on 64-dimensional feature that combines color moments, auto-correlogram and

color texture moments. Candidate keywords are extracted from the metadata of the visual neighbors and ranked by combining two scores: 1) image-to-image relevance (IIR) between the target image and the visual neighbors. Here, the Euclidean distance between the corresponding feature descriptors is computed. And 2) a word-to-image relevance (WIR) which determines the importance of a candidate annotation based on its similarity to the keywords associated with the visual neighbors. For this purpose, two strategies were proposed. Word *prominence* measure which indicates how distinctive a keyword for a given image is. It is estimated according to the number of occurrences of the keyword in the metadata of the associated web image. The second strategy is the image frequency-inverse keyword frequency (*IFIKF*) which is based on the *TF-IDF* (term frequency-inverse document frequency) measure [Baeza Yates et al., 1999]. The authors further improved their approach in [Wang et al., 2006c]. A main extension is a new ranking procedure based on the techniques of Random Walk with Restarts (RWR). First, a fully connected graph is constructed, in which each candidate keyword represents a node. The edges are weighted according to the similarity between the connected nodes. To that end, the authors introduce a measure for word-to-word relevance (WWR). For two words, each of them is used to query a Web image searcher (e.g. Google Image [Google, 2014]) and the number of returned results for each of them is counted. Next, the co-occurrence of the two words is calculated by using both of them together to query the search engine. Finally, the similarity between the two words is calculated by dividing their co-occurrence count on that of the word of the minimum occurrence. After generating all weights for the graph, the RWR algorithm is applied to determine the final relevance scores.

The work of [Rui et al., 2007a] adopted the same measures of WWR and WIR proposed by [Wang et al., 2006b,c] for ranking keywords extracted from the metadata of Web images. However, they investigated using Dempster-Shafer theory [Shafer, 1976] to combine WWR and WIR scores for ranking the final annotations. In their later work [Rui et al., 2007b], the authors follow the same procedure as in AnnoSearch [Wang et al., 2006a] to extend the initial keywords using the search result clustering algorithm (SRC) [Zeng et al., 2004]. Their main contribution is a procedure for re-ranking the initial as well as the extended keywords based on a bipartite graph reinforcement model (BGRM). To achieve this, two disjoint graphs corresponding to the initial and the extended keywords (which were acquired by SRC), respectively, are created. The two graphs are then connected to build a bipartite graph. An edge is drawn between two nodes of the two disjoint graphs based on their similarity. More specifically, an initial keyword x (from the first graph) is linked to an extended keyword y (from the second graph) if y is extracted from the search results of x or if x is *similar* to y . Similar words are identified according to WordNet [Miller, 1995] and the Jiang & Conrath measure (JCN) [Jiang

and Conrath, 1997]. Finally, a graph reinforcement learning algorithm is applied on the generated bipartite graph to boost the ranks of the candidate annotations.

[Liu et al., 2007] proposed the dual cross-media relevance model (DCMRM) which assumes that a predefined lexicon exists, from which candidate keywords can be extracted. DCMRM aims at maximizing the joint probability distribution of images and words based on the expectation over the words of the predefined lexicon. Specifically, DCMRM assumes that the probability of observing a keyword w and the target image I_u is mutually independent given a keyword $v \in V$ from the predefined lexicon V as given in the following formula:

$$w^* = \arg \max_{w \in V} \sum_{v \in V} P(I_u|v)P(w|v)P(v) \quad (2.1)$$

Where $w^* \in V$ is the word which best describes I_u , $P(I_u|v)$ is the word-to-image relevance, $P(w|v)$ is the word-to-word relevance and $P(v)$ is the importance of the word v . The annotation procedure starts by assuming that the target image is associated with initial keywords. Next, a two-phase search is applied to identify the set of visual neighbors of the target image. To calculate WIR (i.e., the term $P(I_u|v)$), two scores are combined: 1) the semantic similarity between the candidate keyword and the set of keywords associated with the visual neighbors and 2) the similarity between the target image and the visual neighbors. To calculate the semantic similarity between two words, the authors present a measure based on the idea of the Normalized Google Distance (NGD) [Cilibrasi and Vitanyi, 2007]. To compute $P(w|v)$ and $P(v)$ the two terms are rewritten as $P(w, v) = P(w|v).P(v)$ and the proposed NGD-based distance between the two words is used as an estimator.

[Xia et al., 2008] also consider annotating Web images using initial keywords extracted from the associated metadata. In contrast to the discussed works, the authors consider using WordNet and the technique of latent semantic analysis (LSA) [Deerwester et al., 1990] to improve the quality of the generated annotations. First, the initial keywords are ranked using a combination of the standard TF-IDF weight and a score for word visibility based on WordNet. Thereby, the visibility of a word is defined according to its ability to describe the visual contents of the image. Next, additional candidate keywords are obtained by applying latent semantic analysis (LSA) to determine synonyms for the initial keywords. To rank candidate keywords, each of them is first used to obtain images from the Web. Images of each result set are then clustered based on their visual features. Finally, the originating keyword is given a score based on the average similarity between the target image and the centers of the top matching visual clusters.

[Mei et al., 2008] place a special emphasis on the semantic coherence between candidate annotations. An example of semantically coherent words is the pair "cat" and "tiger" while the word pair "indoor" and "sky" is not. Their annotation approach consists of three steps: first, given a set of labeled images (the training set), semantic clusters are built based on the semantic similarity of the associated keywords using WordNet. In the next step, a semantic distance function (SDF) is learned for each cluster. The rationale behind this is that the distance between image features should be adapted according to the semantic of the corresponding group. Finally, a search-based approach is used to annotate a new image. For this purpose, CBIR process is performed where the similarity between the target image and an image of a certain semantic cluster is determined according to the learned SDF. Next, similar images are ranked and their annotations are propagated to the target image.

2.4.2 Community Photos as Annotation Resource

Compared to the World Wide Web, collaborative image sharing platforms (e.g. Flickr) provide a structured resource in which users explicitly define the association between the images and their textual description (tags, titles, etc.). Furthermore, the new platforms provide rich contextual information about images, such as their locations, the users who uploaded them and their social contexts. In what follows, we review works on AIA that leverage community-tagged photos with the associated contextual information.

2.4.2.1 Classification-based Approaches

Works of this category use community labeled photos to create training dataset and apply classification methods to predict labels for un-annotated images.

[Lindstaedt et al., 2008] presented *Tagr* - a system for tag recommendation using a subset of Flickr images. *Tagr* deals with a predefined subset of concepts taken from the "fruit & veg" pool on Flickr. It applies a model-based approach and train a multi-label SVM classifier using the MPEG-7 Color Layout descriptor as a classification feature.

Another work which leverages Flickr for building a training dataset was proposed in [Chen et al., 2008]. The authors create a training dataset for 62 predefined concept using two crawling strategies: 1) a photo-level data collection strategy in which each of the 62 concepts are used to retrieve photos from Flickr and 2) group-level strategy which retrieves photos from Flickr groups that match the predefined concepts. In the next phase, training datasets obtained by both data collection strategies are fused and used to train an SVM classifier in which each concept correspond to a class. For each

image 369 dimensional feature vector is created comprising color, texture, and edge features. To automatically annotate a new image the classifier is used to predict the top concepts. Next, the predicted concepts are used to retrieve the best matching Flickr groups. Finally, candidate annotations are extracted from the most used tags of the top groups.

2.4.2.2 Semi-automatic Approaches

A considerable body of works assumes that the target image is partially annotated and aims at extending the initial annotations with tags extracted from community photos.

[Sigurbjörnsson and van Zwol, 2008] presented an approach for extending initial image tags based on tag co-occurrence analysis. To this end, two measures for tag co-occurrence were proposed: a symmetric measure based on Jaccard coefficient and an asymmetric measure which normalizes the co-occurrence of a tag pair based the number of occurrences of one of them. Candidate tags are then re-ranked using a combination of two scores: tag *aggregation* and tag *promotion*. Tag aggregation can be determined according to either of two strategies. A *vote* strategy which determines the importance of a tag based on the number of initial tags which extended it. The second strategy is based on the sum of the co-occurrences of the candidate tag with each of the initial tags. The promotion score attempts to consider the tagging behavior on Flickr to determine the importance of a tag. Accordingly, the authors define three measures: *stability* and *descriptiveness* which penalize tags with low and high usage frequency, respectively, and *rank-promotion* which weights a tag t extended by an initial tag u according to its position in the list of the candidate tags extended by u .

[Anderson et al., 2008] adopted the co-occurrence model of [Sigurbjörnsson and van Zwol, 2008] as a language component in a system for automatic image annotation called *Tagex*. *Tagex* also provides a vision component which is able to produce tags for unlabeled images by applying a model-based AIA approach called ALPIR [Li and Wang, 2008]. To annotate a new image, the two models are run to produce lists of candidate annotations. Finally, *Borda voting* is used to combine the ranks produced by both models.

[Wu et al., 2009] introduced a model for tag recommendation which also exploits tag co-occurrence. The proposed approach extends initial image tags using three kinds of co-occurrence measures. The first measure adopts the tag co-occurrence of [Sigurbjörnsson and van Zwol, 2008]. The second estimates the similarity between two tags based on the correlation between their visual contents. For this purpose, the authors used the visual language model (VLM) algorithm [Wu et al., 2007] to generate a visual representation for each tag. Subsequently, the visual-content similarity of two tags is determined according

to the distance between the corresponding VLMs. The final type of tag co-occurrence determines the similarity between two tags with regard to the target image. This is defined in terms of the distance between the VLM of target image and the VLM of each of the tags. Finally, the three similarity scores are combined using the Rankboost algorithm [Freund et al., 2003] to rank the candidate annotations.

[Li et al., 2009b, 2008] proposed a method for estimating tag relevance (i.e. word-to-image relevance) based on neighbor voting. The goal is to rerank a set of initial tags assigned to a target image. For this purpose, they adopt the same CBIR technique as in [Li et al., 2006, Wang et al., 2006b] to identify a set of visual neighbors from Flickr. Next, a relevance score is given for each tag according to the number of visual neighbors which are annotated with it. To reduce the bias caused by a single user annotating a large number of images with the same collection of tags, the visual neighbor set is restricted to a single image per user. Furthermore, the authors introduced a prior for each tag based on its frequency of occurrence in the complete image set.

A further technique for neighbor voting was proposed by [Makadia et al., 2010]. First, CBIR process based on global image features is used to identify a ranked list of visual neighbors for the target image. The target image is then annotated, by, first, propagating the tags of the top visual neighbor, followed by the tags of the second top one, and so on. The importance of a tag given a visual neighbor is determined based on its frequency in the complete training set.

[Guillaumin et al., 2009, Verbeek et al., 2010] proposed an AIA, called *TagProp*, which also applies the neighbor voting technique. TagProp weights the votes of the visual neighbors using distance-based or rank-based techniques. To compensate tag frequencies, the authors propose a tag-specific sigmoid to boost the probability for tags with low frequency and to lower the probability of very frequent tags. TagProp was extended by [Mensink et al., 2010b] through a pseudo relevance feedback model to boost the ranks of the visual neighbors.

2.4.2.3 AIA using Spatial and Temporal Contexts

The popularity of geotagged community photos inspired interesting applications, such as event discovery [Yuan et al., 2008, Naaman et al., 2004] and automatic landmark identification [Quack et al., 2008, Zheng et al., 2009, Kennedy and Naaman, 2008].

A notable work was presented by [Kennedy et al., 2007] to extract location-based representative tags. For a given geographical area, related geotagged photos are obtained from Flickr and clustered according to their geographical coordinates. The images are

then clustered based on their visual-contents. Next, information about the geographical as well as the visual clusters is used to identify representative tags for the corresponding geographical area using a specialized TF-IDF measure. Furthermore, the authors investigated identifying the semantic of the tags by automatically classifying them into location or event tags. The mined tags are then used by the *World Explore* visualization tool [Ahern et al., 2007].

Among the earliest works which exploit geotagged community photos for automatic annotations is the *LOCALE* system proposed by [Naaman et al., 2003]. *LOCALE* relies on proximity information to propagate the captions of already tagged photos to other photos taken in the same location.

ZoneTag is a mobile phone application which uses location-based contextual information to suggest tags for newly captured photos [Naaman and Nair, 2008]. For this purposes, *ZoneTag* exploits several contextual clues to mine candidate tags. These include location-based tagging history of the user who took the photo, the location-based tagging history of other users and Yahoo services for identifying point of interest (e.g. restaurants, cafes) and possible events in the location of image capture. Finally, the authors apply heuristics to score the candidate annotations according to the resources from which they were extracted.

Another mobile phone application is presented in [Cheng et al., 2010]. The application leverages information provided by the GPS and the compass sensors of the mobile phone for the tag mining process. First, the boundaries of the geographical area corresponding to the target image are determined and divided into a grid of overlapping cells. Next, images corresponding to each cell are obtained from Flickr and clustered based on their visual contents. For this purpose, agglomerative clustering and the bag of visual word (BoW) representation for image features are used. Next, representative tags are extracted for each cluster by applying a TF-IDF-based measure. Thereby, the term frequency for a tag associated with a given cluster corresponds to its number of occurrences among the images in that cluster. The inverse document frequency corresponds to the ratio of tag occurrence in the whole set of clusters. To annotate a new image its visual neighbors are first determined. Next, GPS and compass information are fused to determine the best matching cluster. Finally, representative tags of top clusters are transferred to the target image.

MonuAnno is a system with a focus on automatically annotating landmark photos [Popescu and Moëllic, 2009]. For this purpose, a dataset of 5,000 landmarks was created by crawling geotagged photos with the associated user-tags from Flickr and Panoramio. The photos in the dataset were indexed by applying the BoW (bag of visual words)

technique on SIFT features. The annotation of new unlabeled images is performed using two-step K Nearest Neighbor (KNN) algorithm. First, photos of landmarks found in the geographical neighborhood (up to 1 KM) of the target image are identified. Next, candidate landmarks are determined based on the visual similarity between the target image and the geographical neighbors. Finally, a verification procedure is applied to eliminate false positives and the name of the top matching landmark is propagated to the target image.

[Abbasi et al., 2009] proposed an approach for automatic image annotation by clustering community geotagged photos based on the geographical location, the user-supplied tags, and the visual contents. The annotation of an unlabeled image is done as follows: the top matching cluster is determined based on the location of the target image and the same visual features which are used to build the clusters. Consequently, tags of the matching cluster are chosen as candidate annotations. Finally, the tags are ranked according to the number of users which applied them.

An approach which is most relevant to this thesis is the *SpiritTagger* system proposed by [Moxley et al., 2008]. The automatic annotation process of SpiritTagger consists of two phases: a geographical mining phase, in which images taken in the spatial proximity of the target image are obtained from Flickr. In the second phase, CBIR process is applied to identify the visual neighbors of the target image out of the geographical neighbors. Next, candidate tags are extracted from the annotations of the visual neighbors and ranked by combining their *local* as well as *global* occurrences. Local occurrence of a tag indicates its frequency in the collection of geographical neighbors, while the global occurrence refer to its total occurrence in the complete set of Flickr images. Finally, tags with higher local frequency are favored and proposed as annotations for the target image.

Two additional works that follow a procedure similar to that of SpiritTagger [Moxley et al., 2008] were proposed by [Silva and Martins, 2011] and [Mitran et al., 2013], respectively. Both works focus, however, on investigating methods for combining different kinds of tag relevance estimators. For instance, [Silva and Martins, 2011] explored supervised (e.g. RankBoost) as well as rank aggregation methods (e.g. CombSUM and CombMNZ [Shaw et al., 1994]) to combine relevance scores which are derived from tag usage pattern, geospatial proximity, and image visual similarity.

In contrast to the presented works which ignore or implicitly use the temporal context, some researches explicitly investigated the importance of temporal information for AIA. [McParlane and Jose, 2013] proposed using tag temporal co-occurrence to improve the

annotation generated by state-of-the-art AIA approaches. Thereby, the authors investigated integrating tag co-occurrences over different time-windows (i.e. hours, days, months and years) in the annotation process.

2.4.2.4 AIA using Social Context

Currently, there is an increasing trend to exploit relational information, i.e., the social context provided by photo-sharing websites to assist the process of automatic image annotation. In general the social context is used in a complementary process for extending candidate annotation produced by a given AIA approach (e.g.[Denoyer and Gallinari, 2010, Sawant et al., 2010]).

In their early work, [Garg and Weber, 2008] proposed a method for providing immediate tags proposals for the user annotating some photo based on the entered tags. The suggestions are selected from tags: which 1) were used often by the user in the past, 2) which the user often used with the initial tag (co-occur) to annotate his photos, and 3) from tags of Flickr groups that correspond to user interest and co-occur frequently with the initial tag.

Sawant et al. [Sawant et al., 2010] proposed a method for extending initial annotations obtained by ALPIR image annotation system [Li and Wang, 2008]. In order to automatically annotate the image of a given user, a set of initial annotations A is first produced using ALPIR. Next, the set of tags T_{social} is built out of the annotations applied by the users belonging to the social network of the initial user. Finally, a candidate tag $t_i \in T_{social}$ is ranked based on its co-occurrence with the initial annotations, A .

[Rae et al., 2010] also considered extending initial image annotations. For each extended tag four scores are calculated based on its co-occurrence with the initial tags. These scores are derived based on four layers for social context: 1) the user personal context which corresponds to user tagging history, 2) the social context of the user which consists of tags used by all users found in the contact list of the initial user, 3) the social group context which contains all tags used in groups to which the initial user belongs, and 4) the collective context which contains tags corresponding to photos posted by all users. The final tag ranks are then determined by combining the four calculated scores using Borda voting.

To annotate unlabeled image, [Elahi et al., 2010] applied social network analysis to identify the *central user*, which has high annotation activity, from the social network of the uploader of the image. The idea is to identify images of the central user which were taken in the same location as the input image. Next, the tags of the central users are propagated to the input image.

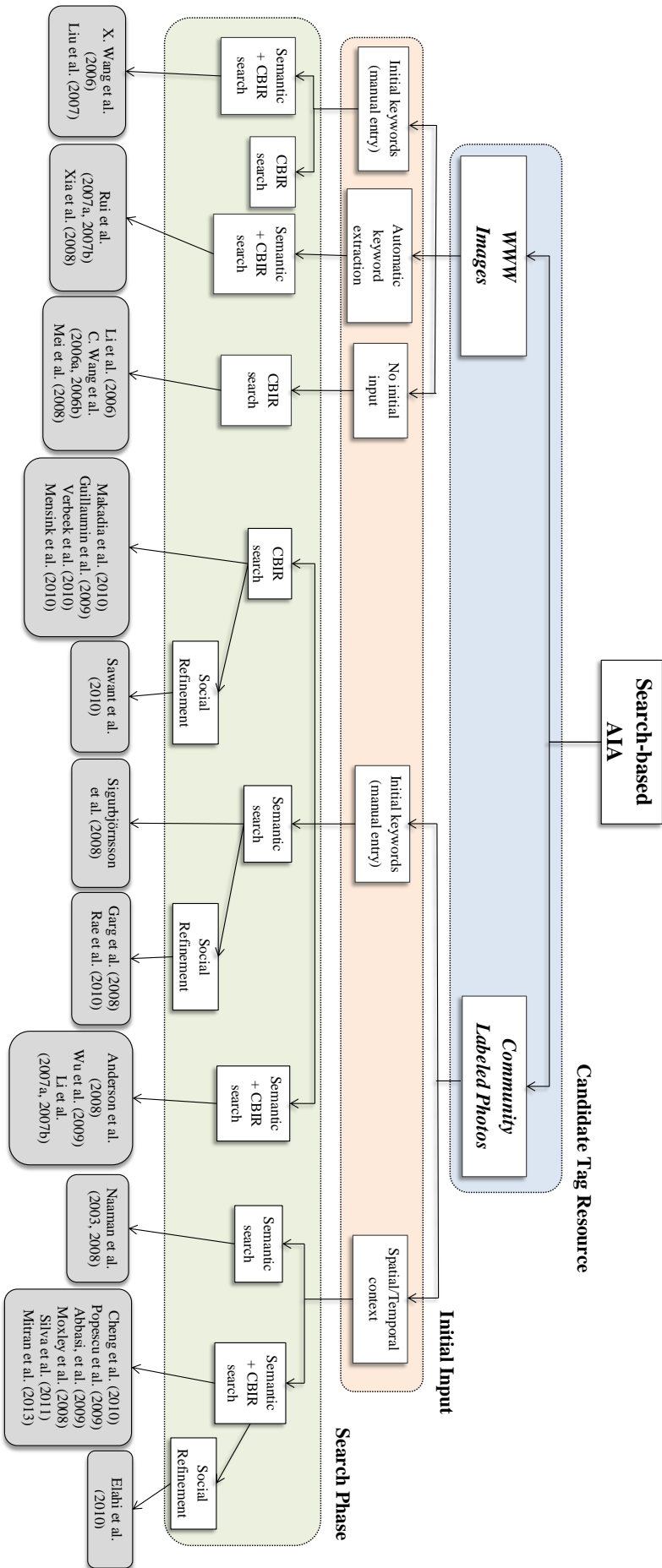


FIGURE 2.2: Taxonomy of the search-based AIA approaches reviewed in this chapter

2.5 Related Work around Search-based Image Annotation

Based on the above review, we can identify three main aspects which influence the quality of search-based image annotation approaches. The first aspect is concerned with the repository of labeled images from which the semantic and visual neighbors of the target image are retrieved. In order to enable annotating images of different natures, the images in the repository have to be representative, i.e., they have to cover various topics, locations, events, etc. Second, search-based AIA approaches deal with user-tags as a main resource for candidate annotations. However, user-tags are noisy and inherently ambiguous. To ensure the quality of the mined tags, there is an urgent need to provide solutions for resolving tag ambiguity. Finally, CBIR is a core component for the majority of automatic image annotation approaches. Therefore, a special emphasis should be placed on the kind of used image features and the speed as well as the accuracy of the applied CBIR process.

In the following, we discuss the state-of-the-art regarding the mentioned aspects.

2.5.1 Geo-based Photo Crawling

Collecting data from community contributed photos for the purpose of automatic image annotation has been investigated in several works. The goal is to create a dataset of user-labeled images which can be pre-processed and indexed to achieve efficient retrieval. Since the approach of this thesis relies on the location context, we focus on methods for data crawling from online photo sharing platforms based on location information. Currently, the significant amount of geotagged photos available online gave rise to a number of strategies for location-based data crawling.

[Keßler et al., 2009] proposed an approach to crawl geotagged photos based on keyword search. For this purpose, photo sharing services are first queried using keywords (e.g. city names). Next, all geotagged images annotated with these keywords are retrieved. The datasets presented in [Hays and Efros, 2008, Kalantidis et al., 2011, Tolias and Avrithis, 2011, Hollenstein and Purves, 2013, Weyand et al., 2012] were created by using the geographic query feature provided by Flickr API. The queries are built based on the geographic boundaries of specific cities or urban centers. A first effort to build world-scale photo dataset was introduced in [Quack et al., 2008]. Initially, the authors divide the world map into a grid of overlapping tiles. After that, the boundaries of each tile are used to query Flickr. [Crandall et al., 2009] also presented a data collection strategy for creating a world-scale photo dataset from Flickr. The main focus of the proposed method is to create a sample with a real spatial distribution. That means, the

density of the photos collected from a given place should reflect the popularity of that place among photographers. The crawling method starts by randomly selecting a photo identifier from the pool of Flickr photo identifiers. Next, the uploader of that photo is identified and the corresponding geotagged photos are downloaded with the associated metadata. Additional photos are then acquired by traversing the friendship graph of the initial user to identify new users and downloading the corresponding geotagged photos. To obtain additional data, the complete process is repeated by selecting a new photo identifier.

Recently, a number of photo datasets which provide location information (explicitly or implicitly) have been made available for research purposes. For the Photo Annotation and Retrieval Task, *ImageCLEF* initiative [ImageCLEF, 2014] provides a dataset based on *MIRFlickr* [Huiskes and Lew, 2008]. It contains one million Flickr images with a subset of 25,000 manually annotated photos. *MIRFlickr* provides different kinds of metadata about the downloaded images, such as the EXIF files and the associated user-tags. However, by investigating the EXIF descriptors, we found out that location information is either missing or inaccurate for a large part of the photos in the dataset. *NUS-Wide* is another dataset based on Flickr [Chua et al., 2009]. It consists of 269,648 images with the associated user-tags as well as six types of low-level image features. Additionally, the dataset provide a ground-truth for 81 concepts. However, only a small part of the photos in the dataset are geotagged (around 50,000). Additional dataset of about one million photos was introduced in [Kalantidis et al., 2011]. The data were crawled from Flickr and correspond to 22 *European cities*. The dataset was extended in [Tolias and Avrithis, 2011] to 40 world cities with a total of about 2,23 million images. However, these datasets provide only the photos without the associated metadata. The authors of [Weyand et al., 2012] provide a script for a dataset called *Paris500k*. The dataset contains more than 500 thousands photos taken in the city of Paris. A further dataset with a main focus on reverse geotagging is presented by the *MediaEval* benchmarking initiative [MediaEval, 2014]. The dataset, named *MediaEval Placing Task 2013 Data Set* contains around nine million geotagged images crawled from Flickr [Hauff et al., 2013]. User tags are also provided, however, in their raw "noisy" form. Additionally, the authors did not give any information on the applied crawling strategy and the spatial representativeness of the data.

2.5.2 Resolving Tag Ambiguity

Collaborative tagging as a common practice of Web 2.0 led to complex networks of users, tags and resources which are currently known under the name folksonomy. The term folksonomy is a portmanteau of the two words "folk" and "taxonomy" and it was first

introduced by [Vanderwal, 2010]. According to the degree of user collaboration, Vanderwal classifies folksonomies into two main categories: broad and narrow folksonomies. While in broad folksonomies, such as the bookmarking website del.icio.us [del.icio.us, 2014], multiple users annotate the same resources with a variety of terms, narrow folksonomies show a lower degree of user interaction. Indeed, in narrow folksonomies the tagging activity is mainly performed by the content creators. Image folksonomies like Flickr belong to this category.

Tag Nosiness

Due the freedom and uncontrolled manner of tag creation, tags suffer from several intrinsic problems. [Mathes, 2004] defines two main problems of user-supplied tags: *ambiguity* and *lack of synonym control* a.k.a *redundancy* [Gemmell et al., 2009]. We use the term tag *noisiness* to encompass both problems.

Ambiguity arises when the same tag is used to indicate different meanings. [Weinberger et al., 2008] defined different kinds of tag ambiguity:

- **Word-sense ambiguity:** in this case, the right sense of the word is determined according to the context. For example, the word "bank" can mean a financial institution or a river bank.
- **Language ambiguity:** here, the word in one language can mean something different in another language. For example, the word "Gift" means poison in German and present in English.
- **Temporal ambiguity:** the exact meaning of a temporally ambiguous word is determined according to a certain date or time period. For example, "World Cup" of 2006 or 2010.
- **Geographic ambiguity:** the meaning of the word is defined according to a given geographical context. For example, "Cambridge" is a city in the UK as well as in Massachusetts, USA.

Tag redundancy emerges when different tags are used to describe the same thing. Here, we can also recognize different types:

- **Lexical variations:** users tend to adopt different lexical styles to express the same thing. The main cause of this problem is that most annotation systems allow using single-word tags only. To overcome this limitation, multi-word tags

are combined in different ways. For instance, a tag referring to the "United States" can be expressed using a hyphen "united-states", an underscore "united_states", or the camel case "UnitedStates".

- **Different naming conventions:** it is common that people use different naming styles to refer to the same thing. For example, to annotate an image taken in New York some people use "New York" while other use the abbreviation "NYC".
- **Vernacular geography:** it is very common to use informal geographical vocabulary to refer to geographical entities. For instance, to indicate a neighborhood, a statue, a pub, a hill or an area, people might use terms which are not recorded before on maps or gazetteers. For example, the landmark building called "Spinnaker Tower" is known among the people of in Portsmouth as "The Pregnant Pin". Another example is the term "city business center" which is expressed using different local names. People in North America prefer the word "downtown", while "city-centre" is widely used in England and the abbreviation "CBD" (Central Business District) is most common in Australia.
- **Multilingualism:** refers to using different languages to denote the same thing, e.g., "tower" in English and "Turm" in German.

Mining Tag Relatedness for Resolving Tag Ambiguity

In order to resolve tag ambiguity¹, researches worked on techniques for identifying related tags by analyzing their usage patterns. Generally, the focus of most solutions is to exploit tag co-occurrence statistics and clustering techniques to identify groups of tags which share similar semantics. Consequently, tag relatedness information is used to identify the real meaning of individual tags.

Formally, a folksonomy F is defined as a tuple $F = \{T, U, R, A\}$ [Hotho et al., 2006a] where T is the set of tags contributed by a set of users U to annotate a set of resources R . A co-occurrence of two tags $t_1, t_2 \in T$ indicates that they are used by one or more user to describe the same resource $r \in R$. This is captured by the assignment relation, $A \in U \times T \times R$.

With respect to the three dimensions of the folksonomy T , U and R , a tag can be represented as a vector in one (or in combination) of three possible real vector spaces: $\mathbb{R}^{|T|}$, $\mathbb{R}^{|U|}$ and $\mathbb{R}^{|R|}$, respectively [Cattuto et al., 2008]. In the following we will refer

¹Most researchers use the term "tag ambiguity" to refer to tag ambiguity and tag redundancy at the same time.

to these real spaces as \mathbb{R}^T , \mathbb{R}^U and \mathbb{R}^R , respectively. In \mathbb{R}^T representation, called *tag-context*, a tag t is defined as a vector $v(t) \in \mathbb{R}^T$ with the entries corresponding to the unique tags in the folksonomy. The value of an entry corresponding to another tag $t' \in T$ is given by the number of resources that have been tagged with t and t' at the same time. The second kind of tag representation is called *user-context*. The entries of the tag vector $v(t) \in \mathbb{R}^U$ correspond to the unique users in the folksonomy. The value of an entry related to a user $u \in U$ indicates how often u has used t in his annotation activity. The last kind of tag representation is the *resource-context*. The entries of the tag vector $v(t) \in \mathbb{R}^R$ correspond to the unique resources in the folksonomy. The value of an entry related to a resource $r \in R$ corresponds to the number of times in which t was used to annotate r .

Approaches for tag relatedness use one or more of the presented tag contexts to identify related tags (e.g. [Cattuto et al., 2008, Hotho et al., 2006b, Specia and Motta, 2007, Begelman et al., 2006]). The choice of the context is usually determined according to the type of the folksonomy, i.e., narrow or broad. In image folksonomies like Flickr, where low user collaboration is observed, tag-context is widely used.

[Begelman et al., 2006] proposed a tag relatedness measure which is based on tag co-occurrence counts (i.e. the tag-context). In that approach, the co-occurrence of each tag pair is computed and a cut-off threshold is used to decide whether two tags are related. This threshold is determined using the first and the second derivatives of the tag co-occurrence curve. Finally, tag clusters are identified by organizing the tags in a similarity matrix and applying spectral bisection clustering algorithm on it. [Specia and Motta, 2007] also exploit the tag-context representation to identify groups of similar tags. First, the tags are organized in a co-occurrence matrix with the columns and the rows corresponding to the tags. The entries of the matrix represent the number of times two tags were used together to annotate the same resource. Next, each tag is represented by a co-occurrence vector and the similarity between two tags is calculated by applying the cosine measure on the corresponding vectors. Finally, the tag similarity matrix is fed into a clustering algorithm to identify groups of similar tags. In a similar manner, [Gemmell et al., 2008a,b] apply clustering on the tag similarity matrix to identify groups of related tags. However, to create the similarity matrix, the authors represent the tags as vectors where the entries correspond to TF-IDF scores. Thereby, resources (e.g., images, web pages) are considered as documents while the tags are considered as terms. Finally, agglomerative clustering is applied to identify groups of similar tags. [Simpson, 2008] investigated using graph techniques to identify groups of similar tags. For that purpose, the author propose normalizing tag co-occurrences using Jaccard coefficient and organizing the tags in a similarity graph. Subsequently, an iterative divisive clustering algorithm is applied on the graph to identify clusters of related tags. Another

graph-based approach was introduced in [Papadopoulos et al., 2010]. Thereby tags are organized in a weighted graph, in which the nodes correspond to tags while the edges are weighted according to the structural similarity between the nodes. Consequently, the similarity between two nodes is defined in the terms of their common neighbors. In line with the tag disambiguation approach proposed in this thesis, [Weinberger et al., 2008] investigated statistical techniques for identifying ambiguous tags. For this purpose, the tags are first represented as probability distributions based on their co-occurrence with the top frequent tags in the folksonomy. Subsequently, ambiguous tags are determined according to the change in their probability distributions as a consequence of adding new tags. To quantify this change a weighted version of Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] is used.

2.5.3 Improving Image Matching

Applications for finding similar images have proven very successful since the introduction of the groundbreaking SIFT algorithm in 2004 [Lowe, 2004]. To speed up the process of identifying similar images, earlier works focused on providing efficient mechanisms for indexing the huge number of produced keypoint descriptors. For instance, [Ke et al., 2004] proposed using locality-sensitive hashing to index SIFT descriptors. Recently, the Bag of Words (BoW) representation has seen much use since it provides a faster matching [Nister and Stewenius, 2006, Philbin et al., 2007].

An additional step to speed up image matching is made by works that deal with the problem of reducing the number of generated keypoints. [Foo and Sinha, 2007] proposed a strategy for reducing the number of SIFT keypoints based on their contrast (intensity) values. They showed that an efficient matching can be achieved by ranking the keypoints according to their contrast values and selecting the top N . Another method for identifying if a keypoint is useful for the matching was introduced by [Turcot and Lowe, 2009]. The usefulness of a SIFT keypoint corresponding to a query image is determined through its counterparts in other matching images. This is achieved by using RANSAC algorithm to determine if the features are geometrically consistent. In more recent work on near-duplicates, [Dong et al., 2012] showed that SIFT features with near-empty regions are a major source of false positives. Accordingly, the authors propose restricting the process of identifying near-duplicate images to the subset of keypoint descriptors which have rich internal structure. To achieve that, they apply entropy-based filtering on SIFT features.

Another family of works investigated using visual attention and saliency maps [Itti et al., 1998] to prune the number of keypoints generated by local features extraction algorithms.

The idea is to apply visual attention algorithms to identify regions in the image, called salient region, which most attract human’s attention. In order to reduce the number of generated keypoints, the local feature extraction algorithm is then applied only on the discovered salient regions (for more details refer to Chapter 5). [Pimenov, 2009] used saliency maps based on Itti’s model [Itti et al., 1998] to identify salient SURF keypoints. The calculated saliency values are also used to decide whether two keypoints represent a match. In [López-García et al., 2011] the authors proposed another method for generating a retina-optical saliency map by using local phase information of the input data. The saliency maps were used to filter SIFT and SURF features for the purpose of matching video sequences of a robot-navigation system. [Chen et al., 2011] also proposed a method for generating saliency maps. In this work, phase Fourier transform is used to construct the saliency map. To enable faster scene matching, SURF keypoints are extracted only from the salient image regions.

2.6 Thesis Contributions Revisited

In this chapter we discussed research efforts on automatic image annotation which exploits already labeled images. We have shown how research efforts evolved from using the World Wide Web in the early years to leveraging the rich contextual information of the current photo sharing platforms. The work in this thesis also exploit community photo to automatically annotate new images. Similar to other works [Moxley et al., 2008, Silva and Martins, 2011, Mitran et al., 2013] our approach takes advantage of the ever increasing number of geotagged photos as an additional contextual clue to assist the tag mining process. However, in contrast to these approaches, we lay special focus on the efficiency of the annotation process and the quality of the produced tags. More specifically, the novelty of the work presented in this thesis is demonstrated through solutions to the following main AIA challenges:

- **Representativeness of the dataset form which candidate annotations are extracted:** compared to the discussed ad-hoc methods [Hays and Efros, 2008, Kalantidis et al., 2011, Tolia and Avrithis, 2011, Hollenstein and Purves, 2013, Weyand et al., 2012], we propose a data crawling strategy that ensures the representativeness of the annotation resource. The representativeness is defined in terms of the spatial distribution of the collected images and the quality of the associated metadata. Additionally, we provide an efficient mechanism for fast retrieval from the dataset based on location information.
- **Resolving tag ambiguity using relatedness information:** it can be observed that the majority of works on tag disambiguation exploit tag co-occurrence counts

and apply a simple co-occurrences threshold [Begelman et al., 2006, Simpson, 2008] or the cosine measure to identify similar tags [Specia and Motta, 2007, Gemmell et al., 2008a,b]. We aim, in this thesis, to address two important aspects which are less investigated in literature on tag relatedness. First, although we use the same representation for tags as probability distributions as done in [Weinberger et al., 2008], our method deals also with statistical fluctuations in the created probability distributions and propose extension for the well-known Jensen-Shannon Divergence. Second, to best of our knowledge, this work is the first to deal with the problem of feature selection for building tag co-occurrence vectors. In this regard, we propose a solution based on the method of Laplacian score for feature selection [He et al., 2005] and demonstrate its efficiency in identifying related/similar tags.

- **Improving the accuracy and the performance of CBIR based on SURF image features:** whilst other approaches are focused on improving the performance of SIFT-based image matching [Foo and Sinha, 2007, Turcot and Lowe, 2009, Dong et al., 2012], our work is concerned with boosting the performance of SURF [Bay et al., 2008]. SURF is faster to compute than SIFT and at the same time it has comparable matching performance [Juan and Gwun, 2009]. Our approach aim at identifying the subset of SURF keypoints which contribute most to identifying similar images. Similar to [Lepetit and Fua, 2006, Ozuysal et al., 2010] we also investigate classification techniques for keypoint characterization. However, in contrast to those works where multi-label classification is used, we deal with the problem as binary classification and investigate different features for characterizing the keypoints. We also investigate the efficiency of our solution by comparing it to other approaches which use visual attention (e.g. [Pimenov, 2009]).
- **Combining contextual information for tag ranking:** while other works focus on investigating the effectiveness of location information on the produced rankings [Moxley et al., 2008, Silva and Martins, 2011, Mitran et al., 2013], thanks to Bayes' rule, our tag ranking model is scalable and can be easily extended to consider further contextual information. We use our model to combine and to investigate the effectiveness of different kinds of information (i.e., geographical, image-content, user and tag usage information) on the quality of the automatically generated annotations.

Part II

Data Preparation and Tag Disambiguation

Chapter 3

Geographical Crawling and Indexing of Community Photos

This chapter deals with the problem of creating a representative dataset of geotagged images. For this purpose, a strategy for crawling image data from Flickr based on location information is introduced. Additionally, an approach for spatially indexing the collected data using quad-tree data structure is described. Furthermore, a solution for resolving lexical problems of user-tags is presented.

3.1 Introduction

As mentioned in the first chapter, the preparation phase of our automatic annotation approach requires constructing a dataset of geotagged images with the associated metadata. In this thesis, we define global criteria to ensure the quality of geotagged image datasets. First, the dataset must have a high spatial coverage in the sense that it should contain images covering the whole world map. Second, the data should be spatially representative, i.e., the density of the collected photos should vary according to the popularity of the different places. Third, the dataset have to ensure the quality of the provided image metadata. Basically, user-tags represent the main resource for high-level descriptive image metadata. To allow a maximum advantage, the ambiguity of user-tags has to be resolved before they can be used by further applications.

In the following sections, we describe our approach to build a dataset which fulfills the mentioned criteria. First, our data crawling strategy is presented. Then, we introduce a tag cleaning approach. Finally, a spatial data indexing strategy is presented.

3.2 Geo-based Data Crawling

To fulfill the representativeness requirements, we followed a data crawling strategy based on Flickr friendship's graph and the principle of small-world [Milgram, 1967]. The proposed method is inspired from [Crandall et al., 2009], however, instead of creating a random sample of photo identifiers, we generated a sample of identifiers corresponding to users residing in various places in the world. The decision to create such a sample can be explained according to the small-world phenomena. By traversing the friendship's graphs of spatially well-distributed user set, the final set would converge to a collection of users who have taken photos in every part of the world. Consequently, crawling geotagged photos uploaded by those users will lead to a dataset with a spatial distribution that approximates the real-world case. More specifically, a seed set of Flickr users is created by randomly selecting users who are residing in different areas in the world. This set is then extended as follows: first, the friendship's graph of each user in the seed set is obtained from Flickr. After that, breadth-first search is applied on each graph to gain additional users. This process is applied recursively on the newly acquired users until a certain number of unique users is reached. Finally, for each user, the corresponding geotagged photos are crawled with the associated metadata.

During the crawling process we applied two filtering conditions on the downloaded photos. First, we used the metadata provided by Flickr to discard images with poor geographical accuracy¹. Second, since many applications require photos of acceptable resolution, photos with resolution below 320×240 pixels were also removed.

Figure 3.1 shows a plot of the geographical coordinates of a sample of 300,000 photos taken from our dataset. Each image is represented by a point in a two dimensional space of longitude on the x-axis and the latitude on the y-axis. The graphic shows how the coordinates of the crawled images can approximate the world map. Moreover, dark areas indicate densely photographed places. This conforms to several studies on Flickr (e.g. [Crandall et al., 2009]) which shows that certain places in Western Europe and the United States are most popular among photographers.

A closer look on the spatial distribution of the crawled photos is given in Figure 3.2.(b). Photos taken in Paris are represented according to their geographical coordinates in the longitude-latitude space. Dense areas correspond to places which attract photographer at most. Compared to the map of Paris shown in Figure 3.2.(a), we observe dense

¹Flickr defines 16 different accuracy levels for the geographical coordinates of a geotagged image. The highest level 16 indicates that the location is accurate at street-level, while the lowest value 1 corresponds to world-level. For our dataset, we set the minimum accuracy level for the downloaded images to city-level (value 11).

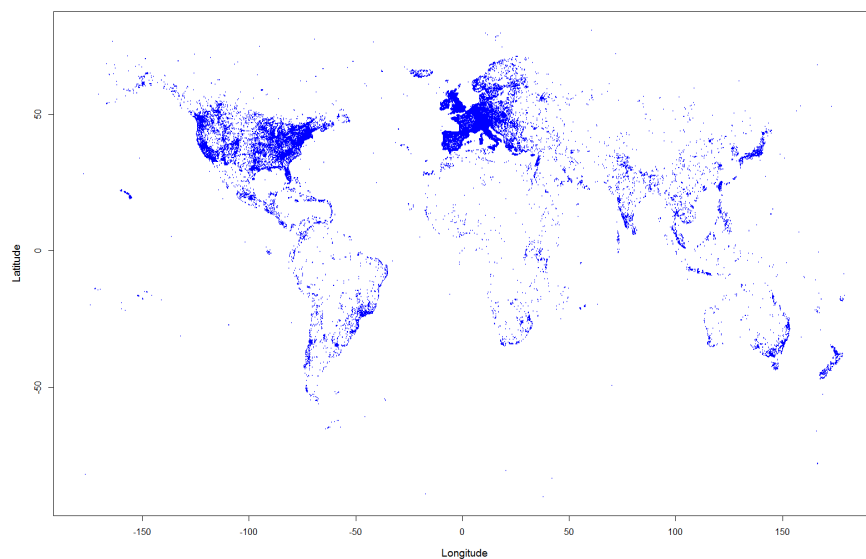
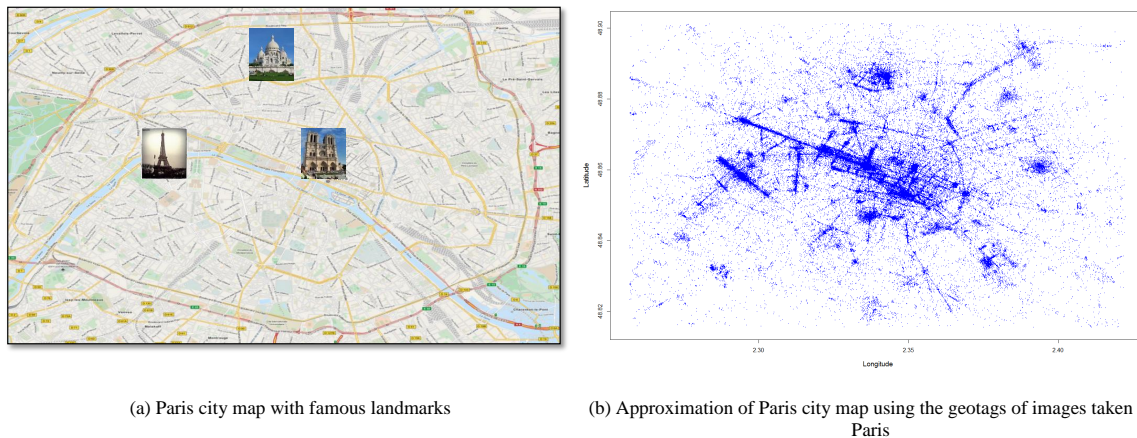


FIGURE 3.1: The geographical coordinates (latitude vs. longitude) of a sample of 300,000 images from our dataset

amounts of photos around touristic attractions, such as the city center, around Eiffel Tower and along the Seine River.



(a) Paris city map with famous landmarks

(b) Approximation of Paris city map using the geotags of images taken in Paris

FIGURE 3.2: Photo density in the city of Paris according to our dataset

We also compared our dataset to the one presented in [Crandall et al., 2009] in terms of the most photographed cities. The authors analyzed a collection of 35 million Flickr photos and demonstrated how cities, such as New York, London, San Francisco and Paris belong to the most photographed cities and in the provided order. The same was observed in our dataset (Figure 3.3).

The final dataset contains a collection of 14.1 million photos with associated metadata. The photos were contributed by more than 200,000 users in the time period from

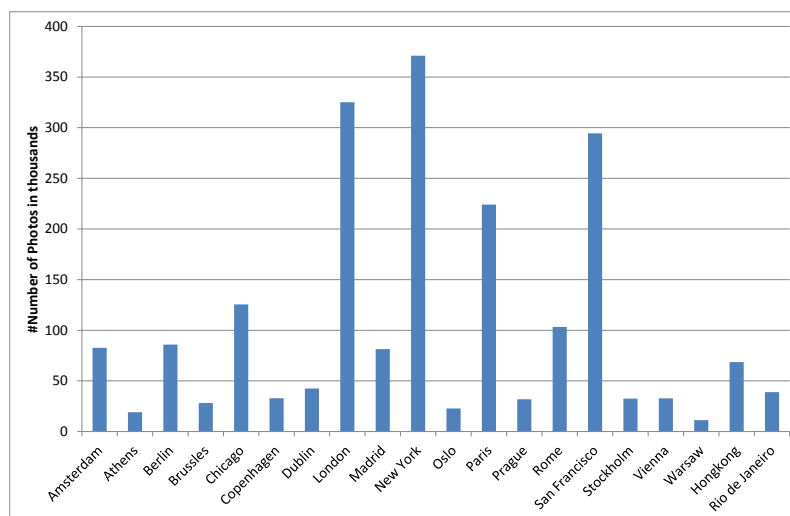


FIGURE 3.3: The number of images per city according to our dataset

14/5/2000 until 01/04/2012. For each photo the following metadata data are provided: the photo identifier, the identifier of Flickr’s user who uploaded the photo, the photo title (if existing), the list of associated user-tags, the location information represented by the longitude and latitude, the accuracy level of the location information as defined by Flickr, the date of photo capture, the date when the photo was uploaded to Flickr’s server, and the information needed to construct the photo URL [Flickr, 2014b].

3.3 Tag Cleaning

In order to reduce the noisiness of the collected user-tags, we applied simple and effective tag cleaning procedures which mainly focus on addressing problems related to the lexical variations of the tags (In Chapter 4 a more sophisticated approach for dealing with tag ambiguity is presented).

Tag Filtering

Before dealing with lexical problems of user-tags, a filtering step is applied to remove tags corresponding to stop words. For this purpose, we manually identified a list of stop words. This includes non-descriptive tags, such as the words *photo*, *picture* and the like. Another kind of stop words includes tags referring to technical terms, such as camera types and camera settings (e.g. canon, longexposure, d40x). Furthermore, tags specific to Flickr, e.g. *flickr.com*, *platinumheartaward*, etc. and other tags referring to dates, web services or photo editing programs are also added to stop word list. Moreover, in

a similar manner to other works on tag cleaning [Specia and Motta, 2007, Mika, 2007, Cantador et al., 2008, García-Silva et al., 2008, Giannakidou et al., 2008], we apply an additional refinement step which filters out tags with low frequency. Tags that are used by a small number of users are usually noisy since they contain user-specific information. Accordingly, we eliminated tags which were used by less than 5 users from the dataset. The final dataset contains 415,369 unique tags with a total occurrence of 100,791,616 and an average of seven tags per photo.

Lexical Cleaning

User-tags suffer from problems, such as misspelling and lexical variations (refer to Section 2.5.2). The latter problem arises when users adopt different lexical forms to express the same term. For example, different users may annotate photos taken in New York with "newyork", "new-york" or "new york". To deal with such problems, we developed an automatic approach based on the correction suggestions provided by Yahoo search engine [Yahoo, 2014]. Each tag t is first used to query Yahoo. In the case where t is misspelled or if it consists of combined words then Yahoo provides proposals for related search terms (see Figure 3.4).

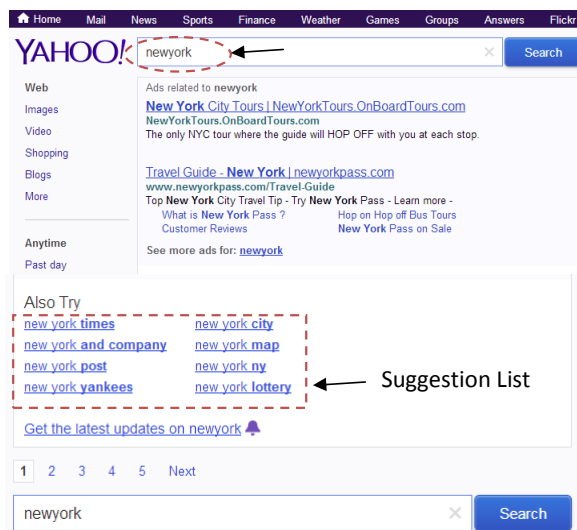


FIGURE 3.4: Search results for the term "newyork" according to Yahoo search engine with suggestions for related search terms

Indicate with $S = \{S_1, \dots, S_n\}$ the set of all suggestions produced for a given query word t . In turn, each suggestion $S_i \in S$ is a an ordered sequence of words denoted as $S_i = (w_1, \dots, w_k)$. Next, the set of unique words is obtained from the union of all suggestion sequences, i.e., $W = \cup_i S_i = \{w_1, \dots, w_m\}$. After that, for each word $w_j \in W$

its number of occurrences, $C(w_j)$, over all suggestions sets is computed. Finally, the set Corr_t that contains the terms which will be used to correct the input word t is determined as follows:

$$\text{Corr}_t = \{w_j | w_j \in W \wedge C(w_j) \geq \theta\} \quad (3.1)$$

In Equation 3.1, θ is a lower bound for word occurrence and can be set experimentally. By conducting several experiments, we recognized that reliable results can be achieved by setting $\theta = 0.8 \times |S|$. That means, in order for a word to belong to the correction set, it must appear at least in 80% of the suggestions $S_i \in S$.

After the correction set has been acquired, a final correction term is created by identifying the right order of the correction terms. To do that, we apply a simple technique which determines the order of the words according to their order in the majority of the suggestion sequences $S_i \in S$. That is, for two words $w_1, w_2 \in \text{Corr}_t$, if w_1 occurs before w_2 in the majority of the suggestions, then w_1 must proceed w_2 in the final correction term.

In Figure 3.4, for example, a correction set for the input tag *newyork* can be built out of the most frequent words in the corresponding Yahoo suggestions. The means, the correction set is given as: $\text{Corr}_{newyork} = \{new, york\}$. As the word *new* occurs before the word *york* in all suggestions, the same order must be followed in the final correction term, i.e., the term *newyork* have to be replaced with new term *new york*. Table 3.1 shows further examples of misspelled as well as tags consisting of multiple words which have been corrected usign the described method.

Original Tag	Corrected Tag
abandoned-building	abandoned buildings
abrahamlincoln	abraham lincoln
portlandmusic	portland music
greatsanddunes nationalpark	great sand dunes national park
sanpedrolalaguna	san pedro la laguna
enviroment	environment
freind	friend

TABLE 3.1: Sample user-tags acquired from Flickr (first column) which have been automatically corrected using the presented tag cleaning algorithm (second column)

3.4 Indexing using Quad-tree

In Chapter 1, we have seen that the very first step of the tag mining phase is to search for images taken in the same location as the input image. Doing this in a naive way requires calculating the geographical distance based on the geographical coordinates of the input image and those of each image in the dataset. This process is time consuming, especially when the search space is huge, such as the one in our case. Therefore, we developed a method which allows fast retrieval of the geographical neighbors by spatially indexing the dataset using the quad-tree data structure [Finkel and Bentley, 1974, Samet, 1984].

Quad-tree is a hierarchical data structure which is based on the principle of recursive decomposition. It is widely used for indexing two dimensional data such as geographical coordinates [Samet, 1990, Gahegan, 1989]. For this purpose, data points are recursively divided into four regions until a stopping condition is met. This condition is defined in terms of the maximum allowed capacity of a single quad-tree region. When faced to a large number of data points (e.g., our data set contains 14.1 million data points) a direct application of the quad-tree algorithm become impractical. Indeed, indexing large number of data points and by using a relatively low maximum capacity threshold leads to immense memory requirements. This is due to the high recursion depth². To address this problem, we propose a method for distributing the computation of the quad-tree. Initially, we divided the world map into tiles. A tile is created only if there are photos in the dataset taken in the area specified by that tile. After that, tiles with a high photo density are further divided into sub-tiles. This process is repeated as long as the number of photos in the tile exceeds a predefined upper bound (Figure 3.5). In a next step, the quad-tree algorithm is applied on each tile (Figure 3.6). The final index consists of the coordinates of each tile and each quad-tree region. To allow flexible retrieval, the index also keeps track of the neighborhood information of each quad-tree region. This can be useful when a specific quad-tree region is sparse. In this case, additional data points can be efficiently retrieved by extending the result set to data points of neighboring quad-tree regions.

We applied the described approach on our dataset using initial 10×10 tiles. The tiles are then shrunk according to the coordinates of the contained data points. Next, tiles containing more than 300,000 photos were further divided. Figure 3.5 shows the results of this phase. The produced tiles show an approximation of the world continents. Additionally, we can see that tiles corresponding to areas of high photo density are further divided into sub-tiles shown as smaller rectangles inside the corresponding tiles.

²On a machine with 8GB RAM and using Matlab, the maximum recursion limit of 500 was reached with a relatively small set of 300,000 data points and a maximum capacity of 2,000 data points per quad-tree region

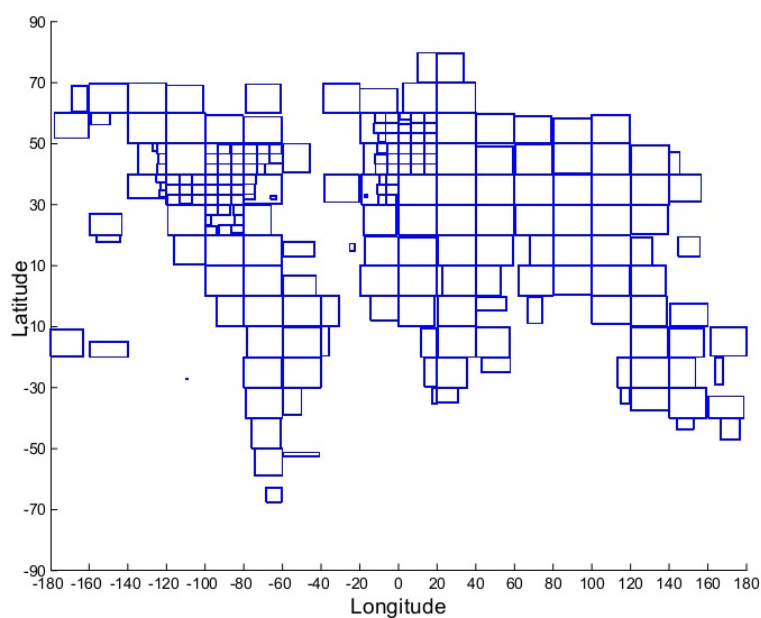


FIGURE 3.5: World map divided into tiles according to the photo density as given by our dataset. Dense tiles are further divided into sub-tiles

For all tiles, we applied the quad-tree algorithm and set the maximum capacity for each region to 800 photos. The corresponding quad-tree regions are shown in Figure 3.6.

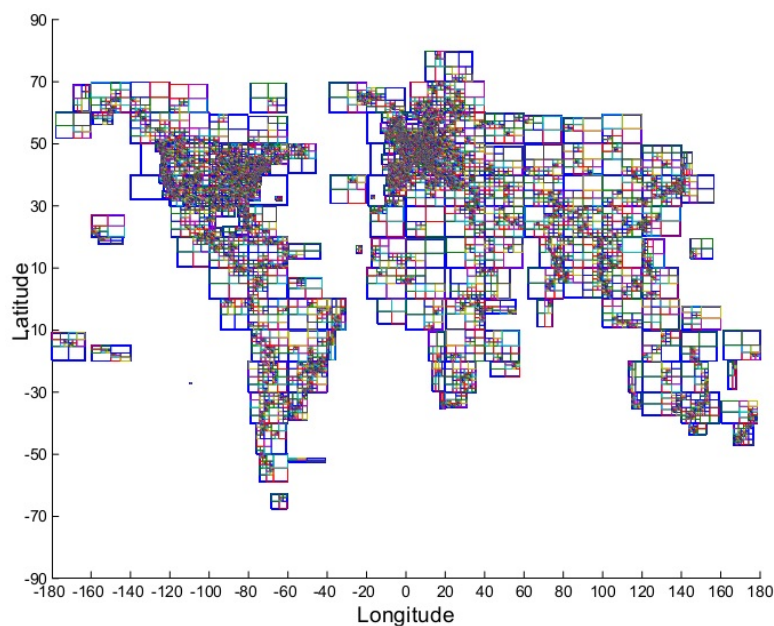


FIGURE 3.6: Quad-tree regions for our dataset. The quad-tree algorithm is applied on each tile separately to allow efficient computation

We collected statistics about the generated tiles and the corresponding quad-trees. Indexing the collection of 14.1 million geographical coordinates resulted in 215 tiles with an average of 312 quad-tree region per tile. Each tile contains about 65,500 data points

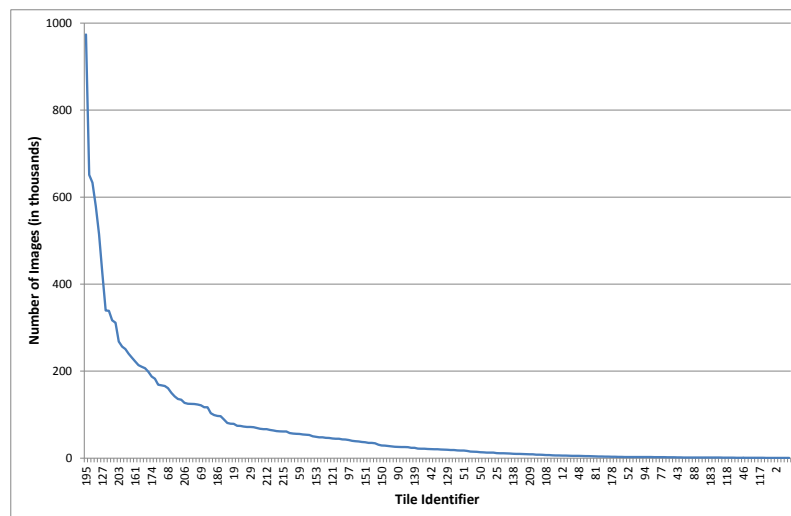


FIGURE 3.7: The number of photos per tile according to our dataset. The x-axis correspond to the tile identifier and the y-axis gives the total number of photos per tile

(i.e., longitude-latitude pairs) on average, however, with a large standard deviation of about 122,000. This is due to the sharp differences in the density of photos from place to place. In fact, there are very few places in the world which are frequently photographed, while quite a large number of places are even less photographed (Figure 3.7).

3.5 Qualitative Insight

To get an impression about the nature of our dataset, we provide a multifaceted visualization of it using the image browsing tool Folkioneer [Mousselly Sergieh et al., 2014a]. Folkioneer applies agglomerative clustering using the CURE algorithm [Guha et al., 1998] to spatially cluster the images according to their geographical coordinates. The output of the clustering algorithm is a dendrogram of the complete set of data points in the dataset. To build the geographical clusters, a cut-off threshold is applied on the dendrogram. The clusters can also be built at different levels of geographical granularity by using different cut-off thresholds.

Figure 3.8 shows a layout of geographical clusters corresponding to geotagged images from our dataset. The clusters are shown in different colors (red, green, violet and yellow) while the light brown color indicates that there is no images in the corresponding area. For instance, compared to North Africa and Eastern Europe, we can see that Western Europe has a larger number of clusters which indicates higher photo density. By zooming in, a new cut-off threshold is applied on the dendrogram and the new clusters are visualized on the map (Figure 3.9).



FIGURE 3.8: Geographical clusters at zooming level 5



FIGURE 3.9: Geographical clusters by zooming in to level 6

In order to take a closer look on the images in the dataset, we used Folkioner to retrieve images corresponding to the area of Paris. Folkioner applies a further clustering step on the images of a given geographical cluster based on the visual similarity. Figure 3.10 shows sample visual clusters containing photos that are typical to the city of Paris, such as Eiffel Tower and Arc de Triomphe. Finally, to get an insight into the collected user-tags, we used Folkioner to visualize tags which are used to annotate images taken in Paris. The tags are represented in the form of a cloud of tag clouds. Figure 3.11 shows the tags which are typical to images taken in Paris with similar/related tags represented as single tag clouds and using a distinctive color.

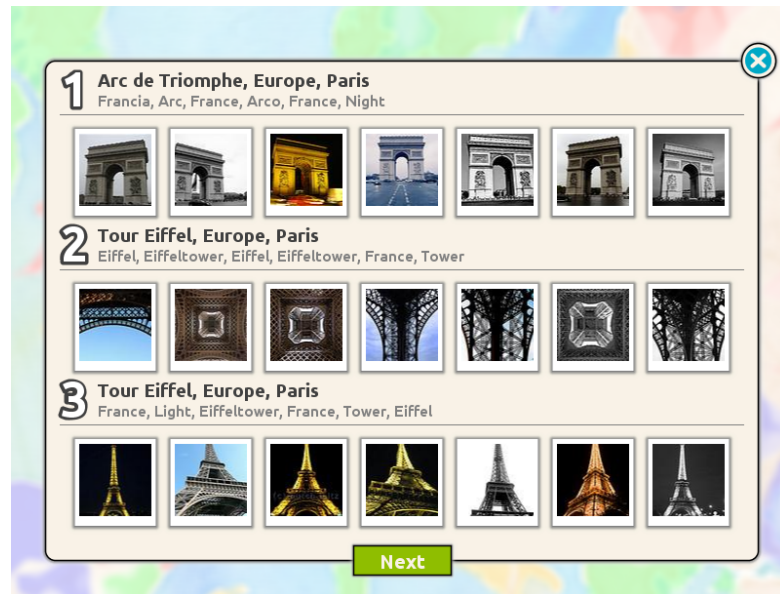


FIGURE 3.10: Sample images from our dataset which are located in Paris

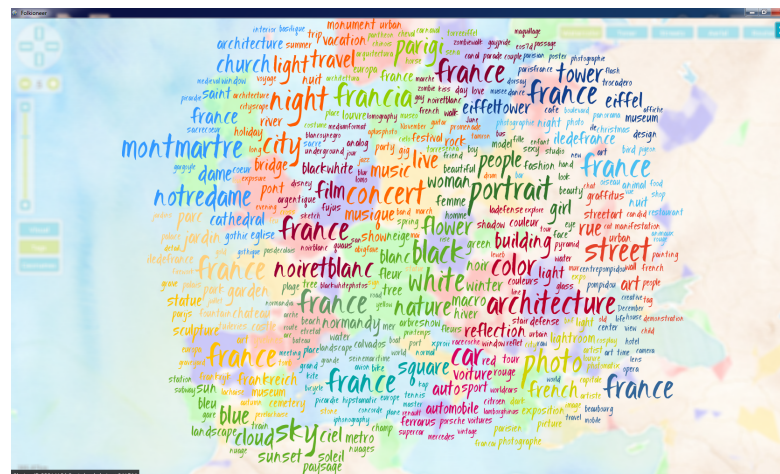


FIGURE 3.11: Tags from our dataset corresponding to images taken in Paris

3.6 Summary

This chapter presented an approach for crawling and indexing geotagged images from Flickr. The approach was used in order to create a dataset of 14.1 million images with the corresponding metadata. The dataset is used by our image annotation approach as a resource for generating tag proposals for new unlabeled geotagged images. The spatial representativeness of the data was ensured through a crawling strategy based on Flickr friendship's graph and the principle of small-world. Additionally, we improved the quality of the associated user-tags using a simple technique for tag cleaning which exploits the feature of Yahoo search term suggestions. To allow efficient retrieval from the dataset, a spatial index is created using an improved implementation of the quad-tree algorithm.

Chapter 4

Mining Tag Relatedness for Resolving Tag Ambiguity

This chapter discusses the effect of noisy user-tags on search-based automatic image annotation and proposes an approach for identifying related tags in folksonomies as a pre-step to resolve their ambiguity. The proposed tag relatedness approach aims at addressing the problem of creating tag representation through an approach for feature selection based on the Laplacian score method. Furthermore, it presents a new measure for calculating the distance between tag probability distribution based on the well-known Jensen-Shannon Divergence (JSD).

4.1 Introduction

As discussed in Chapter 2, tag noisiness represents a unique challenge for search-based AIA approaches. [Gemmell et al., 2009] analyzed the impact of ambiguous and redundant tags on approaches for tag recommendation in folksonomies. On one hand, noisy tags can be misleading and may bother the users with unwanted recommendations. On the other hand, noisy tags make it difficult to judge the quality of recommendation systems. In general, the quality is evaluated according to the ability of the system to predict tags in a holdout set. Accordingly, recommending noisy tags can result in underestimating or overestimating the performance of the recommendation system.

For a better insight into the problem, assume that a given search-based AIA system is used to annotate a photo taken in London as shown in Figure 4.1. Assume also that the system has generated for the input image a ranked list of tags as illustrated in the left-hand side of the figure. At first glance, the collection of the mined tags seems to

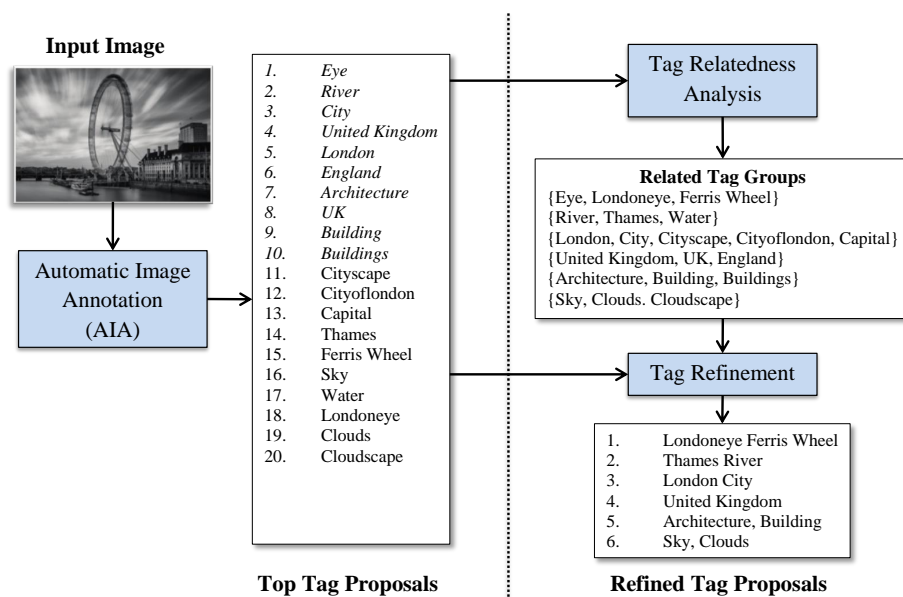


FIGURE 4.1: Tag proposals for a photo of London Eye

fit the contents of the input image. However, it is also evident that the list contains redundant tags. Tags like "UK" and "United Kingdom" refer to the same thing and the same applies to "clouds" and "cloudscape". Moreover, the tag "eye" is ambiguous and could be considered irrelevant by people who do not know about the famous Ferris wheel, called London Eye. Furthermore, it is typical to limit the number of the proposed tags to the best ranked ones, therefore, retaining redundant tags results in discarding useful information. For example, the list of the top 10 tags (shown in italic in Figure 4.1) suffers from redundancy, e.g., the location tags "United Kingdom", "England" and "UK". Additionally, limiting the number of proposed tags to the top 10 will make relevant tags, such as "Thames" and "Ferris Wheel" disappear from the top tag list.

The presented example demonstrates the need for identifying ambiguous and redundant tags in order to improve the performance of automatic image annotation. As we discussed in Chapter 2, resolving tag ambiguity have been considered by several research efforts. The idea is to mine similarity or relatedness information among the tags and use this information to identify their real meaning. For further explanation, refer to the right-hand side of Figure 4.1. If relatedness information is available, redundant tags like "UK" and "United Kingdom" as well as synonyms like "Eye" and "Ferris Wheel" can be determined. Consequently, we can use this information to improve the quality of tag proposals by removing redundant tags (e.g. the tags "UK", "Buildings") and combining other related ones (e.g. "Thames" and "River").

4.2 Tag Relatedness Approach

In Section 2.5.2, we presented different types of representation for tags, namely using the *user-context*, the *tag-context* or the *resource-context*. In this thesis, we mainly deal with folksonomies corresponding to community photos. Such folksonomies are narrow since the provided tags are mainly contributed by the uploaders of the images and the level of user interaction is low. Accordingly, the user-context representation has a limited value regarding the identification of related tags. The same thing applies for the resource-context. Indeed, there are two reasons which make it unsuitable for identifying related tags in image folksonomies. First, in such folksonomies it is unlikely that the same tag will be applied multiple times to describe the same photo. Second, in contrast to textual resources, where further occurrences of the tags can be acquired by analyzing the associated text, images do not have such textual context. As opposed to the discussed representations, the tag-context provides rich information about the pattern of tag usage in folksonomies. Hence, the tag relatedness approach proposed in this thesis is based on the latter representation.

4.2.1 Approach Overview

The workflow of our tag relatedness approach is illustrated in Figure 4.2. We start from a folksonomy represented as a bipartite graph of tags and resources as nodes (user information is not considered since we are only interested in the tag-context). The edges indicate that a tag has been used to annotate a resource. In the tag-context representation (refer to Section 2.5.2) each tag is represented in terms of its co-occurrence with other tags in the folksonomy. Since it is impractical to calculate the co-occurrences with the complete set of unique tags in the folksonomy, we apply a feature selection algorithm to identify a subset of important tags (features) and restrict the tag-context representation to those features. To compute the feature set, we adopt a feature selection approach based on the Laplacian score (LS) method [He et al., 2005]. After that, for each unique tag a probability distribution is created based on the co-occurrence of that tag with the elements of the feature set. Finally, the relatedness between two tags is determined according to the distance between their probability distributions. To calculate this distance, we propose a new metric, called Adapted Jensen-Shannon Divergence (AJSD). As the name suggests, AJSD is based on the well-known Jensen-Shannon Divergence (JSD) [Manning and Schütze, 1999], however, it is characterized by its ability to deal with statistical fluctuations in the generated probability distributions.

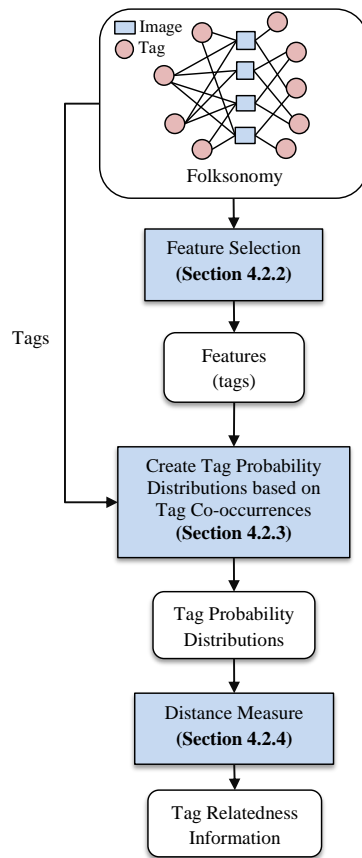


FIGURE 4.2: The workflow of the proposed tag relatedness approach

4.2.2 Feature Selection for Tag Relatedness

First, let us recall the notation of a folksonomy $F = \{T, U, R, A\}$ where T is the tag set, U the set of users, R is the set of resources and $A \in U \times T \times R$ is an assignment relation. Identifying related tags in a folksonomy is an all-pairs-similarity-search problem (APSS) [Bayardo et al., 2007] since each tag has to be compared to all other tags in the folksonomy. Given the set of $|T|$ tags and by considering that each tag is represented by a d dimensional vector, the naive approach will compute the similarity between all tag pairs in $\mathcal{O}(|T|^2 \cdot |d|)$ time. In the case of tag-context, we have $d = |T|$, thus, the algorithm will have $\mathcal{O}(|T|^3)$ complexity. For large folksonomies, performing such computations is impractical. However, the computational cost can be reduced if the tags are represented in reduced vector space, i.e., $\mathbb{R}^{\mathcal{F}}$ where $\mathcal{F} \subset T$ and $|\mathcal{F}| \ll |T|$. Of course, in this case, the challenge is to provide a feature selection approach which can maintain, if not improve, the quality of the tag relatedness measure.

A simple approach to build the feature set \mathcal{F} , is to select a subset of the most frequent tags in the folksonomy (e.g. [Weinberger et al., 2008, Cattuto et al., 2008]). This technique has some effectiveness, but the main issue is that the most frequent tags may

have almost uniform co-occurrence patterns with most other tags in the folksonomy. In this case, all tags would be considered related to each other. Hence, a more sophisticated approach for identifying \mathcal{F} is required.

Identifying discriminative features from data has been investigated thoroughly by researchers, especially in the domains of pattern recognition and computer vision (for surveys refer to [Guyon and Elisseeff, 2003, Molina et al., 2002]). For our particular case, the feature selection process must be applied in an unsupervised way since we are dealing with unlabeled data. For this purpose, we adopt the method of Laplacian score feature selection (LS) proposed by [He et al., 2005] to identify the subset of "important" tags in the folksonomy. The Laplacian score (LS) technique for feature selection is based on Laplacian Eigenmaps [Belkin and Niyogi, 2003] and Locality Preserving Projection [He et al., 2005] techniques. These techniques allow the representation of a dataset, whose data points are characterized by a high dimensionality, by means of a lower dimensional representation, implicitly based on a low dimensional sub manifold of the whole space. These techniques postulate that such a manifold exist and that it can be represented efficiently in terms of a small subset of dimensions (these will be the selected features).

This schema fits the problem at hand: user-tags are the points of our dataset; they are represented initially by high-dimensional vectors (the tag co-occurrence vectors). The results of the application of the method described hereafter confirm ex-post the soundness of the assumptions.

To compute the LS, the data points are first organized in a weighted undirected graph, in which nodes correspond to data points and an edge is drawn between two nodes if they are close to one another according to some predefined similarity measure (such as the cosine measure). The edges are weighted proportionally to the similarity between the connected data points. The Laplacian (matrix) L of such a graph is a square matrix defined by the difference of the degree matrix D and the adjacency matrix S (see below) of the graph. Intuitively, the Laplacian matrix is a discrete analog of the Laplacian operator in multi-variable calculus and serves a similar purpose by measuring to what extent a graph differs at one vertex from its values at nearby vertices. Thanks to such a measure, one can define the Laplacian score for each individual vertex (the less it differs from the neighbors the higher its score) and consequently choose those points who turn out to have the highest scores as representative features.

Formally, the LS feature selection algorithm can be considered as a minimization problem with the following objective function [He et al., 2005]:

$$\mathcal{L}(f) = \frac{\sum_{ij}(f_i - f_j)^2 S_{ij}}{\text{Var}(f)} \quad (4.1)$$

In Equation 4.1, f_i and f_j correspond to the values of the feature f at the data points i and j respectively, while S_{ij} is the similarity between them. $\text{Var}(f)$ is the variance of the of feature f . The minimization of the objective function implies preferring features of larger variances. This conforms to the intuition that features with higher variance are expected to have more expressive power.

The feature selection algorithm and an estimation for the solution of the objective function are summarized in the following steps (more details can be found in [He et al., 2005]):

1. For the set of n data points a k nearest neighbor (KNN) graph is constructed. In that graph, an edge between two data points x_i and x_j is drawn if the points are close to each other, i.e., if x_i belongs to the set of k nearest neighbors of x_j and vice versa.
2. The edges between close nodes are weighted according to a similarity function. To calculate the similarity, there are several options, such as the cosine (Equation 4.6) or the Gaussian similarity which is defined as:

$$S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{2u}} \quad (4.2)$$

where x_i and x_j are two data points and u is a free parameter that can be determined experimentally. Next, pairwise similarities of the data points are combined in a similarity (adjacency) matrix S .

3. For a feature f , defined as a vector over the data points, let:

$$\tilde{f} = f - \frac{f^T D \mathbb{1}}{\mathbb{1}^T D \mathbb{1}} \mathbb{1} \quad (4.3)$$

where $\mathbb{1} = [1 \dots 1]^T$ is the identity matrix and $D = \text{diag}(S\mathbb{1})$ is the diagonal degree matrix, in which each entry d_{ii} corresponds to the sum of the entries of the column i in the similarity matrix S .

4. Let $L = D - S$ be the Laplacian matrix of the similarity graph [Chung, 1997]. The Laplacian score of the feature f is then computed as:

$$\mathcal{L}(f) = \frac{\tilde{f}^T L \tilde{f}}{\tilde{f}^T D \tilde{f}} \quad (4.4)$$

5. The final *feature set* \mathcal{F} contains those features with a Laplacian score greater than a predefined threshold θ :

$$\mathcal{F} = \{ f \mid \mathcal{L}(f) > \theta \} \quad (4.5)$$

In our case, the data points as well as the features correspond to the tags of the folksonomy. Each feature and each data point is represented by a vector using the tag-context representation. If the complete set of data points and features are considered by the LS method, then the representation of a data point and a feature corresponding to the same tag is identical. That means for a feature f and a data point x corresponding to a tag t they are represented as, $x = f = v(t) \in \mathbb{R}^T$, where $v(t)$ is the tag-context representation of the tag t . However, to reduce the time needed by the LS algorithm, it can also be applied on a sample of the data points or only a subset of the tags are considered as features. In both cases, the data points and the features corresponding to the same tag will have varying tag-context representations. This is because they will be represented using different dimensions (subsets of the tags in the folksonomy). Accordingly, by assuming that a subset of the tags $T_k \subset T$ is used to represent the data points, then the set of all data points, X , is given by:

$$X = \{ x \mid x = v(t) \in \mathbb{R}^{T_k} \wedge t \in T \}$$

The same applies to the set of all features \mathcal{F} . Given that each feature is represented as a vector over a subset of tags $T_l \subset T$, then:

$$\mathcal{F} = \{ f \mid f = v(t) \in \mathbb{R}^{T_l} \wedge t \in T \}$$

Illustrative Example

To clarify how the Laplacian score algorithm can be applied to select important features in a folksonomy, consider the tag co-occurrence matrix shown in Figure 4.3. The column and the rows of the matrix correspond to the tags while the entries correspond to the co-occurrence counts of the tag pairs as observed in the folksonomy. The co-occurrence of a tag with itself is set to zero. In this example the tags "France" and "Paris" occur most. Furthermore, both tags show uniform occurrence patterns with the other tags.

Data points as well as the feature can be derived directly from the rows and columns of the co-occurrence matrix, respectively. For example the data point corresponding to the tag "France" is given by $x_{\text{France}} = (0, 30, 30, 30, 30, 30)$, while the feature vector corresponding to the tag "Tower" is given by $f_{\text{Tower}} = (30, 20, 0, 20, 5, 5)^T$. In the next step, we create a weighted nearest neighbor graph from the data points (step 1 and 2

	<i>France</i>	<i>Paris</i>	<i>Tower</i>	<i>Eiffel</i>	<i>Sky</i>	<i>City</i>
<i>France</i>	0	30	30	30	30	30
<i>Paris</i>	30	0	20	20	20	20
<i>Tower</i>	30	20	0	20	5	5
<i>Eiffel</i>	30	20	20	0	10	10
<i>Sky</i>	30	20	5	10	0	5
<i>City</i>	30	20	5	10	5	0

FIGURE 4.3: A sample tag co-occurrence matrix

of the algorithm). Due to the small number of data points, we use a complete graph (instead of *KNN* graph) and chose the cosine similarity (Equation 4.6) to weight the edges (Figure 4.4).

$$\text{sim}(t_1, t_2) = \text{cosine}(v(t_1), v(t_2)) = \frac{v(t_1) \cdot v(t_2)}{\|v(t_1)\| \cdot \|v(t_2)\|} \quad (4.6)$$

For instance, consider the two data points $x_{\text{France}} = (0, 30, 30, 30, 30, 30)$, $x_{\text{Eiffel}} = (30, 20, 20, 0, 10, 10)$ corresponding to the tags "France" and "Eiffel", respectively. The weight of the edge between the corresponding nodes is calculated as:

$$\text{sim}(x_{\text{France}}, x_{\text{Eiffel}}) = \frac{0 \times 30 + 30 \times 20 + 30 \times 20 + 30 \times 0 + 30 \times 10 + 30 \times 10}{\sqrt{0^2 + 30^2 + 30^2 + 30^2 + 30^2 + 30^2} \sqrt{30^2 + 20^2 + 20^2 + 0^2 + 10^2 + 10^2}} = 0.62$$

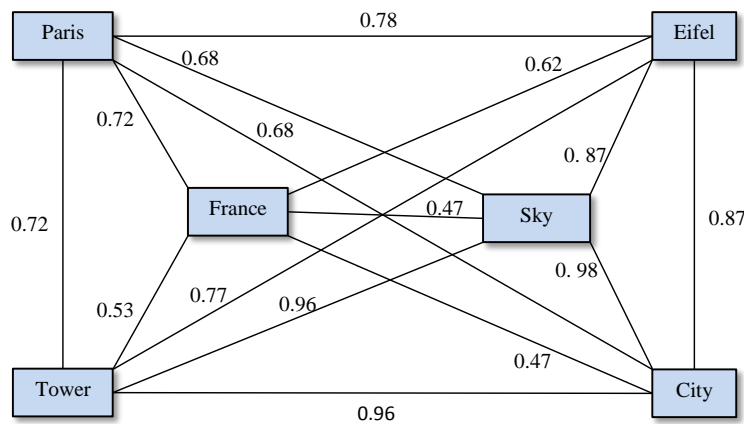


FIGURE 4.4: Similarity graph for the data points corresponding to the rows of the matrix shown in Figure 4.3. The nodes corresponds to the tags with the edges weighted according to the cosine similarity

Next, the nearest neighbor graph is mapped into a similarity matrix S . In the similarity matrix, an entry S_{ij} indicates the cosine similarity between the data points x_i and x_j (Figure 4.5). From S we compute the diagonal matrix D . For instance, the entry $(1, 1)$ of D is calculated as follows:

$$D_{11} = \sum_i S_{i1} = 0 + 0.72 + 0.53 + 0.62 + 0.47 + 0.47 = 2.81$$

Finally, the Laplacian of the graph, i.e., the matrix L , is calculated by subtracting the similarity matrix from the diagonal matrix D . For example the entry $(1, 6)$ of L is calculated as:

$$L_{16} = D_{16} - S_{16} = 0 - 0.47 = -0.47$$

$$S = \begin{pmatrix} 0 & 0.72 & 0.53 & 0.62 & 0.47 & 0.47 \\ 0.72 & 0 & 0.72 & 0.78 & 0.68 & 0.68 \\ 0.53 & 0.72 & 0 & 0.77 & 0.96 & 0.96 \\ 0.62 & 0.78 & 0.77 & 0 & 0.87 & 0.87 \\ 0.47 & 0.68 & 0.96 & 0.87 & 0 & 0.98 \\ 0.47 & 0.68 & 0.96 & 0.87 & 0.98 & 0 \end{pmatrix} \quad D = \begin{pmatrix} 2.81 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3.58 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.93 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.91 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.97 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.97 \end{pmatrix}$$

$$L = D - S = \begin{pmatrix} 2.81 & -0.72 & -0.53 & -0.62 & -0.47 & -0.47 \\ -0.72 & 3.58 & -0.72 & -0.78 & -0.68 & -0.68 \\ -0.53 & -0.72 & 3.93 & -0.77 & -0.96 & -0.96 \\ -0.62 & -0.78 & -0.77 & 3.91 & -0.87 & -0.87 \\ -0.47 & -0.68 & -0.96 & -0.87 & 3.97 & -0.98 \\ -0.47 & -0.68 & -0.96 & -0.87 & -0.98 & 3.97 \end{pmatrix}$$

FIGURE 4.5: The similarity matrix S , the diagonal matrix D and the Laplacian matrix L as generated from the nearest neighbor graph of Figure 4.4

Now, we have all information which enables us to calculate the Laplacian score for the features (tags) of our example according to Equation 4.4. Table 4.1 shows the features and the corresponding LS scores in increasing order of importance. As we can see, the features "City" and "Sky" are considered more important by the LS algorithm than "France" and "Paris". This is because, the tags "Paris" and "France" have uniform co-occurrence patterns with all other tags. Consequently, their influence on identifying groups of related data points is negligible or even biased.

It is important to mention that the presented example is not representative enough; however, it gives an idea about the manner of how the Laplacian score algorithm can be applied to discover important tags in folksonomies. Furthermore, it shows a main

Feature	Laplacian Score
f_{Sky}	-0.07
f_{City}	-0.07
f_{Tower}	-0.09
f_{France}	-0.14
f_{Eiffel}	-0.16
f_{Paris}	-0.23

TABLE 4.1: The feature vectors ordered according to their importance (Laplacian score) from most to least important

characteristic of the LS algorithm which is its ability to determine the importance of the tags independently of their frequency of occurrence as well as to discover features of uniform co-occurrence patterns and reducing their importance.

4.2.3 Tag Probability Distribution

In this processing phase, each tag in the folksonomy is given a representation in terms of an empirical probability distribution. For this purpose, we quantify the co-occurrences of a given tag with each of the elements of the feature set. Recall the notation of the folksonomy $F = \{T, U, R, A\}$ and let $\mathfrak{R} : T \rightarrow \wp(R)$ be a function from the set of tags to the power set of the resource set, that maps a given *tag* to the *set* of resources which are annotated with it. That means, for a tag $t \in T$ we have:

$$\mathfrak{R}(t) = \{ r \mid r \in R \wedge \exists u \in U \wedge \exists (u, t, r) \in A \} \quad (4.7)$$

The measure of co-occurrence of two tags $t_i, t_j \in T$ can be defined by the function C :

$$C(t_i, t_j) = |\mathfrak{R}(t_i) \cap \mathfrak{R}(t_j)| \quad (4.8)$$

Equation (4.8) means that the measure $C(t_i, t_j)$ of co-occurrence of two tags corresponds to the number of resources which are annotated by *both* of them.

To create an empirical probability distribution for a tag t , the co-occurrences of t with each feature $f \in \mathcal{F}$ are counted so as to obtain a histogram in the variable f . Then, by normalizing this histogram, with the total number of co-occurrences of t with the elements of the set \mathcal{F} , a vector representing the empirical co-occurrence probability distribution $P(f|t)$ for the tag t with the elements $f \in \mathcal{F}$ is obtained:

$$P(f|t) = \frac{C(t, f)}{\sum_{f \in \mathcal{F}} C(t, f)} \quad (4.9)$$

where C is the tag-to-tag co-occurrence function given in equation (4.8). Each entry f of the vector $P(f|t)$ corresponds to the set of unique tags in the folksonomy which have been designated as features in the previous phase – the *feature tags* – while the value $P(f|t)$ of each entry corresponds to the measure of normalized co-occurrence of t with the *feature tag* associated with that entry. The empirical probability distribution of the tag t over the complete set of features \mathcal{F} can be denoted in short by $P(\mathcal{F}|t)$.

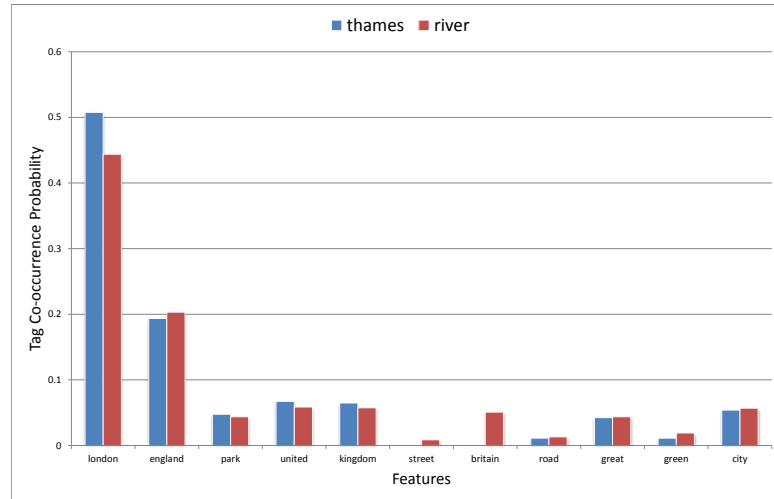


FIGURE 4.6: Empirical probability distributions of two tags "river" and "thames". Each distribution consists of several histogram channels corresponding to the elements of a feature set (x-axis). The value of a histogram channel is given by the normalized tag co-occurrence

Figure 4.6 shows sample segments of the empirical probability distributions corresponding to the tags "river" and "thames" which are found in an image folksonomy corresponding to the city of London. The x-axis corresponds to the elements of the feature set, which in this example consists of a subset of the most frequent tags in the folksonomy. For a given feature (e.g. "london") and a given tag (e.g. "river") the value of the corresponding histogram channel (y-axis) corresponds to their normalized co-occurrence as defined in Equation 4.9. Note, that the two tags "thames" and "river" show similar co-occurrence patterns with the elements of the feature set.

4.2.4 Distance Measure

At this point of the procedure, in order to determine if two tags are related, the distance between their corresponding empirical co-occurrence probability distributions must be computed. The Jensen-Shannon Divergence (JSD) [Manning and Schütze, 1999] is a widely used metrics which has shown to outperform other measures [Ljubešić et al.,

2008]. It is based on Kullback-Leibler Divergence (KL) [Kullback and Leibler, 1951], however, it is symmetric and has always a finite value.

Since the presented tag probability distributions are created from samples (ideally drawn from the true distribution), and are necessarily affected by statistical fluctuations, we propose an extension of the standard JSD measure, called Adapted Jensen-Shannon Divergence (AJSD), based on a Maximum Likelihood (ML) estimate of the JSD which both takes into account fluctuations and provides a measure of the statistical error of the results.

Before introducing the new metric, we review the KL and JSD approaches to calculate the distance between probability distribution. Let us consider two tags $t_1, t_2 \in T$ and the corresponding empirical co-occurrence probability distributions $P(\mathcal{F} | t_1)$ and $P(\mathcal{F} | t_2)$ over the feature set $\mathcal{F} = \{f_1, \dots, f_m\}$. We can simplify the notation by using $P(\mathcal{F}) \equiv P(\mathcal{F} | t_1)$ and $Q(\mathcal{F}) \equiv P(\mathcal{F} | t_2)$. Additionally, the values of P and Q at a specific feature $f_k \in \mathcal{F}$, will hereafter be represented simply by $P(f_k)$ and $Q(f_k)$, respectively.

The most typical metrics for distance between two probability distributions is the Kullback-Leiber divergence D_{KL} , defined as follows:

$$D_{KL}(P||Q) = \sum_{f \in \mathcal{F}} P(f) \log \frac{P(f)}{Q(f)} \quad (4.10)$$

Notice that the expression $D_{KL}(P||Q)$ is asymmetric in its arguments, i.e, in general $D_{KL}(P||Q) \neq D_{KL}(Q||P)$. This problem can be solved by adopting, as a definition of divergence, a symmetrized version of the previous expression:

$$D_{SKL}(P||Q) = \frac{1}{2} \left(\sum_{f \in \mathcal{F}} P(f) \log \frac{P(f)}{Q(f)} + \sum_{f \in \mathcal{F}} Q(f) \log \frac{Q(f)}{P(f)} \right) \quad (4.11)$$

However, KL divergence become infinite as soon as either P or Q vanish in one point of the support set, due to the denominators in the logarithm arguments of the two terms. This problem can be fixed by using the Jensen-Shannon Divergence (JSD), which is given by the following equation:

$$D_{JS}(P||Q) = \frac{1}{2} \sum_{f \in \mathcal{F}} \left(P(f) \log \frac{2P(f)}{P(f) + Q(f)} + Q(f) \log \frac{2Q(f)}{P(f) + Q(f)} \right) \quad (4.12)$$

JSD differs from the SKL divergence of equation (4.11) in that the denominator of the logarithm's argument consists now of the arithmetic average $\frac{P(f)+Q(f)}{2}$ of the two probabilities $P(f)$ and $Q(f)$.

Adapted Jensen-Shannon Divergence (AJSD)

If, as in our case, the probabilities P and Q are not available, we have an estimate of them through a finite sample represented in the form of a histogram for P and a histogram for Q . In this case the divergence computed on the histograms is a random variable. This variable, under appropriate assumptions, can be used to compute an estimate of the divergence between P and Q using error propagation under ML approach, as illustrated hereafter.

For P and Q consider that the channels at a point (feature) f of the corresponding histograms are characterized by the number of co-occurrences with f , denoted as k_f and h_f respectively. We define the following measured frequencies where:

$$x_f \equiv k_f/n \quad y_f \equiv h_f/m \quad (4.13)$$

Here, $n = \sum_f k_f$ and $m = \sum_f h_f$ are the sum of counts for the first and second histogram, respectively.

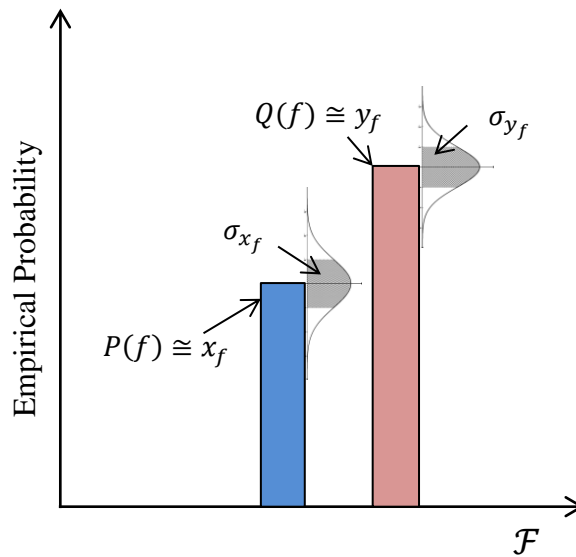


FIGURE 4.7: Two histogram channels corresponding to the feature $f \in \mathcal{F}$ taken from the empirical probability distributions P and Q respectively. Each histogram channel is considered as a normally distributed random variable

When the number of co-occurrences is high enough (large n and m), the quantities x_f and y_f can be considered to have normal distributions around the true probabilities $P(f)$ and $Q(f)$, respectively (Figure 4.7). Consequently, the *measured* JSD, denoted as d , can be considered as a stochastic variable defined as a function of the two normal

variables x_f and y_f . By substituting x_f and y_f in Equation 4.12 we get:

$$d = \frac{1}{2} \sum_{f \in \mathcal{F}} \left(x_f \log \frac{2x_f}{x_f + y_f} + y_f \log \frac{2y_f}{x_f + y_f} \right) \quad (4.14)$$

The value of this expression does not correspond, in general, to the ML estimate of JSD since the variances of the terms in the sum are unequal. In order to find the maximum likelihood estimate \hat{d} of the divergence, we need to proceed through error propagation as in the following steps:

1. Thanks to the normality condition stated above, the ML estimate of $P(f)$ corresponds to $x_f = k_f/n$ with the variance given in a first approximation by $\sigma_{x_f}^2 = k_f/n^2$. Similarly, the ML estimate of $Q(f)$ is $y_f = h_f/m$ with the variance given by $\sigma_{y_f}^2 = h_f/m^2$.
2. We represent the individual addendum term in the sum expression of Equation 4.14 as a random variable z_f :

$$z_f \equiv x_f \log \frac{2x_f}{x_f + y_f} + y_f \log \frac{2y_f}{x_f + y_f} \quad (4.15)$$

If the two variables x_f and y_f are independent, the variance propagation at the first order is given by:

$$\sigma^2(z_f) \simeq \left(\frac{\partial z_f}{\partial x_f} \right)^2 \sigma^2(x_f) + \left(\frac{\partial z_f}{\partial y_f} \right)^2 \sigma^2(y_f) \quad (4.16)$$

$$\simeq \log^2 \frac{2x_f}{x_f + y_f} \sigma^2(x_f) + \log^2 \frac{2y_f}{x_f + y_f} \sigma^2(y_f) \quad (4.17)$$

The variance $\sigma^2(z_f)$ can be easily calculated by substituting the quantities of step 1 in the equation (4.17).

3. Define the (statistical) precision w_f (to be used later as a weight) as: $w_f \sim \frac{1}{\sigma^2(z_f)}$. Then, the maximum likelihood estimate of the quantity d of equation (4.14) is given by the following weighted sum:

$$\hat{d} = \frac{\sum_f w_f z_f}{\sum_f w_f} \equiv D_{AJS}(P||Q) \quad (4.18)$$

With the variance given by:

$$\sigma^2(\hat{d}) = \frac{1}{\sum_f w_f} \quad (4.19)$$

We use \hat{d} as Adapted Jensen-Shannon Divergence (AJSD). Note that due to the statistical fluctuations in the samples, AJSD gives, in general, values greater than zero even when two samples are taken from the same distribution, i.e., even when the true divergence is zero. However, by weighting the terms according to their (statistical) precision, the scores produced by AJSD are expected to provide better estimate of the divergence than JSD does (refer to Section 4.3).

4.3 Evaluation

4.3.1 Dataset

In order to evaluate the performance of the proposed tag relatedness approach, we performed several experiments on a folksonomy extracted from Flickr. The folksonomy corresponds to images taken in the area of London. To avoid bulk tagging, we restricted the dataset to one image per user. The final dataset contains around 54,000 images with 4,776 unique tags occurring more than 10 times and a total of 544,000 tag assignments.

4.3.2 Qualitative Insight

For each of the 4,776 unique tags in the dataset, we identified its most related tags. Table 4.2 shows sample tags (first column) with the corresponding related tags ordered according to their degree of relatedness from left to right. The related tags are obtained by the cosine (COS), JSD and AJSD measures, respectively, and by using the top 1000 Laplacian features (more details in the next section). First, one can notice the overlap among the groups of related tags corresponding to the same initial tag. This can be explained because the compared approaches are based on the same tag representation, namely the tag-context. Second, we have recognized that, in general, the groups of related tags which are identified by AJSD have a higher cardinality than their counterparts which are identified using JSD and the cosine approaches (e.g. the tags "Car" and "Garden" in Table 4.2). That is because AJSD generates non-zero similarity even the two tags have different empirical probability distributions.

To investigate the effect of feature selection, we applied the Laplacian score method on the dataset to identify the most important tags. To generate the tag similarity graph, we set the number of nearest neighbors to 10 and used the Gaussian similarity function (Equation 4.2) with $u = 0.5$.

Figure 4.8 shows a plot of the top tags according to the calculated LS scores against their frequencies, i.e., their total occurrences in the folksonomy. Additionally, the plot

Initial Tag	Method	Related Tags
Airport	COS	Heathrow, KLM, duty, check, airports, runway
	JSD	Heathrow, runway, African, international, ramp
	AJSD	Heathrow, ramp, departures, president, restaurants
Car	COS	automobile, Citroen, driving, rolls, pit, wreck
	JSD	cars, classic, motor, Sunday, Ford, Mini, BMW, driving
	AJSD	cars, classic, Sunday, Ford, Mini, BMW, driving, Caterham, pit
Garden	COS	Covent, jardin, ING
	JSD	flower, gardens, rose, Covent, jardin
	AJSD	flower, gardens, Covent, jardin, pots, Nicholson, rocks
Thames	COS	path, Kingston, river, mud, embankment, Sunbury, shore
	JSD	river, path, Kingston, riverside, Greenwich, ship, embankment
	AJSD	river, water, riverside, path, Kingston, Greenwich, embankment
Music	COS	musician, bands, records, fighting, acoustic
	JSD	concert, rock, stage, festival, pop, jazz, song, records
	AJSD	concert, rock, festival, stage, pop, jazz, Simon, song
Olympics	COS	triathlon, men's
	JSD	Olympic, men's, arena, venue, women's, athlete
	AJSD	Olympic, men's, center, athlete, women's, venue, game, triathlon

TABLE 4.2: Sample tags with the corresponding most related tags

illustrates the most frequent tags in the folksonomy (*italic*). According to LS, the importance of a tag is determined according to its locality preserving power (refer to Section 4.2.2) which is independent from the frequency. For example a tag like "potter", which is much less frequent than the tag "england", has a higher Laplacian score, thus, it is considered as more important. In fact, since the folksonomy contains images taken in London, it is very likely that most images will be tagged with the word "england" disregarding their contents. Consequently, the tag "england" is expected to have a uniform co-occurrence with all other tags in the folksonomy. Therefore, it is less discriminative, i.e., it has a lower LS than a more specific tag like "potter", which is expected to have non-uniform co-occurrence patterns.

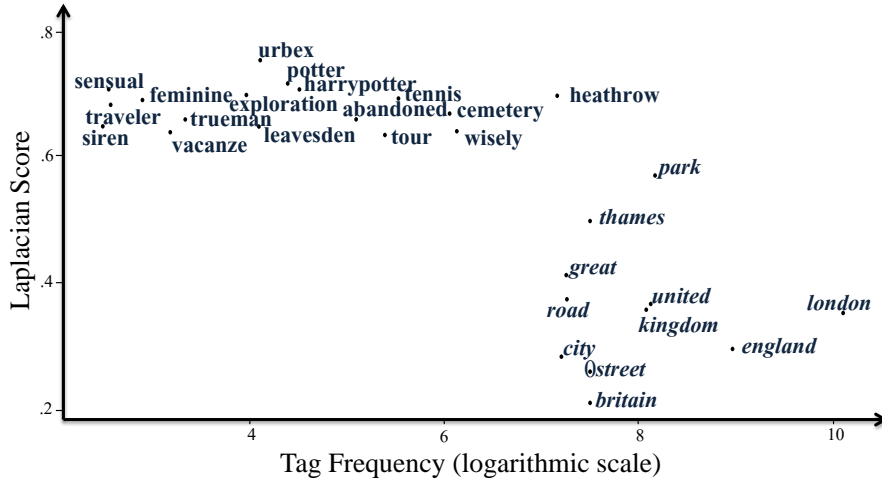


FIGURE 4.8: Tag importance according to LS algorithm vs. tag frequency

4.3.3 Semantic Grounding using WordNet

To provide a quantitative evaluation, we performed additional experiments using WordNet [Miller, 1995]. WordNet has been used by several works as a tool for semantically grounding tag relatedness measures [Cattuto et al., 2008, Srinivas et al., 2010, Markines et al., 2009]. The goal is to assess how a given tag relatedness measure approximates a reference measure. For our study, we used the Jinag & Conrath (JCN) measure as a reference since it showed a high correlation with human judgment [Jiang and Conrath, 1997]. Since the similarity score of JCN takes a value in the range $[0, \infty)$, we applied the transformation proposed by [Li, 2005] to convert the similarity into a distance measure in the range $[0, 1]$ according to the following function:

$$f(x) = \begin{cases} 1 & x \leq 0.06 \\ 0.6 - 0.4 \sin\left(\frac{25\pi}{2}x + \frac{3}{4}\pi\right) & 0.06 < x < 0.1 \\ 0.6 - 0.6 \sin\left[\frac{\pi}{2}\left(1 - \frac{1}{3.471x+0.653}\right)\right] & x \geq 0.1 \end{cases} \quad (4.20)$$

where f is non-increasing function and x refers to the JCN similarity between two words.

The goal of the evaluation is to assess the effectiveness of the proposed distance measure AJSD and the to investigate the effect of the Laplacian score feature selection on identifying related tags. For this purpose, we created a gold standard by identifying the most similar tag pairs from our dataset using WordNet and the JCN measure. After that, the relatedness between the tag pairs of the gold standard is calculated using different configurations: 1) our proposed distance measure *AJSD*, 2) the standard *JSD*

measure and 3) an approach based on the cosine of the tag co-occurrence vectors, denoted as *COS*. More specifically, the COS approach defines the relatedness between two tags t_1, t_2 based on the cosine distance between their corresponding tag-context vectors, i.e., $1 - \text{cosine}(v(t_1), v(t_2))$. Furthermore, we also examined the performance of these distance measures by using 1) feature selection based on the most frequent tags in the folksonomy (*FRQ*) and 2) features selection using the top Laplacian features (LS). The configurations of the investigated tag relatedness approaches are summarized in Table 4.3.

NR	Distance	Features
1	AJSD	FRQ
2	AJSD	LS
3	JSD	FRQ
4	JSD	LS
5	COS	FRQ
6	COS	LS

TABLE 4.3: Configurations of the evaluated tag relatedness approaches

The performance of a tag relatedness approach is evaluated according to the average JCN distance (Equation 4.20) over the set of most similar tag pairs identified by that approach. Furthermore, to deal with possible differences in the distributional properties of the investigated measures, we followed the same method as in Markines et al. [2009]. Thereby, the performance of a tag relatedness approach is evaluated based on the correlation between the rankings it produces and that of a reference measure. Specifically, we used Kendall τ -correlation coefficient to verify how the ranking (of the most similar tag pairs) generated by a tag relatedness approach correlate with the reference ranking according to the JCN measure.

For each of the six configurations, we calculated the number of unique tag pairs which are most related. On average each method produced 4,100 unique tag pairs. The average percentage of tag pairs with both tags having corresponding entries in WordNet amounts to 40%.

Effectiveness of Feature Selection

In order to investigate the effect of distance measure, we identified the top 1,000, 2,000 and 3,000 features according to the frequency (FRQ) as well as the LS feature selection methods. We used the top features to build a probability distribution for each tag in the

test folksonomy and applied each of the three distance measure AJSD, JSD and COS to identify the most similar tag pairs.

Figures 4.9(a), 4.9(c) and 4.9(e) show the average JCN distance over the sets of most similar pairs as identified by AJSD, JSD and COS, respectively. In general, LS outperforms FRQ disregarding the applied distance measure. That is because LS feature selection results in a smaller average JCN distance under the three applied distance measures and with varying number of features. One exception is the case where the COS method is used with the top 2,000 LS feature. Thereby, FRQ feature selection performs slightly better.

With regards to Kendall τ -correlation, the same conclusion can be made. When LS feature selections is applied, the rankings of most similar tag pairs, which are generated by the three distance measures, have higher correlation with their rankings according to the reference measure JCN. (Figures 4.9(b), 4.9(d) and 4.9(f)).

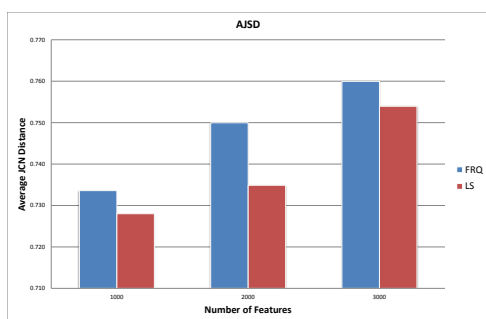
Although it seems that LS outperforms FRQ feature selection, it is important to note that the difference in the performance between the two methods and according to both evaluation measures is relatively small. However, in contrast to the FRQ method, LS has the potential to generate even better results by tuning its parameters, such as the selected similarity measure and the size of the nearest neighbor graph (refer to Section 4.2.2).

Effectiveness of Distance Measures

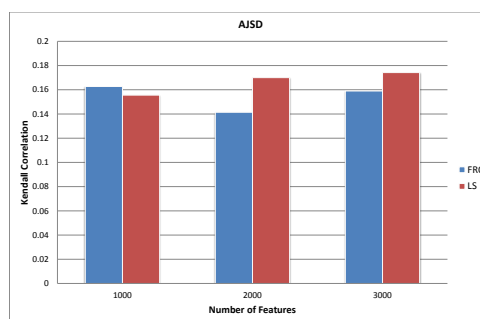
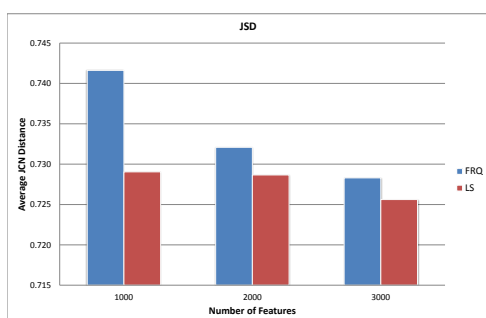
We compared the performance of AJSD, JSD and COS using the the top 1,000 FRQ as well as LS features. Figure 4.10 shows that the proposed AJSD measure have a positive effect of on identifying similar tags. Indeed, AJSD outperforms the two adversary methods JSD and COS with respect to the average JCN distance (Figure 4.10(a)) and rank correlation (Figure 4.10(b)). This observation holds for both feature selection methods, i.e., FRQ and LS. The second performing measure is JSD, which in turn outperforms the COS measure.

4.3.4 Evaluation using Large Scale Co-occurrence Statistics

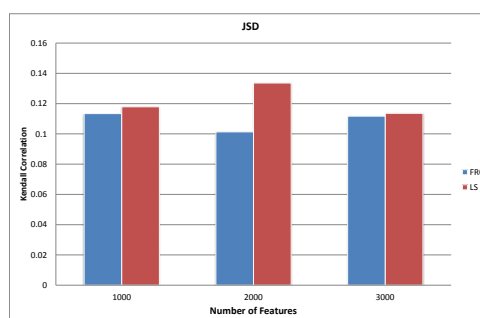
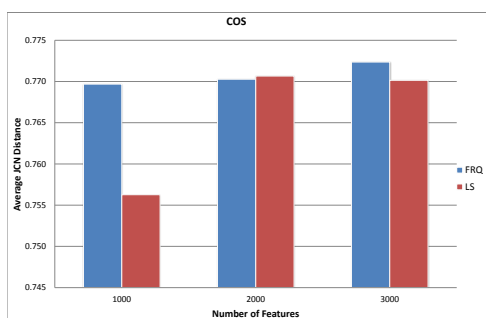
Although WordNet provides a well-established ground truth, it is unable to cover the diversity of user tags. As mentioned before, only 40% of the tag pairs, which were identified by the investigated tag relatedness approaches, have corresponding entries in WordNet. This is because WordNet is limited to the English language and it does not consider entities like abbreviations, proper names or terms in colloquial language. Few



(a) AJSD Average JCN

(b) AJSD Kendall τ -correlation

(c) JSD Average JCN

(d) JSD Kendall τ -correlation

(e) COS Average JCN

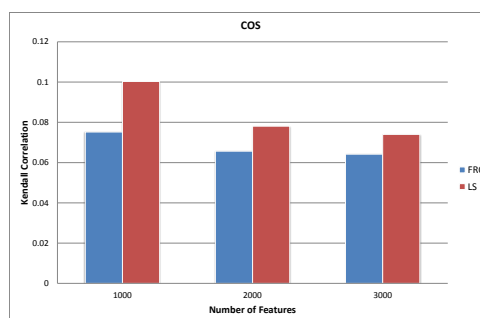
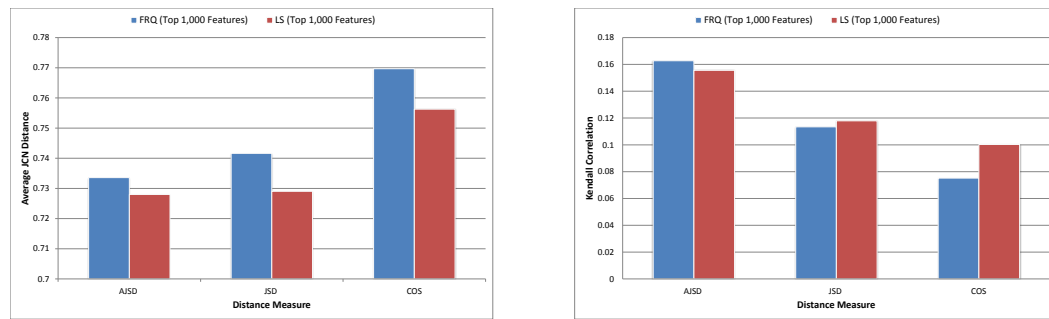
(f) COS Kendall τ -correlation

FIGURE 4.9: Comparison between FRQ and LS feature selection methods

examples of related tag pairs which could not be identified by WordNet are shown in Table 4.4.

To deal with the limitations of WordNet, we evaluated the relatedness between the pairs based on their distributional co-occurrence in a large-scale corpus. For this purpose, we used *DISCO* (DISTRIBUTIONALLY related words using CO-occurrence) measure of semantic relatedness [Kolb, 2008]. We configured the system to use Wikipedia [Wikipedia, 2014]



(a) Average JCN distance achieved by AJSD, JSD and (b) Kendall τ -correlation between the ranking produced by the investigated tag relatedness approaches and the cosine (COS). Tag representations are created according to the top 1,000 features determined by the frequent tags-based (FRQ) and the LS-based (LS) feature selection methods

compared distance measures are AJSD, JSD and the cosine (COS). Tag representations are created according to the top 1,000 features determined by the frequent tags-based (FRQ) and the LS-based (LS) feature selection methods

FIGURE 4.10: Comparison between tag relatedness distance measures

Similar Tag Pairs	Description
<i>Hadid Zaha</i>	Iraqi-British architect
<i>Bradley Wiggins</i>	English professional road and track racing cyclist
<i>Aer Lingus</i>	Irish airline
<i>Boeing Airbus</i>	Aircraft manufactures
<i>Bianco Nero</i>	Italian for white and black
<i>Warner Bros</i>	American producer of film, TV, and music entertainment.
Olympische Spelen	Dutch for Olympics
Verenigd Koninkrijk	Dutch for United Kingdom
Anish Kapoor	Indian sculptor
Psittacula Krameri	Gregarious tropical Afro-Asian parakeet species

TABLE 4.4: Similar tag pairs which are identified by the proposed tag relatedness approach. The listed tags do not have corresponding entries in WordNet. Tag pairs shown in italic are identified as related by DISCO. The description of each tag pair were obtained from Wikipedia

as a corpus. The advantage of Wikipedia is that it offers a much wider vocabulary than WordNet, e.g., it is more likely to find Wikipedia articles containing proper names.

With respect to the similarity measure, it has been shown that DISCO similarity correlate with JCN similarity by a coefficient of 0.38 [Kolb, 2008], however, this value is not high enough to replace human judgment. For example, consider the two word pairs "humour, humor" and "colour, color", the similarity between the words of each pair

is almost 1 according to WordNet and the JCN measure. However, DISCO produces two different similarity values for both pairs. According to DISCO the similarity between "humour" and "humor" amounts to 0.65 while the similarity between "colour" and "color" is 0.81. The example shows that it is not practical to apply the same evaluation measures (i.e., the average similarity and Kendall τ -correlation) as we did with WordNet. Still, DISCO can be used as a reference to evaluate the relatedness between two words. Therefore, instead of grounding the performance on the real similarity values, we use the number of tag pairs falling in certain similarity range as an indicator for the quality of the tag relatedness approach.

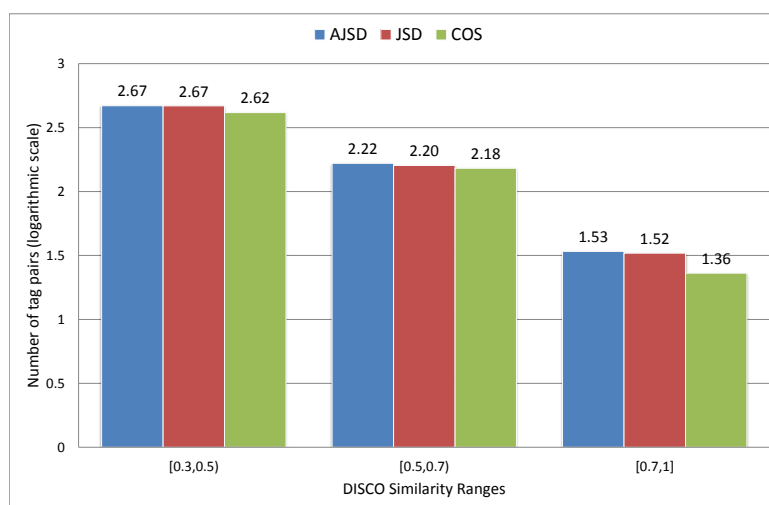


FIGURE 4.11: Histograms of the number of tag pairs identified by each of the distance measures AJSD, JSD and COS using the top 1,000 LS Features. Each channel of the histogram corresponds to the number of tag pairs in a predefined range of DISCO similarity

To this end, we determined three types of ranges for DISCO similarity corresponding to low, middle and high distributional co-correlation between words. Consequently, for each tag pair generated by a tag relatedness approach, we count the number of tag pairs with a DISCO similarity falling into each of the three defined similarity ranges.

Figure 4.11 shows the results of applying the distance measures AJSD, JSD and COS using the top 1,000 LS features. It can be seen, that disregarding the similarity range the COS method produces the smallest number of similar tag pairs. We can also observe that AJSD produce the highest number of tag pairs for all three similarity ranges. This gives an additional clue on the advantage of AJSD over the adversary measures.

4.4 Resolving Tag Ambiguity

In this section, we want to demonstrate how tag relatedness information combined with semantic resources can be used to resolve tag noisiness in the context of automatic image annotation. Figure 4.12 illustrates the procedure which we apply. First, tag relatedness information are extracted from the folksonomy using the approach described previously. After that, most related tag pairs are determined based on the distance between them according to the applied tag relatedness approach. Next, the type of the semantic relationship between the two tags is identified (more details below). The tag pairs are then stored in a database with the corresponding semantic relationship. Furthermore, we define two actions to disambiguate the tag pair: *tag removal* and *tag combination*. The tag removal action is used to remove redundant tags as in the case of lexical variation and synonym tags. For example, for tag pairs like "color" and "colour" or "flower" and "rose" only one of the two tags is kept. The tag combination action assumes that the tags are semantically related according to a relationship different from lexical variation or synonymy. Therefore, it is likely that a more specific term will emerge when the two tags are combined. For example, combining the two related tags "Eiffel" and "Tower" results in identifying the name of the landmark "Eiffel Tower".

To refine tag proposals generated by an AIA approach, we first build tag pairs from the proposal list. Next, each tag pair is used to query the database of semantically similar tag pairs. Finally, the tag pair is replaced by the result of the corresponding action.

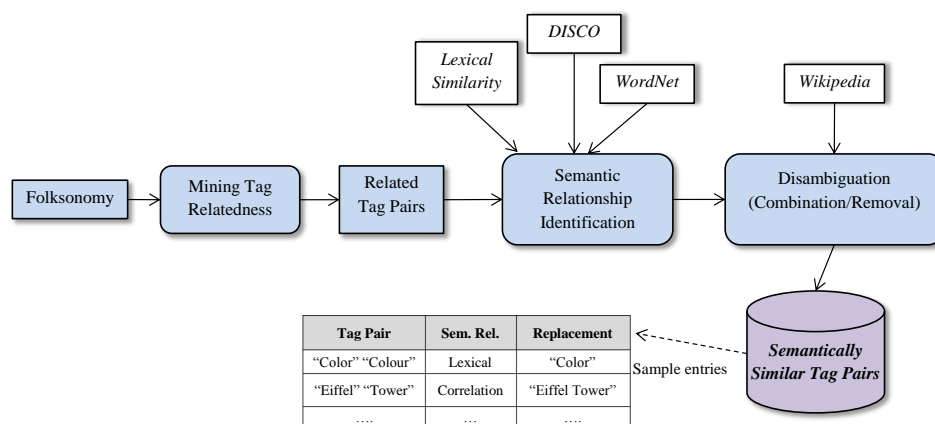


FIGURE 4.12: Workflow of the proposed tag proposals disambiguation procedure

Semantic Relationship Identification

The presented tag relatedness approach determines related tags without specifying the kind of the semantic relationship between them. In the context of automatic image annotation, further information are needed to refine the list of generated tag proposals. For this purpose, we developed a simple approach to identify the type of semantic relationships between related tag pairs. Thereby, we deal with three kinds of semantic relationships: lexical variation, synonymy and correlation.

To determine whether two tags are lexical variation of each others (e.g., "olympics" and "olympic"), we apply a lexical similarity measure based on the widely used algorithm of *edit distance* or *Levenstein distance* (*Lev*) [Marzal and Vidal, 1993]. Edit distance, is defined as the minimum number of elementary edits (e.g., insert, delete, replace) which are required to transform one string into another. We use a variant of the the edit distance called *normalized edit distance* (*NED*) [Marzal and Vidal, 1993] which is calculated by normalizing the edit distance between two words by the length of the longest word.

$$NED(w_1, w_2) = \frac{Lev(w_1, w_2)}{\max(Len(w_1), Len(w_2))} \quad (4.21)$$

where w_1 and w_2 are two words and *Lev* is a function that computes the edit distance between two words based on Levenstein proposal and *Len* is a function that returns the number of characters (length) in a word. Words with NED less than a predefined threshold θ will be classified as a lexical variation of each other. In our experiments good estimations of the lexical variation were provided by $\theta = 0.35$.

To determine whether two tags are synonyms, the taxonomic shortest-path length between the two tags in WordNet is calculated. In WordNet a path length of one corresponds to direct synonyms. Sample word pairs with path length of one are ("cab", "taxi"), ("dawn", "morning") and ("stadium", "arena") .

For both synonyms and lexical variation tags, we apply the tag removal action. That means, the two tags are reduced to only one of them. Thereby, we keep the tag which occurs most in the folksonomy. In the case where the two tags are neither lexical variation nor synonyms, we check the distance between them according to the WordNet-JCN measure. For tag pairs that do not have corresponding entries in WordNet, we calculate their DISCO distance [Kolb, 2008]. If the distance between the two tags according to either or both measure (i.e. JCN and DISCO) is below a predefined threshold, they are considered to be correlated. As mentioned before, correlated words can be combined to obtain more specific terms. To identify such terms, we use the tag pair to query

Wikipedia for articles which contains both words in their titles. If a corresponding article was found, we use the title of that article to replace the tag pair. Otherwise, the tag pair is kept unchanged.

Table 4.5 shows sample of related tag pairs which were identified by applying AJSD measure and LS feature selection on the folksonomy presented in Section 4.3.1. The semantic relationships between the pairs are identified according to the above described procedure. We can see, in addition to determining redundant tags the described method helps to identify named entities such as persons, e.g., "Edwin Lutyens" and landmark names, e.g., "Big Ben".

Related Tag Pairs	Semantic	Final Term
"hamstead" "hampstead"	lexical variation	"hampstead"*
"apartment" "apartments"	lexical variation	"apartment"*
"archaeology" "archeology"	lexical variation	"archaeology"*
"burn" "glow"	synonymy	"burn"*
"overcast" "cloud"	synonymy	"cloud"*
"dawn" "sunrise"	synonymy	"dawn"*
"fall" "autumn"	synonymy	"autumn"*
"metro" "tube"	synonymy	"metro"*
"stadium" "arena"	synonymy	"arena"*
"tube" "underground"	synonymy	"underground"*
"san" "francisco"	correlation	"San Francisco"**
"chapel" "abney"	correlation	"Abney Park Chapel"**
"battersea" "power"	correlation	"Battersea Power Station"**
"ben" "big"	correlation	"Big Ben"**
"camden" "town"	correlation	"Camden Town"**
"wharf" "canary"	correlation	"Canary Wharf"**
"airport" "heathrow"	correlation	"London Heathrow Airport"**
"city" "london"	correlation	"City of London"**
"edwin" "lutyens"	correlation	"Edwin Lutyens"**

* *Tags with higher occurrence in the folksonomy than there related counterparts*

** *entries corresponding to titles of Wikipedia articles*

TABLE 4.5: Sample of tag pairs extracted from the folksonomy presented in Section 4.3.1 with the corresponding semantic relationships and their disambiguation terms

Note that, the presented procedure can be extended to deal with triples or even clusters of related tags. However, investigating such options is out of the scope of this thesis.

In Chapter 8 we will demonstrate experimentally the effectiveness of the presented tag relatedness approach and the presented tag disambiguation procedure on improving the performance of search-based AIA.

4.5 Summary

In this chapter, we demonstrated the effect of noisy tags on the performance of search-based AIA. Furthermore, we introduced an approach for tag relatedness based on Laplacian feature selection and a novel distance measure based on Jensen-Shannon Divergence. The effectiveness of our approach has been demonstrated experimentally using a folksonomy extracted from Flickr. The evaluation is performed using WordNet as a ground truth. An additional evaluation metric is also presented based on the distributional similarity of words according to Wikipedia and the DISCO semantic similarity measure. We also introduced a procedure for refining tag proposals generated by AIA by using the extracted tag relatedness information.

Part III

Improving SURF-based Image Matching

Chapter 5

Classification-based Keypoint Pruning

This chapter deals with the challenge of efficiently matching huge number of images based on the local image feature SURF. To accelerate the matching process, an approach for reducing the number of compared feature descriptors is presented. The approach explores the problem of keypoint characterization and propose a solution using classification techniques.

5.1 Introduction

As presented in Chapter 1, the fourth phase of our automatic annotation approach requires identifying the set of visual neighbors of the image which we want to annotate. To achieve this, we decided to perform image matching based on the SURF method [Bay et al., 2008]. SURF is characterized by its fast computation and a matching performance compared to that of SIFT [Lowe, 2004]. Furthermore, the standard SURF descriptor is half as small (64 dimensions) as that of SIFT. Still, in the case of our AIA approach identifying the visual neighbors implies comparing huge number of SURF descriptors.

To address this problem, we propose an approach for accelerating SURF-based image matching based on machine learning techniques. The goal is to construct a binary classifier which is able to identify, for each image, the subset of keypoints that are important for the matching. Subsequently, unimportant keypoints can be excluded from the matching process. As a result, the matching can be done more efficiently by comparing much smaller subsets of keypoint descriptors.

The motivation for our approach is based on two observations. First, applications that employ CBIR techniques - such as most search-based AIA approaches – perform image matching by following two types of processing phases: an offline phase, in which image features are extracted and stored, and an online phase, in which a search is initiated to discover the visual neighbors of a given input image. This is done by comparing the feature descriptors of the input image with those of the images stored in the database. Hence, reducing the number of compared feature descriptors is a key requirement for boosting the performance of the online phase. Second, it can be observed that a considerable part of the keypoints discovered by algorithms for local feature detection (including SURF) contributes a little to identifying similar images [Turcot and Lowe, 2009]. Additionally, in most cases, it is enough to have a small amount of keypoint correspondences to decide whether two images are similar or not. For instance, it has been shown in [Ke et al., 2004, Jones et al., 2010] that a small number of five keypoint correspondences provides a strong clue that the associated images are similar. Please refer to the images shown in Figure 5.1 for more explanation on the latter observation.

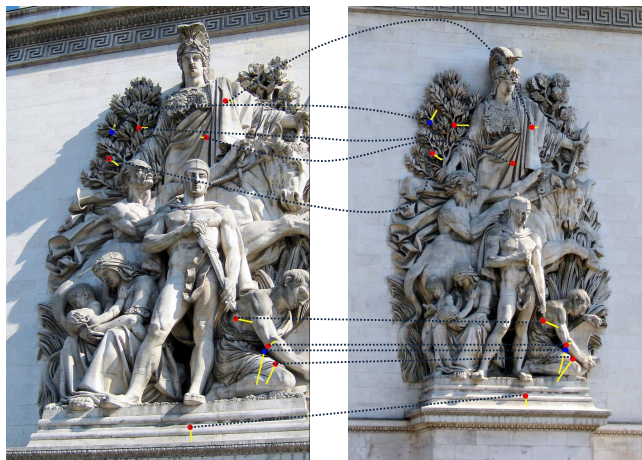


FIGURE 5.1: Two images depicting the same scene from different perspectives. Keypoint correspondences between the two images are connected using dotted lines

The two images illustrate the same scene from different perspectives. By matching the two images using the SURF algorithm, only 10 keypoint correspondences were identified. This number is quite small compared the total of 1602 and 1300 keypoints extracted from the left and the right image, respectively. In the naive case, identifying the common keypoints requires performing 1602×1300 comparisons among 64-dimensional feature vectors.

Based on the above discussion, the performance of the matching phase can be improved if we are able to identify a subset of "significant" keypoints and restrict the matching accordingly to the corresponding feature descriptors. The literature refers to this process

as *keypoint pruning* and it has been the topic of several research studies (refer to Section 2.5.3). Before we present our solution in this regard, we briefly introduce a state-of-the-art method for keypoint pruning based on visual attention and saliency maps. The presented method will serve as a basis to investigate the effectiveness of our approach.

5.2 Keypoint Pruning using Visual Attention Models

Visual attention is a research field which aims at simulating the way in which humans perceive photos. In general, humans analyze a visual scene by selecting specific areas in the image, called *salient regions*, that seem most relevant to them. Then, finer/higher-level activities, such as object recognition are conducted on the salient regions only. An early computational model for simulating this process was introduced in a widely cited paper of [Itti et al., 1998]. The proposed visual attention model aims at identifying regions in images which correspond to human perception of saliency. The processing stages of this model are illustrated in Figure 5.2. First, for a given image simple color, intensity and orientation features are extracted using the center-surround technique. After that, for each feature a map is created. Finally, feature maps are fused in a saliency map which topographically codes conspicuous locations in the input image.

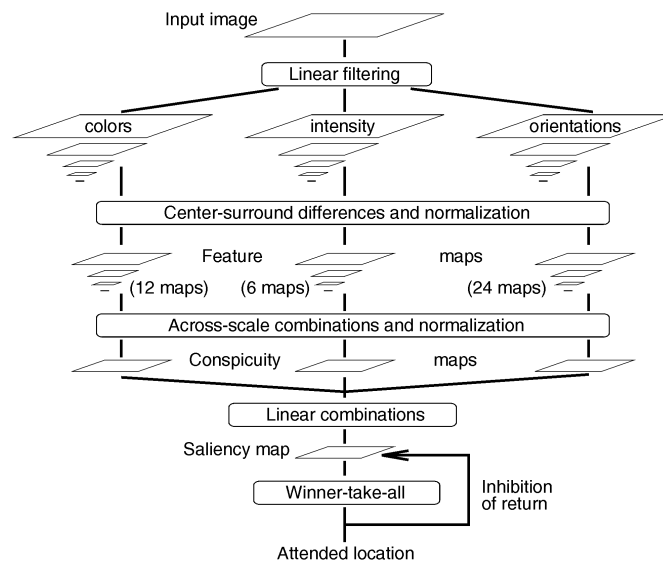


FIGURE 5.2: The model of saliency-based visual attention according to [Itti et al., 1998]

As discussed in Section 2.5.3, saliency maps have been successfully used to perform keypoint pruning [Pimenov, 2009, López-García et al., 2011, Chen et al., 2011]. The assumption of such approaches is that the importance of a keypoint can be determined according to its location in the image. That is, keypoints which fall in salient regions

would have more importance than those lying in non-salient regions. Subsequently, the number of keypoints can be reduced by only considering those which belong to the salient image regions. Consequently, image matching can be performed more efficiently by comparing smaller subsets of feature descriptors which correspond to the keypoints of the salient regions. This process is illustrated in Figure 5.3. A saliency map is generated for the input image shown in Figure 5.3.a. Usually, saliency maps are represented by a gray-scale version of the input image. Bright pixels (values near 1) represents salient areas in the image, while dark ones (values near 0) are of low saliency (Figure 5.3.b). In Figure 5.3.c the complete set of keypoints as discovered by SURF are shown, while Figure 5.3.d illustrates the results of reducing the keypoints to those which correspond to salient areas (brightness values above 0.4).

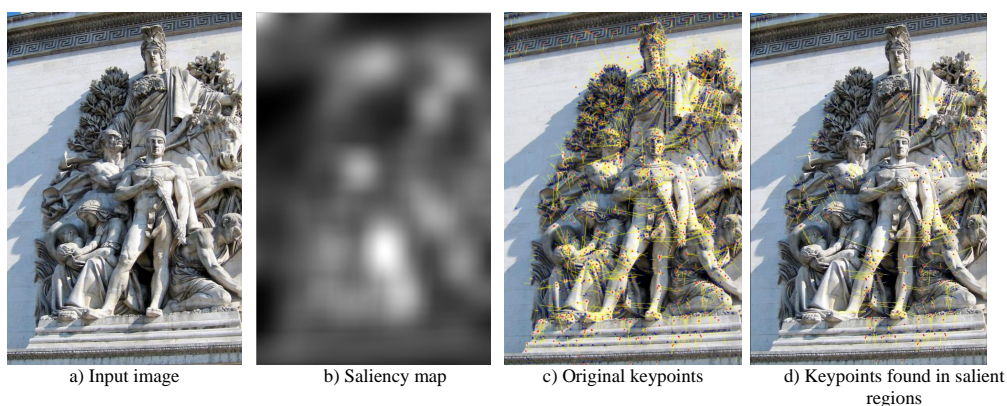


FIGURE 5.3: Keypoint pruning using saliency maps. a) The input image, b) the corresponding saliency map, c) the image with the identified SURF keypoints (without keypoint pruning), d) the image with the subset of SURF keypoints corresponding to salient regions

Note that by using this approach the ratio of detected keypoints differs according to the nature of the image, i.e., the size of the salient regions. Furthermore, this number can also be controlled by using different values for the saliency threshold. In the example above, 1,300 keypoints were detected from the input image with a subset of 499 keypoints corresponding to salient areas (about 38% of the original set).

5.3 Keypoint Pruning as Classification Problem

Our solution for keypoint pruning is based on building a classifier which is able to identify two categories of keypoints. *Significant* keypoints which represent a crucial indicator for image similarity, and, *insignificant* keypoints, which are usually noisy, i.e., they have low or even biased impact on identifying visually similar images. By categorizing keypoints

in this manner, the matching performance can be significantly improved by limiting the matching to the set of significant keypoint descriptors.

To explain how such a classifier can be constructed, suppose we have an input image I with the set of the associated SURF keypoints KP . Each keypoint $kp \in KP$ can be then described by a set of features F extracted from a patch of width w centered on the keypoint, denoted as $Q_w^F(kp)$. Additionally, assume that there exists a perfect labeling function, Y , that assigns a label for each keypoint based on the associated feature patch as given below:

$$Y(Q_w^F(kp)) \in L; L = \{-1, 1\} \quad (5.1)$$

The value -1 corresponds to the insignificant class and 1 to the significant one. In practice, Y cannot be directly derived, however, an estimation of it \hat{Y} can be learned by applying classification techniques. The quality of the classifier is estimated by its ability to predict the correct classes for test keypoints. Accordingly, a best performance can be achieved if the classifier is able to predict the right classes for test keypoints with a very low error rate ϵ . In other words, the probability that the class predicted for a given keypoint $kp \in KP$ differs from the true one is low:

$$P(Y(Q_w^F(kp)) \neq \hat{Y}(Q_w^F(kp))) < \epsilon \quad (5.2)$$

To build the classifier \hat{Y} , a training dataset with labeled instances has to be created. A training instance corresponding to a keypoint can be defined as tuple (x, y, Q_w^F, L) in which (x, y) are the coordinates of the keypoint in the corresponding image, Q_w^F is the characterizing feature patch and L is the class of the keypoint. This set can then be used to train the classifier. Subsequently, the learned model can be used to determine the usefulness of a test keypoint kp' for the matching. To achieve this, the same set of features which were used in the training process are extracted from a patch Q_w^F centered on kp' and passed to the classifier. Finally, kp' is considered by the matching if $\hat{Y}(Q_w^F(kp')) = 1$, otherwise, it is excluded from the matching.

In the following sections the process of building the keypoint classifier and the associated challenges are discussed in detail.

5.3.1 Training Dataset

We created a training dataset from a collection of 1,100 groups of images taken from the Object Recognition Benchmark dataset [Nister and Stewenius, 2006]. Each group consists of four images that are visually similar, i.e., depicting the same objects, but

taken from different perspectives and under different illumination conditions (Figure 5.4). The dataset was used to provide training instances, i.e., a collection of labeled keypoints. For this purpose, an image was selected randomly from each group and matched with the other three images in the same group using the SURF algorithm. A keypoint was labeled as *significant* if it has a correspondence in each one of the other three images. Otherwise, the keypoint is assigned the label *insignificant*.

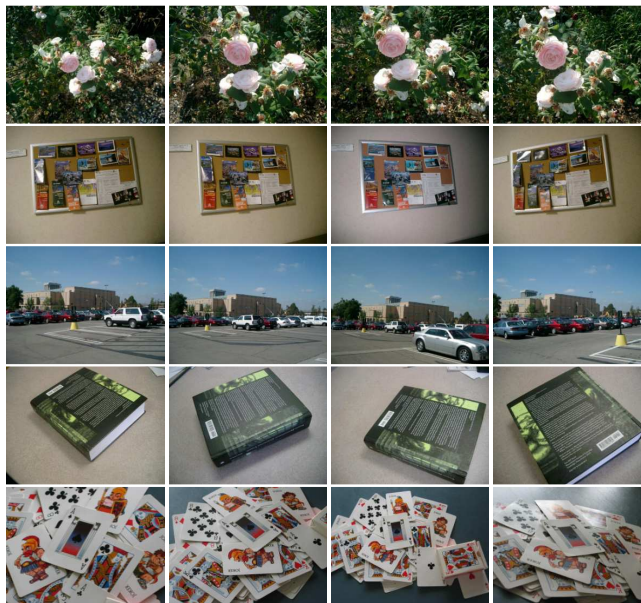


FIGURE 5.4: A sample of image groups, which were used to create the training dataset. The image groups were taken from the Object Recognition dataset [Nister and Sweeney, 2006]

Diversity of the Training Set

While building the training dataset, selecting keypoint instances from the adjacent areas in the image can lead to redundancy. This is due to possible overlap among the corresponding keypoint patches. Training the classifier with redundant instances may lead to bias in the produced model. To deal with problem, we apply a distance based filtering on the training instances. More specifically, for two training instances extracted from the same image and having the same label, the two instances must be spatially far enough from each other. To achieve this, the pairwise Euclidean distance between the coordinates of the keypoint instances is calculated. Next, an instance is discarded from the training dataset if there is already an instance which is close to it in the training dataset, i.e., the distance between the two instances is less than a threshold d . In our

experiments, we followed the recommendation of [Calonder et al., 2008] and used $d = 5$ pixels.

Dealing with Imbalance

Labeling keypoints using the above described approach leads to an imbalanced training dataset, in which the number of insignificant instances greatly exceeds the number of significant ones. In our training dataset less than 5% of the training instances were labeled as significant. In general, using an imbalanced training dataset leads to a poor classification performance [Provost, 2000] because the classifier tends to assign all instances to the majority class (in our case the insignificant class).

To address this problem, we applied a sampling approach on the obtained training dataset. We investigated different sampling configurations and realized that *random sampling without replacement* [Cochran, 2007] is most convenient in our case. In contrast to other sampling methods, such as SMOTE [Chawla et al., 2002], sampling without replacement ensures that the classifier is trained using real instances. While in SMOTE for example, the training dataset can again suffer from high redundancy according to replicating instances from the minority class to achieve the balance.

It is also important to note that a main drawback of sampling without replacement is that it can lead to a sharp reduction in the size of the training dataset. Using this approach, the maximum number of training instances is as twice as the number of instances in the minority class. Therefore, it is important to make sure that there are enough instances of the minority class in the training dataset before applying this kind of sampling. Otherwise, the learned model will be highly biased due to the small size of the training dataset.

5.3.2 Classification Features

The training instances must be described according to discriminative features before they can be fed into the learning phase. The quality of the features has a direct influence on the performance of the classifier. For this purpose, we investigated different kinds of image features extracted from squared patches centered on the keypoints. The set of features that we investigated includes the SURF descriptor itself, as well as other color and texture features, which we describe below.

SURF Descriptor: SURF descriptor has 64 dimensions and is generated by calculating Haar wavelet responses in 4×4 oriented square sub-regions centered on the keypoint [Bay et al., 2008]. We used the SURF descriptor to characterize the keypoints and

extended it with four additional attributes: 1) keypoint strength which represents the intensity of the corresponding blob. Here, positive values indicate dark blobs while negative values indicate light ones, 2) the Gaussian scale [Lindeberg, 1994] at which the keypoint was discovered, 3) the trace of the Gaussian matrix which was used to discover the keypoint blob and 4) the orientation of the keypoint. The final feature vector has 68 components.

Reduced SURF Descriptor: we applied attribute selection on the generated SURF descriptor to identify the most important classification attributes. For this purpose, we used the RELIEF-F [Liu and Motoda, 2007] method for feature selection. RELIEF-F is an effective feature selection algorithms which starts by selecting instances at random from the training set and determining their nearest neighbors. Subsequently, feature weights are adjusted, so that higher weights are given to features that discriminate each instance from its neighbors which belong to different classes. The algorithm identified 48 distinctive attributes of the SURF descriptor. We coined the new feature as Reduced SURF and used it to train a keypoint classifier.

Color Histogram: color histogram is a global image feature which represents the color distribution of an image (or a region of it) [Swain and Ballard, 1991, Jain and Vailaya, 1996]. It is calculated by defining a number of ranges (bins) for each color component according to a given color space. After that, the number of pixels falling in each bin is determined. We used the RGB color space and for each of the three color channels we used 8 bins. The final histogram consists of $8^3 = 512$ bins in total.

Color and Edge Directivity Descriptor (CEDD): this feature combines both color and texture features in one histogram [Chatzichristofis and Boutalis, 2008a]. It allows fast image retrieval since it is limited to 54 bytes per image. CEDD is created by splitting the image in a predefined number of blocks and calculating the color histogram of each block in the HSV color space. A 24-bins histogram of five different colors is generated for every block. After that, five filters are used to extract the texture information related to the edges present in the image. The extracted edges are classified in vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edges. The final descriptor consists of 144 bins.

Fuzzy Color and Texture Histogram (FCTH): similar to CEDD, the FCTH feature combines color and texture information in one quantized histogram [Chatzichristofis and Boutalis, 2008b]. First, a 10-bins histogram is extracted for 10 colors in the HSV color space. These colors are preselected based on the positions of the vertical edges in each channel. The histogram is extended to 24 bins by separating each color in 3 hues: dark color, color and light color. Then, Haar Wavelet transform is applied on the luminance

component Y of YIQ color space. Finally, the extracted texture information is used to enrich the feature resulting in feature histogram of 192 bins.

Joint Composite Descriptor (JCD): the idea of JCD is based on the fact that the color information of FCTH and CEDD are extracted from the same fuzzy system. Accordingly, JCD combines the texture area of the two features [Chatzichristofis et al., 2009]. JCD is built from 7 texture areas with each area made up of 24 sub regions corresponding to color areas. The feature histogram consists of 168 bins.

5.3.3 Classification Using Random Forest

After the training dataset was constructed, we built for each keypoint in the training set a feature descriptor based on each of the feature presented in the previous section. Next, for every single feature we trained a Random Forest classifier.

Random Forest (RF) [Breiman, 2001] is a state-of-the-art machine learning method that belongs to *ensemble learning* algorithms. The prediction of Random Forest is obtained by aggregating the predictions of several other tree classifiers [Rokach, 2010]. The effectiveness of Random Forest can be compared to that of other powerful classifiers, such as support vector machines (SVMs) [Cortes and Vapnik, 1995, Muller et al., 2001]. Moreover, Random Forest avoids overfitting the training data and generates an unbiased estimate of the generalization error. Briefly, a Random Forest classifier works as follows. A "forest" is built from a collection of n tree classifiers. Each tree is built from bootstrapped sample of the training data. In contrast to traditional classification trees [Rokach and Maimon, 2005], in which the best split for a tree is selected from all provided predictors, the trees in Random Forest are grown by choosing the best split predictor out of a random selection of m predictors. The leaves of each classification tree contain the posterior distribution of the classes.

Formally, let $T = \{T_i\}_{i=1}^n$ denote the set of trees in the forest and $N_i = \{N_{i,j}\}_{j=1}^{k_i}$ the set of leaf nodes of the i^{th} tree where k_i is the number of the corresponding leaf nodes. Furthermore, let \mathcal{C} be the set of available classes. Now, a leaf node $N_{i,j}$ contains the distribution of the classes at that node, denoted as $P_{N_{i,j}}(\mathcal{C})$. To predict the class of a new feature vector v , it is dropped over each tree in the forest until it reaches a leaf node $N_{i,v}$ in each tree. The final class $C \in \mathcal{C}$ of the feature vector v is determined by a majority voting of the n trees and is given by the following conditional probability:

$$P(C|v) = \operatorname{argmax}_{C \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^n P_{N_{i,v}}(C)$$

Random Forest classifiers have further interesting characteristics. It provides a robust generalization error estimate called out-of-bag (OOB) error. During the training phase and while building the trees about one third of the training sample is left out. This sub-sample contains what is called the out-of-bag (OOB) instances and is used for testing the corresponding tree classifier. The proportion of misclassified OOB instances indicates the classification error of the corresponding tree. By averaging the OOB errors over the whole set of tree classifiers, a generalization error for the Random Forest classifier is obtained. Consequently, there is no need for a further cross-validation phase to quantify the quality of a Random Forest classifier. Furthermore, Random Forest is easy to configure. Compared to the complicated setting of SVM classifiers, Random Forest have only two parameters. The number of trees n and the number of predictors m from which a split for the tree is randomly selected.

5.4 Experimental Evaluation

In the following sections, the proposed keypoint pruning approach is evaluated from different perspectives. As mentioned before, for each of the presented features, a Random Forest keypoint classifier was constructed. In the following experiments, we first show the classification accuracy of each of the trained Random Forest classifiers. Next, the performance of image matching using our keypoint pruning approach is evaluated and compared to another method which applies visual attention models. Finally, a runtime analysis is provided.

5.4.1 Classification Performance

We built a training dataset consisting of about 10,000 keypoint instances distributed uniformly on the significant and insignificant classes. Each keypoint was characterized by each of the features presented in Section 5.3.2. The features were extracted from squared patches centered on the keypoints. The width of the patch has a great effect on the classification results. Too small patches make the extracted features miss information about the corresponding keypoints while too large ones might overlap. To address this problem, we identified the size of the patch dynamically using the approach described in [Bay et al., 2008]. Accordingly, we set the width of the patch to 6 times the scale σ at which the keypoint has been discovered using SURF detector.

For each individual feature, we trained a Random Forest classifier using 100 trees. Additionally, we followed the recommendation of Weka data mining framework [Hall et al., 2009] and set the number of randomly selected features to $\log_2(m)$, where m is the

number of feature attributes. The accuracy of the classifiers was evaluated according to the OOB error rate. The results shows that the classifier which was trained using the Reduced SURF feature provides the smallest OOB rate of about 24%. The other classifiers which were trained by each of the remaining features (SURF, CH, CEDD, FCTH, and JCD) separately provided a comparable OOB values around 36%.

5.4.2 Effectiveness of Keypoint Pruning

Before discussing the evaluation results, we introduce issues related to using classification as well as saliency maps for keypoint pruning.

Keypoint Classification Considerations

For a given keypoint (described by a feature patch), the classifier tells with which probability the keypoint belongs to each class. Consequently, a keypoint is assigned to the class with the highest probability (in binary classification this corresponds to the class with a probability higher than 50%). However, due to the classification errors, the classifiers cannot provide perfect predictions. Therefore, the probability value at which a decision is made to assign a certain class to a keypoint is crucial for the matching performance. To understand this, let us assume that the classifier predicts that a great part of the keypoints of an image belongs with a probability of 50.1% to the insignificant class. This means that most of the keypoints will be discarded from the matching although the classifier is not quite sure about the membership of the keypoints (only for 50.1%). To address this problem, the matching performance is evaluated under different decision thresholds. Hereby, a keypoint is assigned to a certain class if the classifier predicts that the keypoint belongs to that class with a probability higher than a predefined threshold. Otherwise, the keypoint is considered to belong to the other class.

Saliency Map Considerations

A saliency map is a matrix which contains for each pixel in the image a corresponding saliency value falling in the range $[0,1]$. A saliency value of 1 indicates that the corresponding pixel is very conspicuous while a value of 0 indicates inconspicuous pixel. In turn, saliency map-based keypoint pruning decides whether a keypoint descriptor will be considered by the matching based on the corresponding saliency value. Accordingly, the saliency value (threshold) at which a keypoint is considered significant has a direct effect on the number of pruned keypoints, thus, it affects the matching performance.

Results

The effectiveness of the proposed approach for keypoint pruning is evaluated according to the achieved image matching performance. The results of our method are compared to three other matching approaches:

1. A baseline method, called "Random", that applies a random keypoint pruning according to a predefined threshold before matching the corresponding images.
2. A conventional matching approach, called "Full" matching, which does not apply keypoint pruning.
3. A matching approach, denoted as SM, which applies keypoint pruning according to visual attention and saliency maps.

The accuracy of the matching is measured in terms of precision and recall. To evaluate the matching precision, we used a subset of the Object Recognition Dataset [Nister and Stewenius, 2006] and created two non-overlapping groups of images: the *query* and the *document* groups which contain 100 and 200 images, respectively. We matched every image from the query group to every image in the document group according to each of the above described methods as well as according to our classification based approach. During the matching, two images were considered similar if they shared at least four keypoints. We calculated the precision for each query image and took the average. The results showed that a high precision of 99% is achieved by all approaches. This leads us to the conclusion that the matching precision has a negligible effect on assessing the effectiveness of image matching that applies keypoint pruning (the precision maintains a constant value under different approaches). In contrast, the matching recall as well as the keypoint reduction ratio (after a pruning method is applied) are crucial factors in judging the quality of the matching. On one hand, the matching recall indicates the ability of the matching algorithm to retrieve images relevant to a given input image. On the other hand, the keypoint reduction ratio gives a clue on the extent to which the matching can be made faster.

The matching recall and the ratio of reduced keypoints were evaluated using a test set consisting of 200 groups (different from the training set) of images taken from the Object Recognition dataset. From each group an image was randomly selected and matched to the other three images in the same group. In the case of classification-based keypoint pruning, the matching recall as well as the ratio of reduced keypoint were calculated using different decision thresholds on the significant class. For saliency maps different values for the saliency threshold were investigated.

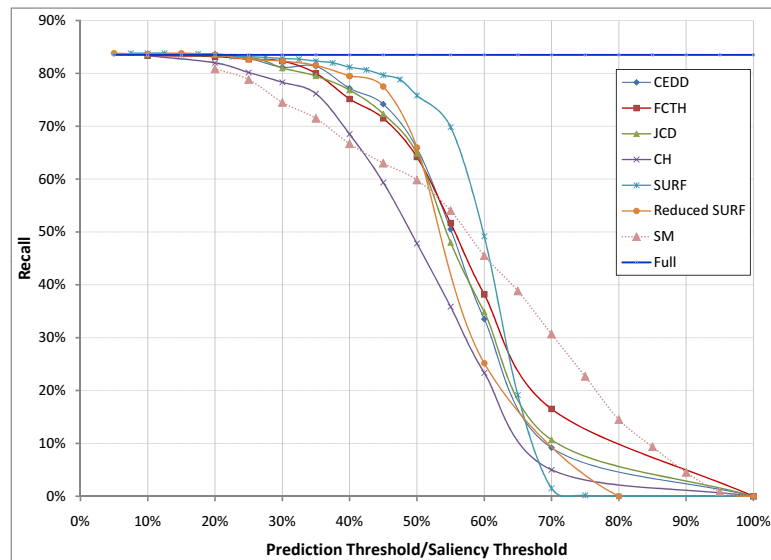


FIGURE 5.5: Recall of different keypoint pruning approaches as a function of the applied decision/saliency thresholds. The test dataset includes 200 groups from the Object Recognition dataset [Nister and Stewenius, 2006]

Figure 5.5 shows that for the different classification approaches, the matching recall increases to reach that of the full matching with a decreasing prediction thresholds. A lower prediction threshold means that more keypoints will be used by the matching and therefore a higher recall can be achieved. The same applies for the saliency threshold¹. That means, with a smaller saliency threshold less keypoints are pruned, so a higher recall can be achieved. With respect to ratio of reduced keypoints, with increasing decision/saliency thresholds, more keypoints are pruned, thus, higher keypoint reduction ratios are obtained.

To get a better insight into the relation between the matching recall and the ratio of reduced keypoints, Figure 5.6 shows the plot of the recall, relative to the recall of full keypoint matching (method 2), versus the ratio of reduced keypoints². First, it can be observed that, in general, our approach (using different classification features) as well as the SM approach for keypoint pruning evidently outperform the random approach. At a very low keypoint reduction ratio of 5% the random keypoint reduction causes a drop in the matching recall of more than 30%. With respect to the best performing approach, Figure 5.6 demonstrates that pruning the keypoints using a classifier trained with the Reduced SURF feature performs best followed by a classifier trained with the standard SURF descriptor. Both approaches provide a high keypoint reduction ratio of more than 40% while at the same time the corresponding drop in the recall stays below 5%.

¹Since the decision and the saliency threshold values fall in the same range, the two thresholds were shown on the same axis

²For better visibility only a subset of the investigated keypoint classifiers is shown.

Compared to classification approaches using SURF and Reduced SURF, the competitor method SM results in a higher drop in the recall, and that is even if a small keypoint reduction ratio is used. In fact, the performance of SM can be compared to that of our approach where the features FCTH, CEDD or JCD are used to train the keypoint classifier. In that case, all these approaches show close performance. For instance, they can reduce the amount of compared keypoints to more than 30% with a drop in the recall below 10%. Finally, the classifier which is trained with the color histogram feature (CH) showed the poorest performance.

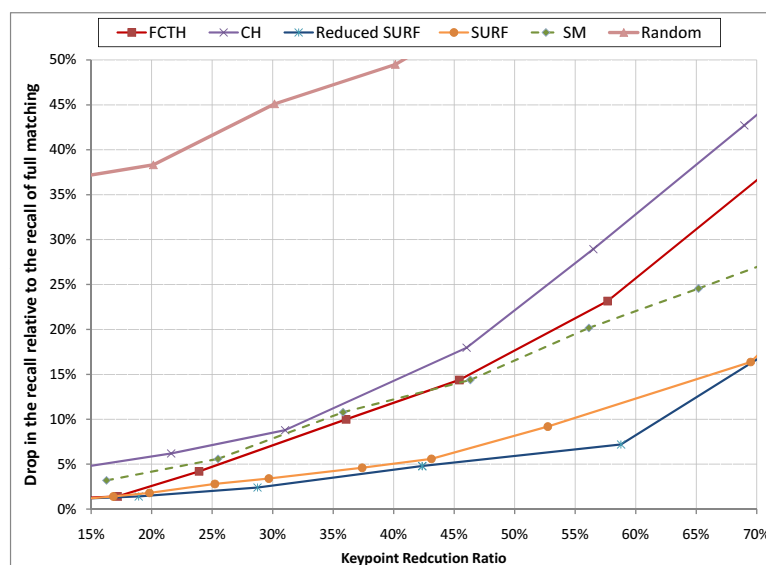


FIGURE 5.6: The drop in the matching recall (relative to the recall achieved by a full matching) as a function of the keypoint reduction ratio achieved by different keypoint pruning methods. The results are obtained from 200 groups of the Object Recognition dataset

5.4.3 Runtime Evaluation

Performing image matching while applying our classification-based approach for keypoint pruning consist of four steps:

1. Detecting keypoints in each image.
2. Describing the keypoints by features extracted from a patch centered on the keypoints.
3. Filtering the keypoint based on their predicted labels.
4. Matching a reduced set of the keypoints.

Similarly, image matching that applies visual attention-based keypoint pruning applies the following procedure:

1. Generating the saliency map for the input images.
2. Detecting keypoints in salient regions.
3. Filtering the keypoint according to their saliency values.
4. Matching a reduced set of the keypoints.

For many applications, the first three steps of both approaches can be performed offline. For example, in CBIR systems, these steps can be executed as background processes before inserting a new image into the database. On the contrary, the matching step must be done online.

The benefit of keypoint pruning on reducing the matching runtime can be estimated analytically as follows. Suppose that we have two images I and J . Let N and M be the numbers of keypoints extracted from I and J , respectively. For a matching algorithm such as the Linear Nearest Neighbor (LNN) [Knuth, 1998], finding correspondences between I and J implies calculating the pairwise distances between the keypoint descriptors of both images. Refer with C to the matching cost. C can be estimated in terms of the total number of compared keypoint descriptors. According to the LNN approach, this cost is given by:

$$C = N \times M \quad (5.3)$$

Now, suppose that the *average* keypoint reduction ratio of a keypoint pruning approach is α . Then the matching cost after applying the pruning C' can be estimated as:

$$C' = (1 - \alpha)N \times (1 - \alpha)M \quad (5.4)$$

Consequently, the order in which the matching with keypoint pruning is faster than the full matching is calculated as:

$$\text{Runtime Ratio} = \frac{C}{C'} = \frac{1}{(1 - \alpha)^2} \quad (5.5)$$

Experimentally, we evaluated the runtime requirements of each of the presented approaches using a ground truth dataset created from a personal collection of images with image resolution up to 3264×1840 pixels and about 1 MB per image on average. This dataset was selected for two reasons. First, it enables us to investigate runtime requirements with a "realistic" dataset, which is not specifically designed for a particular

retrieval task. Second, this dataset allows us to determine how our keypoint pruning approach generalizes to image datasets unrelated to the one that was used in the training. The dataset contains 27 groups and each group has seven images on average. Images in the same group depict the same scene from different perspectives, at different scales and under different illumination conditions (Figure 5.7).



FIGURE 5.7: A sample of the image groups of our manually created dataset. The images have a high resolution (3264×1840 pixels) and an average volume of 1 MB

For each image in the collection, we extracted and pruned the keypoints using our approach as well as saliency maps. Table 5.1 summarizes the runtime taken by each of the pre-matching processing steps averaged on all images in the dataset. The test was carried out using a computer with Intel(R) CORE i5 and 8GB RAM.

Task	Average Runtime (sec)
Keypoint Extraction	4.14
SM Generation and Filtering	2.07
SURF Prediction and Filtering	3.95
CEDD Prediction and Filtering	5.89
FCTH Prediction and Filtering	5.75
JCD Prediction and Filtering	6.03

TABLE 5.1: Average runtime of the pre-matching (offline) phases

Furthermore, we evaluated the matching accuracy as well as the runtime ratio (Equation 5.5) on the new dataset. Similar to the process followed in the last section, we randomly selected an image from each group and matched it to the other images in the same group. Figure 5.8 shows the trade-off between the drop in the recall and the percentage of the

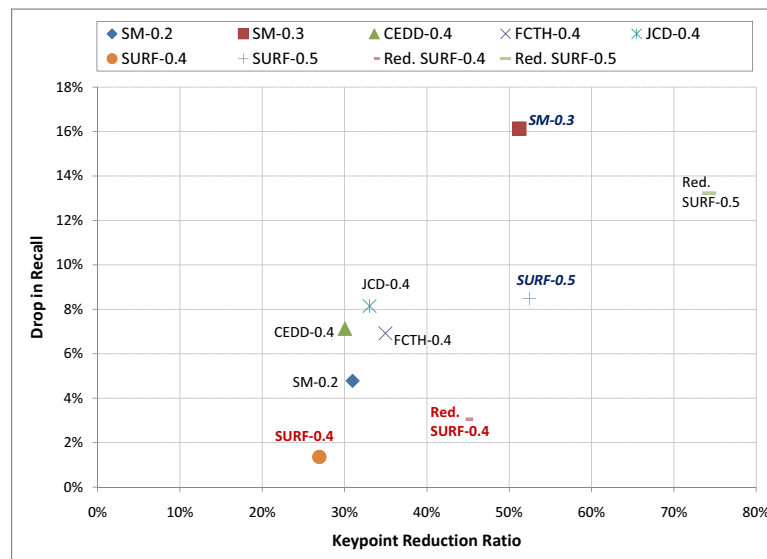


FIGURE 5.8: The drop in the matching recall (relative to the recall achieved by a full matching) as a function of the keypoint reduction ratio achieved by different keypoint pruning methods. The results are obtained using a manually created test dataset (Figure 5.7)

reduced keypoints achieved by our approach and the SM method. We performed the test by using eight different configurations of our approach (in Figure 5.8 the numbers on the right of the classification features correspond to the used prediction thresholds). Furthermore, we compared our approach to saliency map keypoint pruning with saliency thresholds values of 0.2 and 0.3 (denoted as SM-0.2 and SM-0.3 in Figure 5.8). First, it can be observed that for the new dataset, the trade-off between the keypoint reduction ratio and the drop in the matching recall is in accordance with the results obtained from using a subset of the Object Recognition dataset (Figure 5.6). Similarly, at keypoint reduction ratio in the range $[30\%, 40\%]$ the drop in the recall does not exceed 10% disregarding the applied pruning method. Furthermore, the results show the superiority of the Reduced SURF and the SURF features. For example, Reduced SURF-0.4 (i.e., a classifier trained with the Reduced SURF feature and using a threshold of 0.4 as a decision threshold) achieved a keypoint reduction ratio of more than 45% with a drop of only 3% in the matching recall. Moreover, the results emphasize again that our approach outperforms the saliency map approach. For instance, the configurations SM-0.3 and SURF-0.5 (shown in italic in Figure 5.6) both select less than 50% of the keypoints for the matching. However, the saliency map approach (SM-0.3) results in twice as much drop in the recall as in the case of our approach (SURF-0.5).

In the same experiment, we also reported the actual matching runtime and averaged it according to the total number of the performed image matching operations. This is done using full matching as well as matching with keypoint pruning. The actual as well as

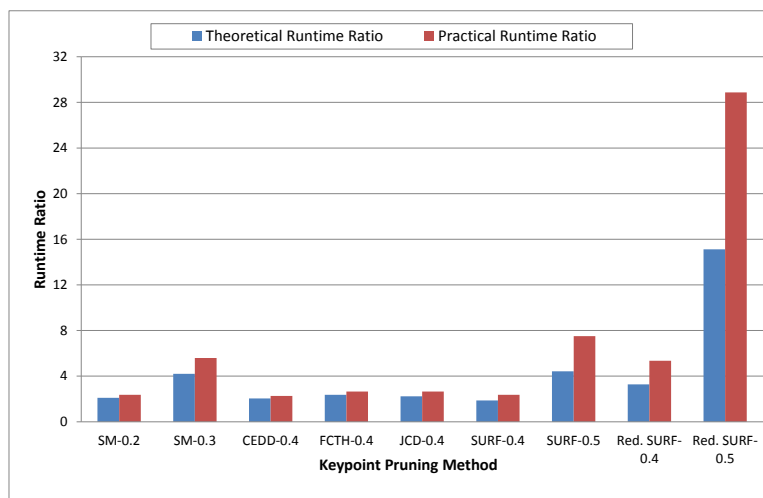


FIGURE 5.9: Comparison between the theoretical and the actual (practical) runtime ratio under different keypoint pruning configurations. The advantage of image matching using keypoint pruning is measured by how much it is faster than the full matching. The x-axis corresponds to the analyzed keypoint configurations. The y-axis represents the ratio between the runtime (the theoretical as well as the practical one) required by keypoint pruning-based image matching and that which is required by the full matching

the theoretical runtime ratio (Equation 5.5) for each method is reported in Figure 5.9. The graph shows the advantage of keypoint pruning on reducing the matching runtime. In general, all investigated approaches can speed up the matching by at least two times compared to the full matching. Indeed, the improvement in the runtime is directly proportional to the ratio of reduced keypoints. For instance, by considering the actual runtime time ratio, the method "Red. SURF-0.5" is around 30 times as faster as the full matching. Figure 5.9 also shows that in all cases the actual runtime ratio is higher than the theoretical one. This can be justified according to the Laplacian sign check which is performed while matching SURF descriptors [Bay et al., 2008]. The SURF algorithm avoids comparing two descriptors when they have different signs, so that faster matching can be achieved. The theoretical analysis, however, assume that all descriptors will be compared disregarding their signs.

5.5 Summary

In this chapter, we presented a classification-based approach for characterizing SURF keypoints. The ultimate goal is improve SURF-based image matching by reducing the number of compared SURF descriptors to the subset of descriptors corresponding to most salient keypoints. For this purpose, we investigated different image features for characterizing SURF keypoints. Furthermore, we compared the performance of our

method to another approach for keypoint pruning based on saliency maps and visual attention. The results show that our approach outperforms the adversary and is able to provide high keypoint reduction ratio with a slight drop in the matching recall. Furthermore, the evaluation shows that the execution time of image matching can be efficiently improved using the presented approach.

Chapter 6

SURF-based Iterative Image Matching

In this chapter, an approach for improving the accuracy of SURF-based image matching is presented. This approach aims at improving the matching recall by iteratively applying the matching process. To reduce the computation complexity, an efficient method for image clustering is presented.

6.1 Introduction

In the previous chapter, we explained the main limitation of image matching using local features due comparing huge amounts of keypoint descriptors. In this chapter, we deal with issues related to SURF and its capacity to discover similar images with various distortions in the matched images. In fact, SURF shows a poor ability to discover similarities between images when they are distorted, due to extreme rotation or illumination changes. In the terms of information retrieval this leads to a low *recall*. To get better insight into this problem, consider the example shown in Figure 6.1. In this figure, we are using SURF to find similar images for an input photo, Figure 6.1(a), (church Notre Dame de Paris) in a given collection of images. SURF succeeded to find a match for the input image as shown in Figure 6.1(b). However, due to high distortions, SURF could not discover that the other two images shown in Figures 6.1 (c)-(d) are also similar to the input image. Such scenarios are typical to SURF-based image matching and they represent a serious disadvantage for our image annotation approach. As mentioned in Chapter 1, tags of the visual neighbors represent the candidate annotations, from which tag proposals are mined. Consequently, if the associated images cannot be discovered,

the corresponding tags will be excluded from the mining process. Hence, the quality of the automatic image annotation approach can be negatively affected.

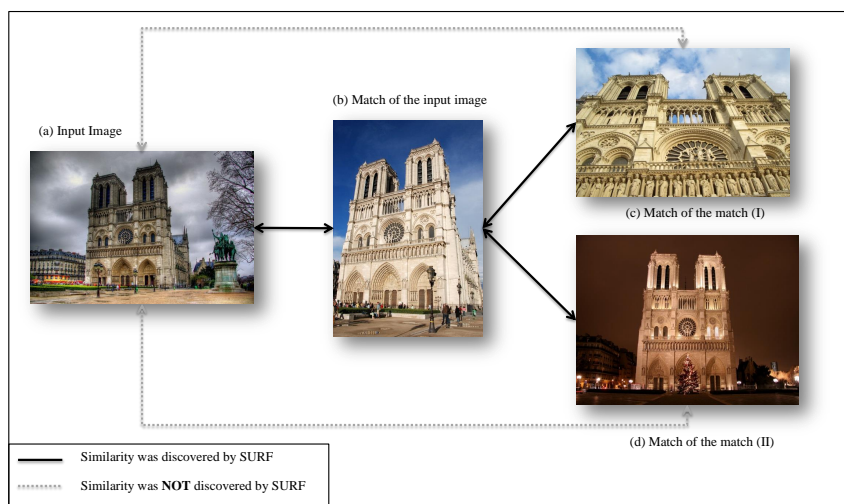


FIGURE 6.1: Example: image matching using SURF

In the next sections, we present our proposal for dealing with this problem through an iterative matching approach. First, a detailed description of our solution is given. After that, the efficiency of the proposed approach is evaluated experimentally.

6.2 Iterative Image Matching

Our method for improving the matching *recall* is based on *iterative* application of the matching algorithm. Thereby, images found in an initial matching step are used as input for further matching phases. The idea can be directly inferred from the example of Figure 6.1. To obtain the missed visual neighbors, we can use those which has been already identified (e.g. Figure 6.1 (b)) to query the image collection to determine further possible matching images. The result of this step can then be added to the list of visual neighbors. This process can be repeated as long as further similar images are discovered.

Obviously, this iterative process implies additional computations. To address this challenge, we propose a solution which applies a clustering strategy to reduce the number of images, which are used as input for the next matching phases. First, let us consider the following representation, in which the results of the iterative matching approach are modeled using a tree data structure (Figure 6.2(a)). The root of the tree corresponds to the input image (e.g. the one we want to automatically annotate). Images that are visually similar to the input image are connected to it via edges. Next, each of the identified visual neighbors is used to initiate a further matching phase. The new visual neighbors

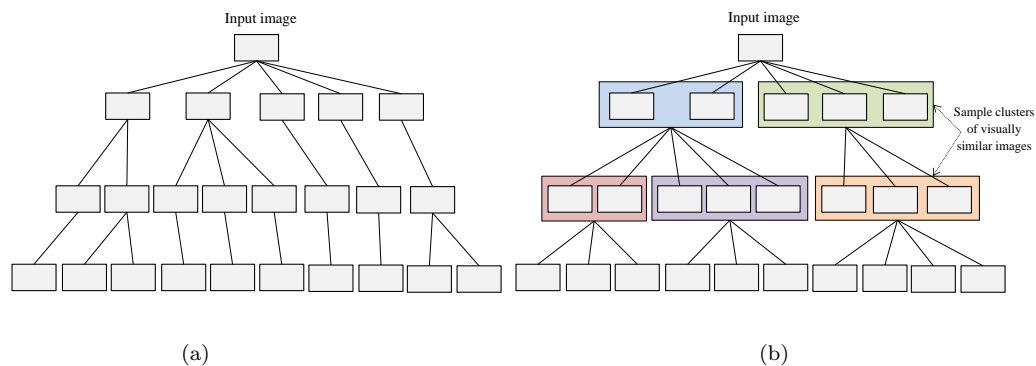


FIGURE 6.2: (a) A tree representation for a collection of images which are similar to a given input image (the root). The images are obtained through an iterative application of the image matching algorithm on each image at each level of the tree. (b) To reduce the computation cost, images at each level in the tree are first clustered and a representative image of each cluster is used to identify further possible matching images

are then connected to the initiating image. Thereby, each image that already has a match is excluded from the search space. Under the assumption that visual similarity is transitive and by applying the described matching process several times, we obtain a tree data structure which represents the relationships between initial input image (the root of the tree) and the complete set of visual neighbors.

This naive approach of building the tree implies intensive computation. To improve the performance, imagine that for each tree level starting from the first one (the first level corresponds to the direct visual neighbors of the input image), we can efficiently determine clusters of similar images (Figure 6.2(b)). Consequently, a better performance can be achieved if a representative image of each cluster is used to initiate the next matching phase, instead of using the whole collection of images at that level.

To cluster the images of a given tree level, we exploit the similarity information between those images and their common father node. Let $\mathcal{L} = \{L_1, \dots, L_n\}$ be the set of images at a given level of the tree with the common father indicated as I_{in} . To identify clusters of similar images inside \mathcal{L} , the elements of \mathcal{L} have to be matched to each other. To reduce the computation, we take advantage of the already calculated pairwise similarities between the father node I_{in} and the elements of \mathcal{L} . Accordingly, two images $L_1, L_2 \in \mathcal{L}$ are considered similar if the father node I_{in} shares the same or overlapping sets of keypoints with each of them.

Formally, we define KP a function that returns for a given image I the set of keypoints kp_i which have been extracted from it using the SURF algorithm:

$$KP(I) = \{kp_1, \dots, kp_n\} \quad (6.1)$$

Additionally, for two keypoints $kp_i \in KP(I)$ and $kp_j \in KP(J)$ extracted from the images I and J , respectively, we use the equal sign to indicate that they are a "correspondence". Furthermore, we define CKP , a function which gives for two images I and J the set of keypoints of I that have correspondences in the set of keypoints of J :

$$CKP(I, J) = \{kp_i | kp_i \in KP(I) \wedge \exists kp_j \in KP(J) \wedge kp_i = kp_j\} \quad (6.2)$$

Now, the similarity between the two images $L_1, L_2 \in \mathcal{L}$ can be determined in terms of the similarity to the father node I_{in} as follows:

$$sim(L_1, L_2 | I_{in}) = |CKP(I_{in}, L_1) \cap CKP(I_{in}, L_2)| \quad (6.3)$$

The presented similarity calculation approach has a lower complexity than the standard approach. In the standard approach, keypoints descriptors of both images have to be compared to each other. Suppose that, on average, n keypoints can be extracted from each image (according to the size of the image n can be estimated in hundreds to thousands of keypoints). Subsequently, the complexity of finding common keypoint correspondence between two images is given by $\mathcal{O}(n^2)$. Concerning our approach (Equation 6.3), the cost of determining the similarity between two images is reduced to finding the intersection between two subsets of keypoints of the common father node. In practice, the average cardinality of those subsets, m , is much smaller than the average number of the keypoints extracted from each image, i.e., $m \ll n$. Accordingly, the complexity of our approach in worst case is $\mathcal{O}(m^2)$, which is lower than the complexity of the standard approach discussed above, i.e., $\mathcal{O}(m^2) \ll \mathcal{O}(n^2)$ since $m \ll n$.

Finally, after the pairwise similarities between the elements of \mathcal{L} have been calculated, the produced similarity matrix can be fed into a clustering algorithm to build the clusters. In contrast to the naive approach, the time needed by the clustering represents a further computational cost for our method. However, our experiments show (see next Section) that the impact of the clustering on the runtime is low.

6.3 Experimental Evaluation

6.3.1 Dataset

To evaluate the recall of image matching under the iterative approach, a ground truth dataset is required. We built a dataset of groups of similar images based on manual search using Google Image Search [Google, 2014], Flickr and a subset of the European

Cities 50K dataset [Avrithis et al., 2010]. Each group of the dataset contains images depicting the same scene; however, they are taken from different perspective and under different levels of illumination. Additionally, the image groups cover different categories, such as outdoor, indoor, buildings, etc. (Figure 6.3). The final dataset consists of 69 image groups, and each group 70 images on average.

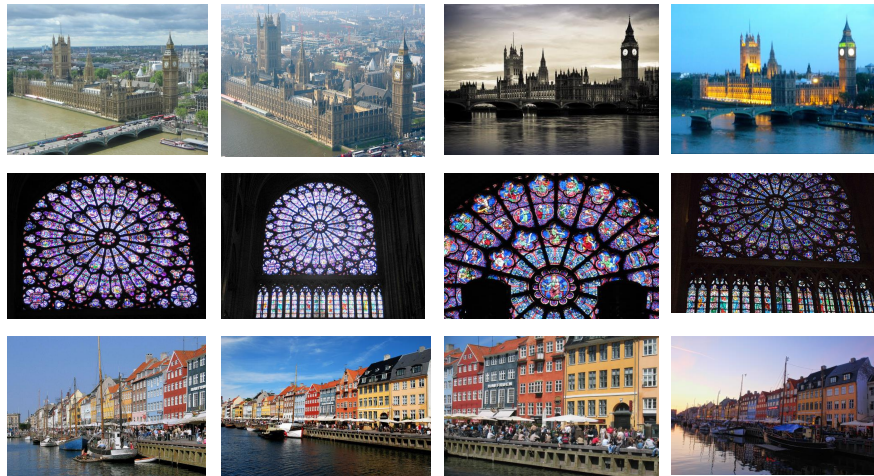


FIGURE 6.3: A sample from the image dataset which we used to evaluate the iterative image matching approach. Each row corresponds to a subset of visually similar images

6.3.2 Evaluation Methodology

Our goal is to investigate the effect of the proposed iterative image matching on the matching recall. Additionally, we want to evaluate the runtime performance when the introduced image clustering is integrated into the matching process with that of the naive approach, which does not apply clustering.

To achieve this, the similarity between two images has to be defined. As discussed before, this is done based on the number of keypoint correspondences between the matched images. In the literature, it has been shown that a threshold of five keypoint correspondences is adequate to consider that two images are similar [Ke et al., 2004, Jones et al., 2010]). We also investigated the same problem and analyzed the effect of different threshold values on the matching precision. For this purpose, we took an image from each of the 69 groups of the dataset and matched it to a collection of 2,000 dissimilar images. We then evaluated the rate of false positives at different thresholds of keypoint correspondences. The results showed that a value of four common keypoints led to 100% precision (similar to the results presented in Section 5.4.2 of the previous chapter). We used this value as a similarity threshold and applied the following procedure to evaluate the matching recall. For each group of the dataset, an image is selected randomly and

matched with the remaining images of the same group. Consequently, the recall is defined by the percentage of the identified similar images. Finally, the average matching recall is calculated over the complete collection of image groups in the test dataset.

6.3.3 Results

We compared our approach which applies image clustering according to the proposal described in the last section to the naive method which considers all images found by a previous matching iteration as input for further matching phases. We used the standard *agglomerative* clustering approach [Kaufman and Rousseeuw, 1990] to cluster the images at each level of the tree. Agglomerative clustering is convenient in this case since there is no need to fix the number of the clusters. After building the clusters, a representative image is selected randomly from each cluster and is used to initiate a further image matching phase.

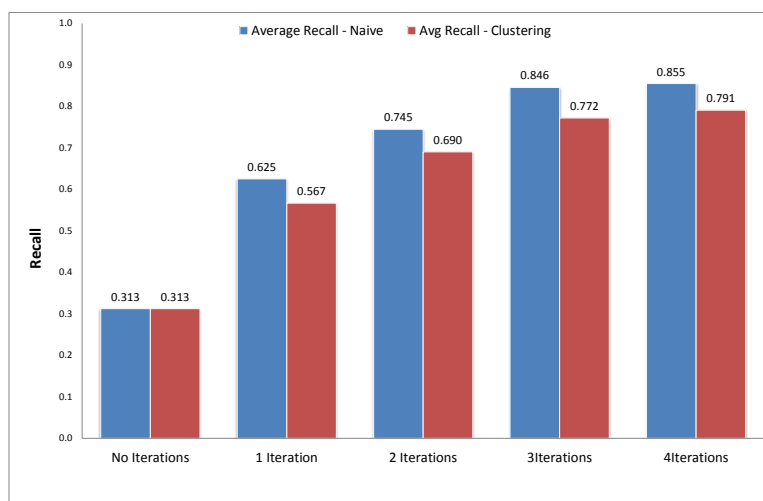


FIGURE 6.4: The average recall achieved by the naive as well as the clustering-based approaches at different matching iterations

Figure 6.4 shows the average matching recall for zero (corresponds to a single matching without iterations) to four iterations. It can be seen, that the iterative approach results in significant recall improvement with an increasing number of iterations. Furthermore, the experiments show that for 76% of the image groups, three iterations were enough to achieve the maximum recall. However, the clustering based approach shows slight drop in the average recall compared to the naive approach.

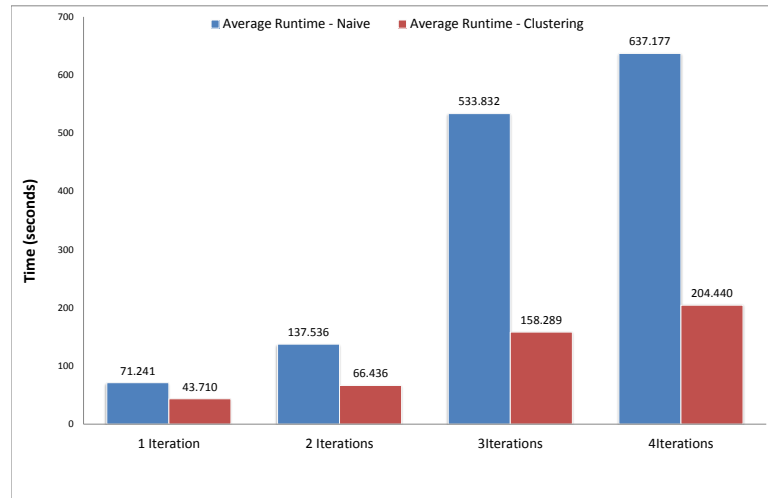


FIGURE 6.5: Clustering-based vs. naive iterative matching: Average matching runtime required by each approach at different matching iterations

With respect to the required runtime¹, Figure 6.5 shows that the clustering results in significant performance gain. Iterative image matching using clustering requires in general less time than the naive method. Furthermore, with increasing number of iteration, the difference in the performance between the two methods becomes more evident. For instance, by applying four iterations the matching with clustering is three times faster than the naive approach.

Finally, it is important to note that the utility of the clustering is dependent on the number of image at each level of the tree (Figure 6.2(a)). With a small number of images, the clustering might lead to additional overhead which makes it less efficient than the naive approach. In this case a hybrid approach can be more suitable. This can be achieved by setting a lower bound on the number of images required to trigger the clustering; otherwise the naive approach is applied. Furthermore, the choice of the clustering algorithm might also affect the performance of the iterative approach, thus, it deserves further investigation.

6.4 Summary

In this chapter, a method for improving the recall of image matching based on SURF is introduced. The method extends the traditional matching through additional matching phases. To reduce the computational overhead, an approach for image clustering is proposed. The efficiency of the presented method was experimentally evaluated. The

¹The experiments were performed on machine with Intel Core i5 CPU with 4x processor, 2,40 GHz and 8 GB RAM and

results show that our method leads to a significant improvement in the matching recall. Furthermore, the evaluation demonstrates that the proposed image clustering approach outperforms the naive approach in terms of the required runtime.

Part IV

Tag Ranking and Global Evaluation

Chapter 7

Probabilistic Model for Tag Ranking

This chapter features a probabilistic model for tag ranking based on Bayes' rule. The model combines data from different modalities, i.e., image contents, tags and user information to rank tag proposals according to their relevance to the to-be-annotated image.

7.1 Introduction

As we have seen in Chapter 2, ranking candidate annotations according to their relevance to the target image is an essential component for the majority of AIA approaches. Indeed, the applied ranking method is a crucial factor to judge the quality of AIA solutions. This thesis addresses the problem of tag ranking by proposing a statistical model for combining the information extracted by the previous phases of the proposed annotation approach (refer to Section 1.4). The model grounds on Bayes' rule [Gelman et al., 2003] and provides a scalable framework for combining information extracted from different modalities to score candidate annotations.

7.2 Problem Statement

For sake of clarity, a simplified version of our automatic annotation approach with a focus on the tag ranking phase is illustrated in Figure 7.1. Let I_{in} be a geotagged image, which we want to annotate using our approach. First, a repository of community tagged images is queried to identify the geographical neighbors $\mathcal{I}_{geo} = \{I_1, I_2, \dots, I_m\}$

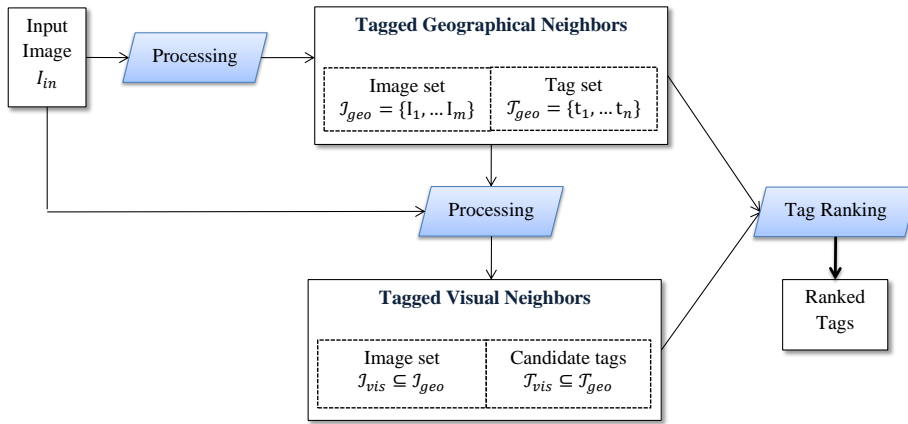


FIGURE 7.1: A simplified model showing the input for the tag ranking phase of our image annotation approach

of the target image I_{in} . We indicate with $\mathcal{T}_{geo} = \{t_1, t_2, \dots, t_n\}$ a lexicon from which the tags of the geographical neighbors were selected. By ignoring user interaction, a simplified folksonomy can be obtained based on the sets \mathcal{I}_{geo} and \mathcal{T}_{geo} . The folksonomy can be represented as a bipartite graph $G(V, E)$ with two vertex classes. The first class corresponds to the geographical neighbors, \mathcal{I}_{geo} , while the other one corresponds to the associated tags, \mathcal{T}_{geo} , such that $V = \mathcal{I}_{geo} \cup \mathcal{T}_{geo}$. An edge $e_{ij} \in E$ indicates that the tag $t_i \in \mathcal{T}_{geo}$ was used to annotate the image $I_j \in \mathcal{I}_{geo}$ (Figure 7.2.(a)).

Now, the problem of mining tag proposals for the input image I_{in} can be formulated by means of the folksonomy graph. For this purpose, we extend the graph by an additional vertex corresponding to the input image. Subsequently, tag proposals for the input image can be provided, if we are able to infer a kind of *virtual* edges between the vertex of the input image and the tag vertices. The virtual edges cannot be derived directly, however, the subset of geographical neighbors which are visually similar to the input image (visual neighbors), denoted as \mathcal{I}_{vis} can be used to establish the connection (Figure 7.2.(b)). In the optimal case, the input image I_{in} can be indirectly connected to a tag $t_i \in \mathcal{T}_{geo}$ if there is an image $I_j \in \mathcal{I}_{geo}$ annotated with t_i and at the time visually identical to I_{in} , i.e., I_j is a replica of I_{in} . However, this assumption is too strict and to loosen it, we assume that I_{in} can be linked to t_i if the two images I_{in} and I_j are visually similar. Finally, to determine the degree to which t_i suits I_{in} , a ranking system that combines and quantifies information along the paths between the two nodes is required. In the next section, we describe our proposal in this regard.

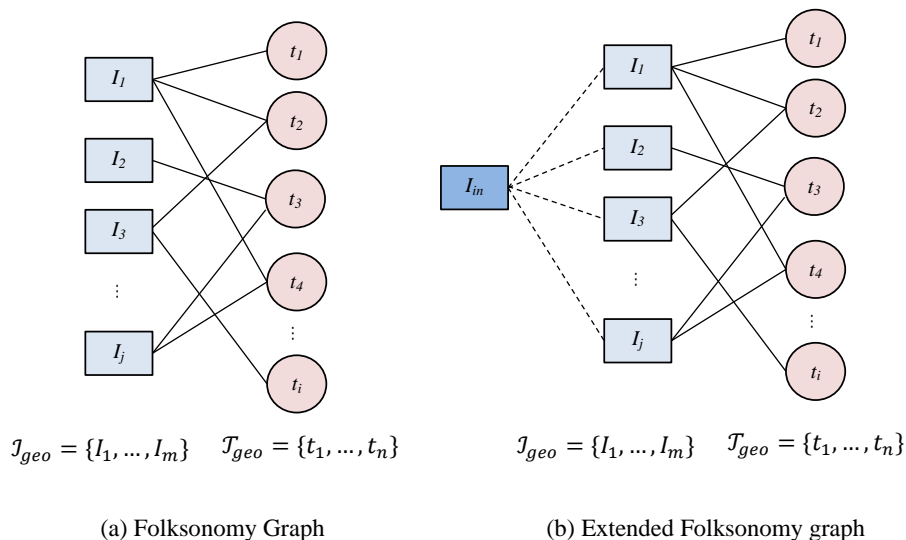


FIGURE 7.2: (a) A Graph representation of a folksonomy corresponding to annotated images found in the geographical proximity of an input image I_{in} . (b) The graph extended by an additional vertex corresponding to the input image. The dashed edges connecting the input image to the geographically close images correspond to the visual similarity

7.3 Pseudo-generative Statistical Model

We frame the tag ranking problem within a probabilistic model based on Bayes' rule (a.k.a. Total Probability Law) [Gelman et al., 2003]. Generally, Bayes' rule is used in settings, where from a generative model linking some causes to some effects, one wants to get the probability of the effects given a cause. In general, the causes are supplied with a priori probabilities. The generative model is typically represented by means of a tree, where the leaves correspond to the effects (or by means of a more compact graph where the effects can be shared among different causes). Complying with this paradigm, we can identify, for sake of convenience, the input image I_{in} with the root of the tree, and the tags of the geographical neighbors with the leaves. The geographical neighbors play the role of intermediate nodes, "caused" by the root and "causing" the leaves.

Within this pseudo-generative model the input image I_{in} can be thought of to "yield" each of the geographical neighbors $I_j \in \mathcal{I}_{geo}$ with a probability $P(I_j|I_{in})$. In turn, each image $I_j \in \mathcal{I}_{geo}$ can be thought of to "yield" the tags $t_i \in \mathcal{T}_{geo}$ with probability $P(t_i|I_j)$. This model in a compact version is shown in the graph of Figure 7.3.

Assume that the values of the conditional probabilities $P(I_j|I_{in})$ and $P(t_i|I_j)$ are available. We can then define the strength of the relationship between the input image I_{in} and a candidate tag t_i to be equal to the conditional probability $P(t_i|I_{in})$. This latter probability, can be calculated by applying Bayes' rule as given by the following formula:

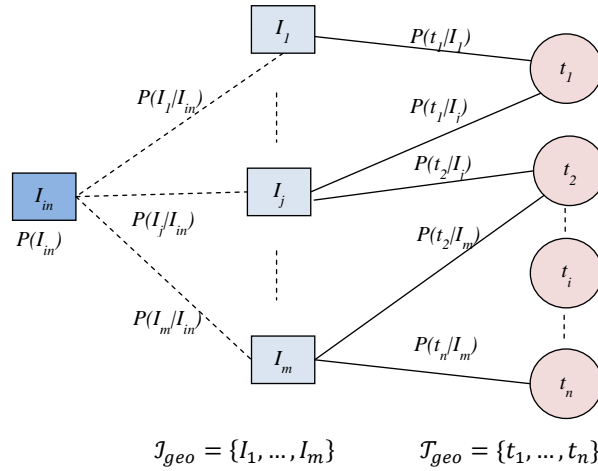


FIGURE 7.3: A Bayesian model for tag ranking: I_{in} represents the input image, $I_j \in \mathcal{I}_{geo}$ are the geographical neighbors and $t_i \in \mathcal{T}_{geo}$ is the set of the associated tags

$$P(t_i|I_{in}) = \sum_{j=1}^m P(I_{in})P(I_j|I_{in})P(t_i|I_j) \quad (7.1)$$

$P(I_{in})$ is the prior distribution of an image and it is uniform and identical for all the images, thus, it can be eliminated from Equation 7.1, so the final tag ranking model is defined as:

$$P(t_i|I_{in}) \approx \sum_{j=1}^m P(I_j|I_{in})P(t_i|I_j) \quad (7.2)$$

In Equation 7.1, the value of $P(t_i|I_{in})$ corresponds to the sum over all the paths leading from the input image (the root of the tree) to the tags $t_i \in T$ (the leaf nodes). Thereby, the probability of an individual path is computed by chain product of the corresponding probabilities along that path (refer to Figure 7.3).

Note that, by construction, only tags that are reachable from the input image are considered as candidate tags. The probabilities $P(t_i|I_{in})$ of the candidate tags provide a natural ordering of the importance/relevance of the tags to the input image.

Discussion

The assumption of the availability of the values of the conditional probabilities $P(I_j|I_{in})$ and $P(t_i|I_j)$ needs further consideration. Although it is a strong assumption, it can be easily relaxed. Strictly speaking, neither the generative process from the input image I_{in} to the images in \mathcal{I}_{geo} nor the generative model from the elements of \mathcal{I}_{geo} to the tags \mathcal{T}_{geo} are known or defined precisely. Hence, the above conditional probabilities cannot be known exactly. However, we are not interested in the probability values per se. We

are rather interested in using those values as indicators for ranking the candidate tags according to their relevance to the input image. Therefore, even quantities proportional to (or simply monotonically dependent on) those probabilities will suite the task since they will not change the ordering. Furthermore, if the probability gaps between pairs of images relevant to an input image, denoted as $P(I_j|I_{in}) - P(I_k|I_{in})$ with $I_j, I_k \in \mathcal{I}_{geo}$, are wide enough, then even slightly distorting functions or indicators correlated with the $P(\cdot|I_{in})$ – a *proxy* of $P(I_j|I_{in})$, can suite the task. In a similar way, the proxy for $P(t_i|I_j)$, denoted as $P(\cdot|I_j)$, can be used in place of the actual probability if the probability gaps between pairs of candidate tags, denoted as $P(t_i|I_j) - P(t_l|I_j)$ with $t_i, t_l \in \mathcal{T}_{geo}$, are also wide enough.

For these reasons, even if the conditional probabilities $P(I_j|I_{in})$ and $P(t_i|I_j)$ are not directly available to us, we will adopt the above described ranking procedure. The proxy values for $P(I_j|I_{in})$ will be defined based on the visual similarity between the two images, i.e., the image-to-image relevance (IIR). While the proxies for $P(t_i|I_j)$ will be given by a measure for word-to-image relevance (WIR). The next sections describe in detail our procedure for determining these quantities.

7.3.1 Estimating Image-to-Image Relevance: $P(I_j|I_{in})$

The term $P(I_j|I_{in})$ represents the probability that the input image I_{in} generates the geographical neighbor I_j . In this thesis, we build our estimation based on the visual similarity between the two images. Suppose there is a function Sim which computes the visual similarity between two images. Consequently, a proxy of the probability $P(I_j|I_{in})$ can be obtained by normalizing the visual similarity between I_j and I_{in} according to the total similarity between I_{in} and its geographical neighbors \mathcal{I}_{geo} as follows:

$$P(I_j|I_{in}) = \frac{Sim(I_j, I_{in})}{\sum_{k=1}^m Sim(I_k, I_{in})} \quad (7.3)$$

In our approach, the similarity function Sim is defined in terms of the number of SURF keypoints correspondences between the two images.

Note that $P(I_j|I_{in})$ can be estimated by using or combining further information, such as, temporal information or the geographical distance between the locations of capture of the two images (e.g [Moxley et al., 2008]). However, thanks to advances in computer vision, available algorithms provide robust means for identifying similar images. As we have seen in the last two chapters, SURF is able to identify images which depict similar scenes with high accuracy. Accordingly, a good estimate for $P(I_j|I_{in})$ can be achieved when only the visual similarity between the two images is considered.

7.3.2 Estimating Word-to-Image Relevance: $P(t_i|I_j)$

According to Bayes' rule, the term $P(t_i|I_j)$ refers to the probability of generating the tag t_i based on the image I_j . In the context of our image annotation approach, this can be interpreted as the degree of relevance/importance between each tag of our simplified folksonomy (Figure 7.2) and the associated images. In what follows, we present two methods for deriving a proxy value for $P(t_i|I_j)$.

Weighted Voting

A first derivation for a proxy value for this probability can be directly inferred from the model shown in Figure 7.3. If I_j is annotated with t_i (i.e., they are connected in the corresponding folksonomy graph), then by a simple application of frequentist statistics, the proxy for $P(t_i|I_j)$ can be given as:

$$P(t_i|I_j)_{\text{weighted voting}} = \frac{1}{|\{t_k|t_k \text{ annotates } I_j\}|} \quad (7.4)$$

In Equation 7.4 the descriptive power (relevance) of t_i to the image I_j is defined as the inverse of the total number of tags annotating I_j . Accordingly, all tags annotating I_j are given a uniform importance which is inversely proportional to the total number of tags.

The name "weighted voting" for this measure can be explained if we consider for a given tag t_i the sum of importance values $\sum_{j=1}^m P(t_i|I_j)$. The sum can be interpreted as a weighted voting since for a single image I_j the value $P(t_i|I_j)$ can be considered as a vote given by I_j about the importance of t_i . Consequently, the global importance of the tag is represented by the sum of all votes contributed by the complete set of images $I_j \in I_{geo}$.

TF-IDF-based Approach

Another way to measure tag importance can be derived based on the idea of TF-IDF (Term Frequency-Inverse Document Frequency) [Baeza Yates et al., 1999]. In a similar manner to the proposal of [Naaman et al., 2007], our method exploits the geographical context as well as visual similarity information. In general, the same tag is used only once to annotate a given image. Therefore, one to one correspondence between an image and a document results in the same term frequency for all tags. To address this problem, we define a document as the subset of the visual neighbors $\mathcal{I}_{vis} \subseteq \mathcal{I}_{geo}$ (refer to Figure

7.1). Subsequently, the term frequency of a tag t corresponds to the cardinality of the subset $\mathcal{I}_{vis}^t \subseteq \mathcal{I}_{vis}$ which contains images annotated with t . Finally, the term frequency (TF) of t is given according to Equation 7.5:

$$TF(t) = \frac{|\mathcal{I}_{vis}^t|}{|\mathcal{I}_{vis}|} \quad (7.5)$$

To calculate the inverse document frequency (IDF) for t , we follow the traditional approach, i.e., each image $I \in \mathcal{I}_{geo}$ is considered as a document. Let $\mathcal{I}_{geo}^t \subseteq \mathcal{I}_{geo}$ indicates the subset of geographical neighbors which are annotated with t . Consequently, the IDF of t can be calculated as follows:

$$IDF(t) = \log\left(\frac{|\mathcal{I}_{geo}|}{|\mathcal{I}_{geo}^t|}\right) \quad (7.6)$$

To make the value of the TF-IDF corresponds to a probability, it must be normalized in the range [0,1]. Since the value of the TF is already in that range, we need to normalize the value of the IDF. To achieve this, we identify the minimum and the maximum IDF values, which are denoted as $\min(IDF)$ and $\max(IDF)$ respectively, for all candidate tags and apply a linear transformation as given below:

$$IDF_{norm}(t_i) = \frac{IDF(t_i) - \min(IDF)}{\max(IDF) - \min(IDF)} \quad (7.7)$$

Finally, the TF-IDF value is used as a proxy for the term $P(t_i|I_j)$:

$$P(t_i|I_j)_{TF-IDF} = TF(t_i) \cdot IDF_{norm}(t_i) \quad (7.8)$$

Note that according to the presented TF-IDF approach, the importance of a tag is independent of the individual images $I_j \in \mathcal{I}_{geo}$. Indeed, the TF-IDF value of a tag is determined globally for the folksonomy based on the input image and the sets \mathcal{I}_{geo} and \mathcal{I}_{vis} .

Integrating User Influence

The popularity of tags among the users, who contributed them, can directly affect their importance as candidates annotations. For example, tags which are used by few users are usually too specific or could be misspelled. Hence, they should be discarded or, at least, given a low importance.

Here, we followed the common way to quantify tag popularity, however, we did that by exploiting location information. In other words, we determine a reputation score for each tag according to its popularity among users who used it to annotate images taken in a certain place. Let \mathcal{U}_{geo} be the set of users who produced the set \mathcal{I}_{geo} of images in a given geographical location. Additionally, assume that the set $\mathcal{U}_{geo}^t \subseteq \mathcal{U}_{geo}$ consists of the users who used the tag t in their annotation task. Accordingly, the popularity (*reputation*) of the tag t , referred to as $R(t)$, can be determined according to the proportion of unique users who used it, as given below:

$$R(t) = \frac{|\mathcal{U}_{geo}^t|}{|\mathcal{U}_{geo}|} \quad (7.9)$$

Finally, the introduced methods for calculating tag importance can be further refined by scaling tag importance values according to the popularity of the corresponding tags as given in the following formula:

$$P(t_i|I_j) = P(t_i|I_j)_{\text{weighted voting/TF-IDF}} \times R(t_i) \quad (7.10)$$

7.4 Summary

In this chapter an approach for tag ranking is proposed. For this purpose, a statistical model based on Bayes' rule is introduced and solutions for estimating its components are provided. The model processes information about image similarity and tag importance to classify user tags according to their relevance to the input image. The proposed tag ranking approach serves as a framework for combining the input of the other phases of the automatic image annotation approach proposed in this thesis. In the next chapter, the effectiveness of the proposed tag ranking model within the frame of the whole AIA approach of this thesis is evaluated experimentally.

Chapter 8

Experimental Evaluation

In this chapter the effectiveness of the AIA approach proposed in this thesis is demonstrated. For this purpose, several experimental studies are carried out to evaluate the performance of our approach under different settings. Furthermore, the performance of our approach is compared to other baseline methods.

8.1 Setup and Evaluation Procedure

To demonstrate the effectiveness of the proposed AIA approach, we conducted several experimental studies. For this purpose, we used the dataset presented in Chapter 3 as a resource for mining annotations for test images. For each image in the dataset, we extracted SURF features and applied the keypoint pruning approach presented in Chapter 5 to speed up the process of identifying the visual neighbors. Furthermore, for test images which have less than 10 visual neighbors, we applied the iterative image matching approach introduced in Chapter 6 to extend the visual neighbor list.

To investigate different configurations as well as to compare the performance of different AIA methods a ground truth is needed. We created a ground truth from the *Getty Image Flickr Collection*[Flickr, 2014c]. Images in this group are taken and annotated by professionals. That means, the majority of the provided tags are relevant to image content and the rate of noisy tags is also low.

We selected a subset of 152 images from the ground truth. The images in this test set were taken in different areas of the world. The test images were then annotated using different configurations of our approach as well as according to other baseline methods. For each test image the top 10 tag predictions were collected and compared

to the original annotations as specified in the ground truth. Subsequently, the results were quantified according to the evaluation metrics presented in the next section.

8.2 Evaluation Metrics

To quantify the results of AIA approaches, we use standard information retrieval measures. For each test image, the top k tags are determined and the following measures are applied: precision at k ($P@k$), recall at k ($R@k$) and the reciprocal rank at k ($RR@k$). Let S_k be the set of top k tag predictions, G is the set of ground truth tags and r the rank of the first relevant tag among the top k predicted ones, then:

- **Precision at k ($P@k$):** is the percentage of correctly predicted tags out of the k ones.

$$P@k = \frac{|S_k \cap G|}{k} \quad (8.1)$$

- **Recall at k ($R@k$):** is the fraction of ground truth tags that are correctly predicted.

$$R@k = \frac{|S_k \cap G|}{|G|} \quad (8.2)$$

- **Reciprocal Rank at k ($RR@k$):** corresponds to the inverse of the rank of the first relevant tag in the top k predicted ones. The lower the rank of the first relevant tag the higher the reciprocal rank:

$$RR@k = \begin{cases} 1/r & r \neq 0 \\ 0 & \text{Otherwise} \end{cases} \quad (8.3)$$

For each test image, the above three measures are calculated at $k \in \{1, \dots, 10\}$ and averaged over all test images. Subsequently, two AIA approaches can be compared according to the average precision ($AP@k$), the average recall ($AR@k$) and the average (mean) reciprocal rank ($MRR@k$) at different sizes k of the top predicted tags.

8.3 Compared AIA Models

In the following, we will investigate the impact of different factors on the effectiveness of the AIA approach presented in this thesis. Furthermore, the performance of our approach is compared to that of two baseline methods for location-based AIA.

The compared AIA methods assume that the test images are geotagged. They all start by identifying the geographical neighbors of the test image. Thereby, the geographical neighbors correspond to the images found in the same quad-tree region as the test image (refer to Section 3.4).

Baseline: Geographical Neighbor Voting (BL (geo-voting))

In this method, tag proposals for an unlabeled test image are obtained from the set of most frequent tags which are applied on its geographical neighbors. This method corresponds to geographical neighbor voting. Thereby, each geographical neighbor votes for a candidate tag if it is annotated with it. Accordingly, the tag ranking function is given as:

$$rank_{\text{Geo-Voting}}(t|I_{in}) = \sum_{I_i \in \mathcal{I}_{geo}} vote(t, I_i) \quad (8.4)$$

where t is a candidate tag, I_{in} is the test image, \mathcal{I}_{geo} is the set of geographical neighbors of the test image and $vote(t, I_i)$ is the voting function which is defined as:

$$vote(t, I_i) = \begin{cases} 1 & \text{If } I_i \text{ is annotated with } t \\ 0 & \text{Otherwies} \end{cases} \quad (8.5)$$

Baseline: Geographical TF-IDF (BL (TF-IDF))

Similar to the previous baseline, the one presented here also leverages the annotations of the geographical neighbors to determine candidate tags. However, it uses different ranking function based on the TF-IDF approach. Given a candidate tag t , the term frequency $TF(t)$ is computed based on its occurrences in the annotations of the geographical neighbors (i.e., images found in the same quad-tree region). To calculate the inverse document frequency, $IDF(t)$, we retrieve the neighboring quad-tree regions and consider each of them as a document. Correspondingly, the $IDF(t)$ is given by the inverse of the number of quad-tree regions in which t has been used. Finally, the candidate tags are scored according to the following formula:

$$rank_{\text{TF-IDF}_{Geo}}(t|I_{in}) = TF(t) \times IDF(t) \quad (8.6)$$

8.4 Results Discussion

In the following sections the evaluation results are presented and discussed.

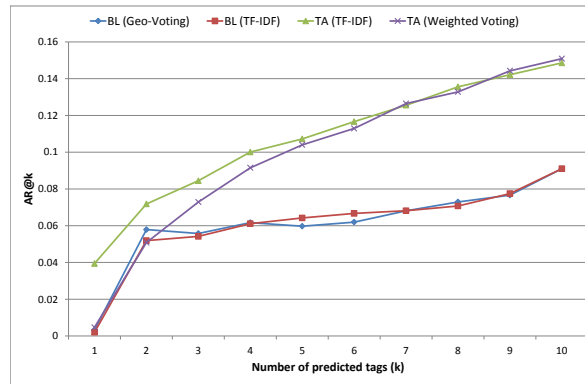
8.4.1 Thesis Approach (TA) vs. Baseline (BL) Methods

First, we want to compare the performance of the tag ranking approach introduced in Chapter 7 to that of the presented baseline methods. For this purpose, we annotated each of the 152 test images using two configurations of our approach. In the first method, TA (TF-IDF), candidate tags are ranked according to the probabilistic model presented in Chapter 7 and using the TF-IDF-based word-to-image relevance (WIR) measure. While in the second method, TA (weighted voting), the weighted voting measure for WIR is used (refer to Section 7.3.2). The test images were also annotated using the presented baselines, BL (TF-IDF) and BL (geo-voting).

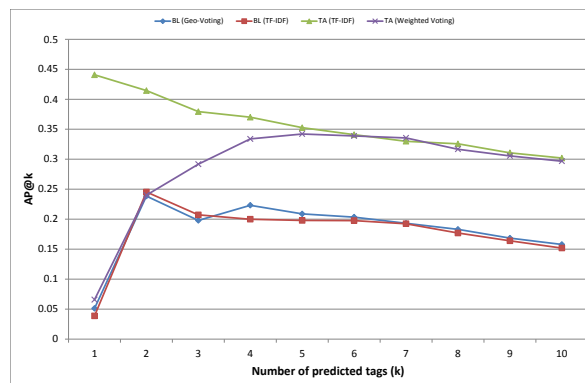
The average recall and the average precision achieved by each method are illustrated in Figure 8.1(a) and 8.1(b), respectively. It can be seen that in general our approach TA (TF-IDF) outperforms the adversary methods in terms of precision and recall. In particular, the average precision of TA (TF-IDF) for the first top tag ($k = 1$) is remarkably high compared to the other methods. With respect to the weighted voting approach, i.e., TA (weighted voting), its performance becomes close to that of TA (TF-IDF) from the fifth predicted tag on. With respect to the baseline approaches they show comparable precision and recall values.

A further investigation of the performance of the compared methods is provided through the mean reciprocal rank (MRR). We calculated the MRR for each method and for the top 10 predicted tags, denoted as $MRR@10$. Again, according to MRR the superiority of TA (TF-IDF) method is evident. For TA (TF-IDF) to provide a correct prediction, it needs to propose a maximum of 2 tags ($1/MRR@10 = 1/0.59 = 1.69$) on average while TA (weighted voting) requires 3 tags. The baseline methods have to propose around 6 tags on average to get a correct tag.

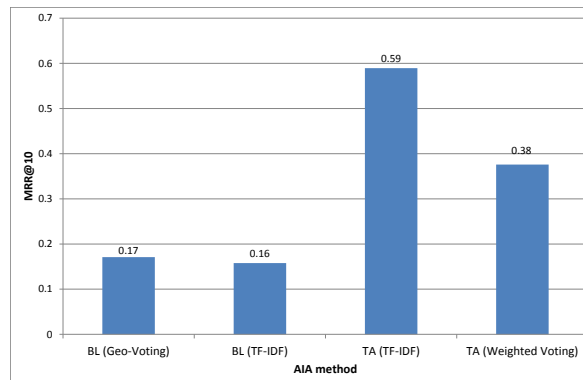
Figure 8.2 shows a sample of the annotated images with corresponding ground truth tags. In contrast to our approach, the baseline method produces the same annotations to images taken in the same location disregarding their visual contents. For instance consider the photos (2) and (3), the collection of suggested tags are identical even though the two images are not. However, the presented baseline methods are useful in the case where no visual neighbors for the input image can be identified. In fact, the presented baseline methods can identify generic tags (e.g., location names or tags like "travel", "tourism", etc.) which could fit any image taken in that specific location. With regard to our approach, the figure shows that it succeeds to identify more specific tags which are related not only to the context but also to the content of the annotated image. In some cases, the tags provided by our approach are even more specific than those provided by the ground truth. For example, for the image number (6) our approach produces for



(a) Average Recall



(b) Average Precision



(c) MRR for the ten top tags (k=10)

FIGURE 8.1: The performance of our AIA approach against the baseline models

the landmark "Hawa Mahal" in India tags which refer to other names of the landmark like "palace of winds" and "pink palace" which are not included in the ground truth. Finally, Figure 8.2 shows the diversity of produced tags as a direct result of leveraging community tags in the mining process.







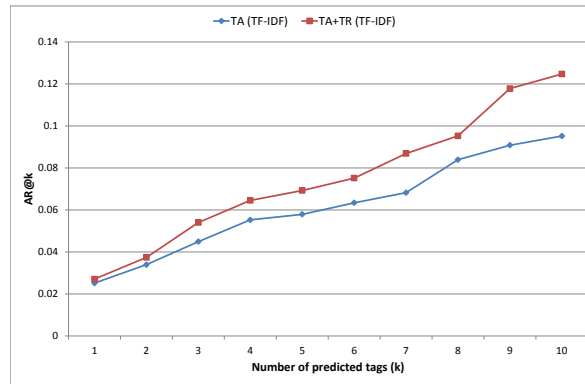
Annotated Image			
	(1)	(2)	(3)
Ground truth	arch, built structure, capital cities, city, column, illuminated, incidental people, Lisbon, motion, night, outdoors, place of interest, square, sky, Praca do Comercio,	Belgium, bell tower, Bruges bell tower, building exterior, city, cold, dawn, dusk, fog, frozen, ice, illuminated, landscape, night, sky, lighting equipment, no people, tree, reflection	bare tree, Belgium, built structure, day, moored, nautical vessel, no people, outdoors, reflection, storm cloud, town, townscape, travel destinations, west Flanders
TA (TF-IDF)	night, arch, Lisbon, Portugal, light, city, triunfal, triumph, people, street	night, canal, Belgium, Belfort, Flanders, belfry, Belgica, Belfort tower, boats, tree	clouds, Belgium, autostitch, dereien, medieval architecture, cel, tormenta, nubes, nuvols, storm
BL (Geo-Voting)	Lisboa, Portugal, liston, street, Lisbona, Europe, Lissabon, city, ilustrarportugal, Portogallo	Bruges, Belgium, Europe, Belgique, canal, Flanders, architecture, travel, Belgio, night	Bruges, Belgium, Belgique, Europe, canal, Flanders, architecture, night, travel, Belgio
Annotated Image			
	(4)	(5)	(6)
Ground truth	capital cities, door, glass, military, people, red, square, state of the Vatican city, Swiss Guard, Vatican city, yellow	circle, diminishing perspective, electric light, empty, indoors, no people, pedestrian walkway, subway	beehive, building exterior, clear sky, day, facade, Hinduism, history, India, Indian culture, window, low angle view, no people, outdoors, palace, sunlight, sunny, tourism
TA (TF-IDF)	Rome, Roma, Italia, Italy, Vatican, St. peters, soldado, Guardia Suiza, Garde Suisse, suisses	ubahn, Marienplatz, underground, subway, orange, München, Munich, metro, tunnel, station	Palace of winds, Hawa Mahal, Rajasthan, architecture, pink city, India, Jaipur, pink palace, palace of the breeze
BL (Geo-Voting)	Rome, Italy, Roma, Vatican, Vaticano, Italia, San Pietro, basilica, Vatican city, church	Munich, Germany, München, Bavaria, Bayern Deutschland, Marienplatz, , munchen, monaco, Viktualienmarkt	India, Jaipur, Rajasthan, palace, Hawa Mahal, travel, city, city palace, architecture, Asia

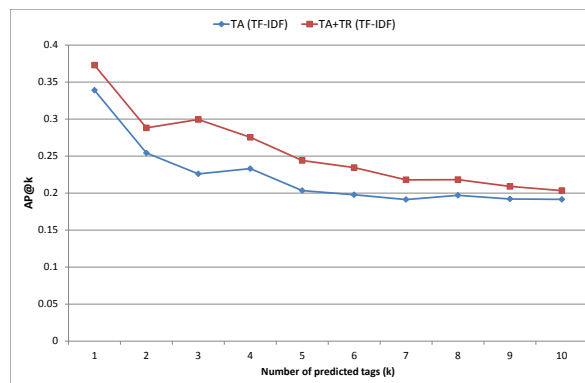
FIGURE 8.2: Sample of the annotated test images with the ground tags and the top tags which were predicted according to our approach TA (TF-IDF) and the baseline method BL (geo-voting)

8.4.2 Effectiveness of Tag Refinement

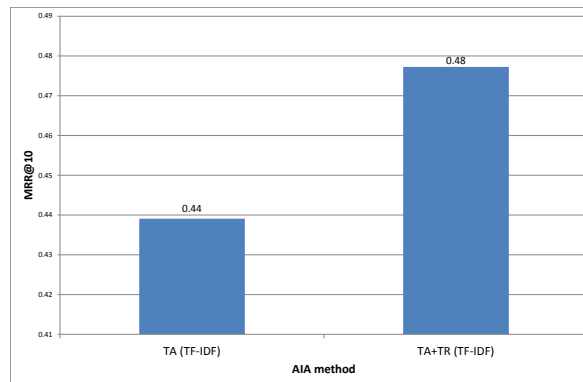
The effect of the tag refinement approach presented in Chapter 4 is evaluated by extracting tag relatedness information from an image folksonomy corresponding to the city of London. Next, the relatedness information is used to disambiguate user-tags as described in Section 4.4. We selected subset of 59 images taken in London from the ground truth (Figure 8.4). Each test image was annotated using our approach TA (TF-IDF) and an extension thereof which applies tag refinement (TR). We refer to the latter methods with TA+TR (TF-IDF). TA+TR (TF-IDF) applies tag refinement on the the tags which are produced by TA (TF-IDF). Thereby, some of the top k tags can be removed (e.g. redundant tags) or combined. Consequently, the tag proposal list is extended by new tag proposals to reach a size of k .



(a) Average Recall



(b) Average Precision



(c) MRR for the ten top tags (k=10)

FIGURE 8.3: The performance of the proposed AIA with tag refinement

Figure 8.3 shows that applying tag refinement have a positive effect on improving the performance of the proposed automatic image annotation approach. Refining tag proposals increase the average recall, precision and the mean reciprocal rank of the annotation approach.

In fact, this effect is less evident in this experiment than it might be in reality. This can be justified according to the nature of the ground truth. By investigating Getty tags,




Annotated Image			
	(1)	(2)	(3)
Ground truth	arch, bridge, british culture, building exterior, capital cities, clock tower, cloud, connection, day, incidental people, outdoors, place of interest, railings, reflection, sea, sky, tourism, transportation, travel destinations, UK	bascule bridge, bridge, British culture, built structure, capital cities, city, connection, dusk, international landmark, no people, outdoors, river Thames, sky, sunset, suspension bridge, transportation, travel destinations, UK	British culture, building exterior, capital cities, city, cloud, dusk, history, Hyde park - London, memories, no people, outdoors, place of interest, rain, reflection, sky, spire, square, sunset, travel destinations, UK, wet
TA (TF-IDF)	Big Ben, London, Thames, Westminster, England, parliament, houses of parliament, river, bridge, sunset	tower, bridge, London, England, Thames, river, Londres, United Kingdom, blue, Europe	Albert, London, England, memorial, Hyde Park, Kensington gardens, United Kingdom, Kensington, Europe, clouds
TA +TR (TF-IDF)	Big Ben, England London , City of London , Westminster North (UK parliament constituency) , houses of parliament, River Thames , sunset, <i>clouds</i> , Westminster Bridge , <i>buildings</i>	Tower Bridge , England London , River Thames , Londres, United Kingdom, blue, Europe, <i>landmark</i> , <i>lights</i> , <i>road</i>	Albert Memorial , England London , Hyde Park, <i>Bedford park</i> , Kensington gardens, United Kingdom, <i>Great Britain</i> , <i>Prince Albert</i> , Europe, clouds

FIGURE 8.4: Sample of test images which were annotated using the approach: TA (TF-IDF) without tag refinement, compared to the same approach under tag refinement: TA+TR(TF-IDF). Tag proposals in bold correspond to refined tags and those in italic results from extending the annotation list to reach $k = 10$

we have noticed that even the provided tags are created by experts, they sometimes suffer from redundancy or some descriptive tags are absent. Furthermore, some of landmark names are not indicated directly. For instance, consider the second photo of Figure 8.4. In that photo the landmark "Tower Bridge" appears, however, the ground truth does not contain the name of the landmark, instead only the tag "Bridge" is provided. Accordingly, an AIA approach which proposes the tag "Bridge" will have better performance than another approach which proposes the tag "Tower Bridge". The same holds for the third photo where the name of the memorial called "Albert Memorial" is not indicated.

8.5 Summary

In this chapter, the performance of the automatic annotation approach presented in this thesis is evaluated. We investigated different configurations of the tag ranking approach presented in the previous chapter and compared the quality of the produced tag proposals to that of two baseline methods for location-based image annotation. The results confirm that our approach outperforms baseline methods. Furthermore, we investigated the effect of the tag refinement approach proposed in Chapter 4 and demonstrated its positive effect on the mined tags.

Chapter 9

Conclusion and Future Work

In this chapter, we will draw together the main conclusions of the work presented in this thesis. We also present potential areas for future work.

9.1 Summary of Research

This thesis addressed the problem of automatic image annotation by following the search-based paradigm and leveraging community photos. It investigated the importance of incorporating location information in the annotation process and presented solutions for boosting the performance and the quality of the produced annotations. More specifically, the thesis proposed a complete framework for automatic image annotation which is built on efficiency and quality considerations. On the one hand, a special focus was given to the computational efficiency of the annotation process. In particular, the thesis considered the problem of data collection and spatial indexing. Furthermore, it succeeded in improving the accuracy and reducing the time required by the phase of visual neighbor's identification. On the other hand, the thesis put a special emphasis on the quality of the mined tag proposals. In this respect, the thesis introduced a novel solution for identifying related tags in folksonomies and successfully incorporated the extracted information in a tag refinement phase. Finally, the thesis presented a scalable model for combining multi-modal information for boosting the ranks of the predicted tags. Below, the specific contributions and findings of the thesis are summarized.

9.1.1 Two-Phase Automatic Image Annotation Approach

In Chapter 1, we described the workflow of our automatic image annotation approach. We divided the annotation process into two phases. A data preparation phase, which

is concerned with issues related to collecting and pre-processing data from community photos and a tag mining phase, which describes the actual annotation process. This separation is useful since the data preparation phase is computationally intensive. It requires querying photo sharing websites, which is inherently a slow process. Furthermore, the process of cleaning and extracting tag relatedness information from the collected data demands high computations. Therefore, we decided to collect and pre-process the data in an offline phase to reduce the computations needed by the actual tag mining process.

9.1.2 Geo-based Data Crawling and Indexing

The thesis presented in Chapter 3 a data crawling strategy for collecting photos and the associated metadata from Flickr using location information. The presented method succeeded in gathering more than 14 million images on a world-wide scale. Thereby, a special attention has been given to representativeness aspects. In particular, our dataset covers the main photographed places in the world and the density of the photos for a given place reflects its real popularity among photographers. Furthermore, we presented a method for indexing the collected data spatially to enable efficient retrieval. For this purpose, we adapted the quad-tree algorithm to deal with huge amounts of data.

9.1.3 Resolving Tag Ambiguity

Tags of already annotated community photos represent the main resource for mining annotations for unlabeled images. However, as we mentioned before, user-supplied tags are noisy in terms of syntax and semantic. To address this problem, we presented in Chapter 3 a simple technique to deal with lexical variations of user tags by using the search engine Yahoo. Although simple, the presented solution is effective. However, it is important to mention that there is a restriction on the number of queries which can be submitted to Yahoo per hour. Hence, only a limited number of tags can be cleaned per day. To handle more sophisticated problems, we presented in Chapter 4 a novel approach for mining tag relatedness information as main step to resolve their ambiguity. We experimentally demonstrated the advantage of building tag representation based on their co-occurrence with the features of the highest Laplacian scores. In fact, our method showed to outperform the standard method of representing tags based on their co-occurrence with most frequent tags in the folksonomy. Furthermore, we investigated the effect of the selected distance measure on identifying related tags. In this respect, we proposed an extension of Jensen-Shannon Divergence which considers sampling errors. The new measure AJSD is promising and it outperforms the standards JSD as well as the cosine measure.

In summary, two main conclusions with regard to tag relatedness approach can be drawn. First, representing tags as probability distributions is more effective than using mere co-occurrence vectors. Second, measures of statistical divergence (e.g. AJSD, JSD) outperform the cosine similarity. However, it is important to note the proposed distance measure AJSD implies more complex computations than JSD and the cosine measure. Furthermore, the Laplacian score algorithm needs more investigation to determine the optimal parameter values. Nevertheless, promising results have been achieved by using the default settings of the Laplacian score method combined with the AJSD measure.

The thesis also presented a method for refining tag proposals based on tag relatedness information. We experimentally demonstrated the positive effect of the proposed method on improving tag proposals. Although, the proposed method is simple, it showed to be successful to predict more relevant tag proposals than in the case where no tag refinement is applied.

9.1.4 Improving SURF-based Image Matching

Image matching is the most computationally expensive component of the automatic annotation process. In Chapter 5, the thesis presented an approach for boosting the performance of SURF-based image matching using a method for keypoint pruning. The proposed approach reduces the number of compared SURF descriptors significantly. Consequently, our image matching approach is notably faster than the standard method. We believe that the proposed method can be further improved by considering and combining additional features for keypoint characterization. Furthermore, the thesis presented a method to improve the recall of SURF-based image matching by an iterative application of the matching process. To reduce the computation required by each new matching phase, we presented a method for image clustering based on the keypoints which they share with a common similar image. Our approach succeeded in increasing SURF-based matching recall by discovering additional images which are similar to a given query image. Furthermore, the proposed image clustering method demonstrated significant reduction in the required runtime compared to a baseline approach. Finally, it is important to mention that there is a trade-off between the matching speed and the matching recall. Therefore, to obtain the most benefit, it is recommended to use the proposed iterative matching approach in settings where the number of visual neighbors identified by a direct matching is reasonably small.

9.1.5 Tag Ranking

The thesis proposed a probabilistic framework for tag ranking based on Bayes' rule in Chapter 7. The presented model combines visual, geographical, and tag usage information in order to rank candidate annotations. In Chapter 8, we demonstrated the effectiveness of this approach experimentally. The results showed that the proposed ranking model outperforms baseline methods for location-based AIA and can be easily extended to consider further contextual clues.

9.2 Future Work

With respect to extracting tag relatedness information, most approaches are focused on exploiting tag co-occurrence between the subset of most frequent tags in the considered folksonomy. However, we have shown in this thesis that this might be not the best solution (Laplacian score feature selection provided better results). Therefore, the topic of unsupervised feature selection for creating tag representation deserves further investigation. In particular, the performance of Laplacian score algorithm can be further improved by investigating non-standard similarity measures for building the nearest neighbor graph. In fact, we expect that incorporating the proposed AJSD measure might increase the effectiveness of the LS algorithm. With respect to the problem of refining tag proposals, the solution presented in this thesis showed to be effective. A possible future work could be to extend the presented idea to triples or clusters of related tags.

Several works considered the problem of tag disambiguation (e.g. [Garcia et al., 2009]); however, fewer efforts were made to address the problem of enriching tag proposals. In fact, works on tag enrichment use a simple method to establish links between tags and semantic entities. The idea is to create a context (group of tags) for each tag (the one we want to enrich) based on other tags which co-occur with it. Subsequently, the context is used to query a semantic resource (e.g. DBPedia [Bizer et al., 2009]), to retrieve a suitable semantic entity. In that process, a more attention should be given to the manner, in which the context is created. For instance, creating the context based on tag co-occurrences only might lead to semantically inconsistent contexts (e.g. a context which contains two contradictory tags like "outdoor" and "sky").

CBIR algorithms are witnessing more and more success; however, existing algorithms are still impractical for applications, such as AIA where a real-time response is expected. To further improve this process, the approach of keypoint pruning presented in this thesis can be combined with indexing techniques like the BoW (the Bag of Visual Words) approach.

There is still no standard benchmark for evaluating and comparing the performance of AIA approaches which exploits location information. Most works follow an ad-hoc method to demonstrate their efficiency. In this thesis, we attempted to use annotations from Getty Flickr Collection which are provided by experts as a ground truth. However, by investigating the ground truth we recognized that provided annotations are incomplete. Therefore, to push research on automatic image annotation there is a need for a complete benchmark. Furthermore, more focus should be given to the measures which are used to evaluate and compare different AIA approaches. In most cases, the effectiveness of an AIA approach is evaluated by comparing its mined annotations to that of the ground truth using *exact* string matching. However, this might not be fair enough. To address this problem, new evaluation measures that incorporate information on the semantic similarity between the mined and the reference annotations are needed. In this regard, semantic resources or even tag relatedness information, which can be extracted according to the approach presented in this thesis, can be integrated in the evaluation process.

The contributions of this thesis can also be advantageous for other research activities. Most recently, there is an increasing interest in automatically identifying the location of non-geotagged multimedia entities (i.e. images and videos) in a process called *reverse geotagging* [Kalogerakis et al., 2009, Trevisiol et al., 2013]. We can support this research activity since the dataset presented in this thesis provides rich information that can be used to train models for predicting the location of non geotagged images. Furthermore, since several works exploits user-provided tags as an important feature in the location mining process, our tag relatedness approach can be used to discover more discriminative textual features by identifying noisy tags and resolving their ambiguity.

Other research efforts considered the problem of automatically generating travel blogs by utilizing contextual information such as location, time, social networks, etc. (e.g. [Cemerlang et al., 2006, Li and Hua, 2010]). The presented automatic image annotations approach can assist such applications by automatically providing rich textual descriptions for images taken in a specific location. This information can be fed into a summarization or a story generation engine to produce automatic report of personal experiences. Moreover, the work presented in this thesis provides a solid basis for touristic recommendation systems. Location information can be analyzed to identify the most popular touristic places and touristic routes. Furthermore, the provided textual information combined with efficient methods for identifying representative images can be used to assist a process like tourist guide generation (e.g. [Kori et al., 2006, Lu et al., 2010]).

Bibliography

Paul Graham. Web 2.0, 2005. URL <http://www.paulgraham.com/web20.html>. Accessed: 23/4/2014.

Tim O'Reilly. What is web 2.0: Design patterns and business models for the next generation of software, 2005. URL <http://oreilly.com/web2/archive/what-is-web-20.html>. Accessed: 23/4/2014.

Facebook, 2014. URL <http://www.facebook.com/>. Accessed: 23/4/2014.

Arthur Brisbane. Speakers give sound advice. *Syracuse Post Standard*, page 19, March 1911.

Flickr, 2014a. URL <http://www.flickr.com/>. Accessed: 23/4/2014.

Flickr, 2014b. URL <http://www.flickr.com/services/api/misc.urls.html>. Accessed: 23/4/2014.

Adi Robertson. Facebook users have uploaded a quarter trillion photos since the site's launch, 2013. URL <http://www.theverge.com/2013/9/17/4741332/facebook-users-have-uploaded-a-quarter-trillion-photos-since-launch>. Accessed: 23/4/2014.

Thomas Vanderwal. Off the top: Folksonomy entries, 2010. URL <http://vanderwal.net/random/category.php?cat=153>. Accessed: 23/4/2014.

Technical Standardization Committee on AV & IT Storage Systems and Equipment. Exchangeable image file format for digital still cameras: Exif version 2.2. *JEITA CP-3451*, April 2002.

G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, November 1987. ISSN 0001-0782. doi: 10.1145/32206.32212. URL <http://doi.acm.org/10.1145/32206.32212>.

Panoramio. Panoramio, 2014. URL <http://www.panoramio.com/>. Accessed: 23/4/2014.

- Stanley Milgram. The Small World Problem. *Psychology Today*, 2:60–67, 1967.
- R.A. Finkel and J.L. Bentley. Quad trees a data structure for retrieval on composite keys. *Acta Informatica*, 4(1):1–9, 1974. ISSN 0001-5903. doi: 10.1007/BF00288933. URL <http://dx.doi.org/10.1007/BF00288933>.
- Xiaofei He, Deng Cai, and Partha Niyogi. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 18:507, 2005.
- Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.
- Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110:346–359, June 2008. ISSN 1077-3142. doi: 10.1016/j.cviu.2007.09.014. URL <http://dl.acm.org/citation.cfm?id=1370312.1370556>.
- Flickr, 2014c. URL <http://www.flickr.com/gettyimages/>. Accessed: 23/4/2014.
- Hatem Mousselly Sergieh, Daniel Watzinger, Bastian Huber, Mario Döller, Elöd Egyed-Zsigmond, and Harald Kosch. Folkioneer: Efficient browsing of community geotagged images on a worldwide scale. In *MultiMedia Modeling - 20th Anniversary International Conference*, pages 361–364. Springer, 2014a.
- Hatem Mousselly Sergieh, Daniel Watzinger, Bastian Huber, Mario Döller, Elöd Egyed-Zsigmond, and Harald Kosch. World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging. In *Proceedings of the 5th ACM Multimedia Systems Conference, MMSys '14*, pages 47–52, New York, NY, USA, 2014b. ACM. ISBN 978-1-4503-2705-3. doi: 10.1145/2557642.2563673. URL <http://doi.acm.org/10.1145/2557642.2563673>.
- Hatem Mousselly Sergieh, Elöd Egyed-Zsigmond, Mario Döller, Gabriele Gianini, Harald Kosch, and Jean-Marie Pinon. Tag similarity in folksonomies. In *INFORSID*, pages 319–334, 2013.
- Hatem Mousselly Sergieh, Mario Döller, Elöd Egyed-Zsigmond, Gabriele Gianini, Harald Kosch, and Jean-Marie Pinon. Tag relatedness using laplacian score feature selection and adapted jensen-shannon divergence. In *MultiMedia Modeling - 20th Anniversary International Conference*, pages 159–171. Springer, 2014c.
- Hatem Mousselly Sergieh, Elöd Egyed-Zsigmond, Mario Döller, David Coquil, Jean-Marie Pinon, and Harald Kosch. Improving surf image matching using supervised learning. In *Signal Image Technology and Internet Based Systems (SITIS), 2012*

- Eighth International Conference on*, pages 230–237, Nov 2012a. doi: 10.1109/SITIS.2012.42.
- Hatem Mousselly Sergieh, Gabriele Gianini, Mario Döllner, Harald Kosch, Elöd Egyed-Zsigmond, and Jean-Marie Pinon. Geo-based automatic image annotation. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12*, pages 46:1–46:8, New York, NY, USA, 2012b. ACM. ISBN 978-1-4503-1329-2. doi: 10.1145/2324796.2324850. URL <http://doi.acm.org/10.1145/2324796.2324850>.
- Xirong Li, C. G M Snoek, and Marcel Worring. Annotating images by harnessing worldwide user-tagged photos. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3717–3720, April 2009a. doi: 10.1109/ICASSP.2009.4960434.
- Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.*, 44(4):25:1–25:24, September 2012. ISSN 0360-0300. doi: 10.1145/2333112.2333120. URL <http://doi.acm.org/10.1145/2333112.2333120>.
- Thomas Mensink, Jakob Verbeek, and Gabriela Csurka. Trans media relevance feedback for image autoannotation. In *Proceedings of the British Machine Vision Conference*, pages 20.1–20.12. BMVA Press, 2010a. ISBN 1-901725-40-5. doi:10.5244/C.24.20.
- Lamberto Ballan, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Social media annotation. In *Proc. of IEEE International Workshop on Content-Based Multimedia Indexing (CBMI)*, Veszprem, Hungary, June 2013. IEEE Computer Society. (Invited).
- Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, December 2000. ISSN 0162-8828. doi: 10.1109/34.895972. URL <http://dx.doi.org/10.1109/34.895972>.
- Ingemar J. Cox, M.L. Miller, T.P. Minka, T.V. Papatomas, and P.N. Yianilos. The bayesian image retrieval system, pichunter: theory, implementation, and psychophysical experiments. *Image Processing, IEEE Transactions on*, 9(1):20–37, Jan 2000. ISSN 1057-7149. doi: 10.1109/83.817596.
- Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 10(1):39 – 62, 1999. ISSN 1047-3203. doi: <http://dx.doi.org/10.1006/jvci.1999.0413>. URL <http://www.sciencedirect.com/science/article/pii/S1047320399904133>.

- Dilip K Prasad. Survey of the problem of object detection in real images. *International Journal of Image Processing (IJIP)*, 6(6):441, 2012.
- David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, November 2004. ISSN 0920-5691.
- Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:506–513, 2004. ISSN 1063-6919. doi: <http://doi.ieeecomputersociety.org/10.1109/CVPR.2004.183>.
- H. Ferhatosmanoglu, E. Tuncel, D. Agrawal, and A. El Abbadi. Approximate nearest neighbor searching in multimedia databases. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 503–511, 2001. doi: 10.1109/ICDE.2001.914864.
- Jing Liu, Mingjing Li, Qingshan Liu, Hanqing Lu, and Songde Ma. Image annotation via graph learning. *Pattern Recogn.*, 42(2):218–228, February 2009. ISSN 0031-3203. doi: 10.1016/j.patcog.2008.04.012. URL <http://dx.doi.org/10.1016/j.patcog.2008.04.012>.
- P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV, ECCV '02*, pages 97–112, London, UK, UK, 2002. Springer-Verlag. ISBN 3-540-43748-7. URL <http://dl.acm.org/citation.cfm?id=645318.649254>.
- J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, SIGIR '03*, pages 119–126, New York, NY, USA, 2003. ACM. ISBN 1-58113-646-3. doi: 10.1145/860435.860459. URL <http://doi.acm.org/10.1145/860435.860459>.
- V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *IN NIPS*. MIT Press, 2003.
- S.L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II-1002–II-1009 Vol.2, June 2004. doi: 10.1109/CVPR.2004.1315274.
- O. Chapelle, P. Haffner, and V.N. Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, Sep 1999. ISSN 1045-9227. doi: 10.1109/72.788646.

- Claudio Cusano, Gianluigi Ciocca, and Raimondo Schettini. Image annotation using svm. *Internet Imaging IV, SPIE*, 5304:330–338, 2003. doi: 10.1117/12.526746. URL <http://dx.doi.org/10.1117/12.526746>.
- Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60, May 2008. ISSN 0360-0300. doi: 10.1145/1348246.1348248. URL <http://doi.acm.org/10.1145/1348246.1348248>.
- Dengsheng Zhang, Md. Monirul Islam, and Guojun Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346 – 362, 2012. ISSN 0031-3203. doi: <http://dx.doi.org/10.1016/j.patcog.2011.05.013>. URL <http://www.sciencedirect.com/science/article/pii/S0031320311002391>.
- Feichao Wang. A survey on automatic image annotation and trends of the new age. *Procedia Engineering*, 23(0):434 – 438, 2011. ISSN 1877-7058. doi: <http://dx.doi.org/10.1016/j.proeng.2011.11.2526>. URL <http://www.sciencedirect.com/science/article/pii/S1877705811053690>. {PEEA} 2011.
- Jia Deng, Alexander C. Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Proceedings of the 11th European Conference on Computer Vision: Part V, ECCV'10*, pages 71–84, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15554-5, 978-3-642-15554-3. URL <http://dl.acm.org/citation.cfm?id=1888150.1888157>.
- Peter G.B. Enser, Christine J. Sandom, and Paul Lewis. Automatic annotation of images from the practitioner perspective. In *4th International Conference on Image and Video Retrieval 2005*, volume LNCS 3, pages 497–506. Springer Berlin/Heidelberg, 2005.
- Xin-Jing Wang, Lei Zhang, Feng Jing, and Wei-Ying Ma. Annosearch: Image auto-annotation by search. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06*, pages 1483–1490, Washington, DC, USA, 2006a. IEEE Computer Society. ISBN 0-7695-2597-0. doi: 10.1109/CVPR.2006.58. URL <http://dx.doi.org/10.1109/CVPR.2006.58>.
- Jing Huang, S.R. Kumar, M. Mitra, Wei-Jing Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768, Jun 1997. doi: 10.1109/CVPR.1997.609412.
- Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. Learning to cluster web search results. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*

- '04, pages 210–217, New York, NY, USA, 2004. ACM. ISBN 1-58113-881-4. doi: 10.1145/1008992.1009030. URL <http://doi.acm.org/10.1145/1008992.1009030>.
- Xirong Li, Le Chen, Lei Zhang, Fuzong Lin, and Wei-Ying Ma. Image annotation by large-scale content-based image retrieval. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 607–610, New York, NY, USA, 2006. ACM. ISBN 1-59593-447-2. doi: 10.1145/1180639.1180764. URL <http://doi.acm.org/10.1145/1180639.1180764>.
- Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. Scalable search-based image annotation of personal images. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, MIR '06, pages 269–278, New York, NY, USA, 2006b. ACM. ISBN 1-59593-495-2. doi: 10.1145/1178677.1178714. URL <http://doi.acm.org/10.1145/1178677.1178714>.
- R. Baeza Yates, B. Ribeiro Neto, et al. *Modern information retrieval*, volume 463. ACM press New York., 1999.
- Changhu Wang, Feng Jing, Lei Zhang, and Hong-Jiang Zhang. Image annotation refinement using random walk with restarts. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 647–650, New York, NY, USA, 2006c. ACM. ISBN 1-59593-447-2. doi: 10.1145/1180639.1180774. URL <http://doi.acm.org/10.1145/1180639.1180774>.
- Google, 2014. URL <http://www.google.com/imghp/>. Accessed: 23/4/2014.
- Xiaoguang Rui, Nenghai Yu, Taifeng Wang, and Mingjing Li. A search-based web image annotation method. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 655–658, July 2007a. doi: 10.1109/ICME.2007.4284735.
- Glenn Shafer. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.
- Xiaoguang Rui, Mingjing Li, Zhiwei Li, Wei-Ying Ma, and Nenghai Yu. Bipartite graph reinforcement model for web image annotation. In *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA '07, pages 585–594, New York, NY, USA, 2007b. ACM. ISBN 978-1-59593-702-5. doi: 10.1145/1291233.1291378. URL <http://doi.acm.org/10.1145/1291233.1291378>.
- George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference on Research in Computational Linguistics*, 1997.

- Jing Liu, Bin Wang, Mingjing Li, Zhiwei Li, Weiyang Ma, Hanqing Lu, and Songde Ma. Dual cross-media relevance model for image annotation. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 605–614, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-702-5. doi: <http://doi.acm.org/10.1145/1291233.1291380>. URL <http://doi.acm.org/10.1145/1291233.1291380>.
- Rudi L. Cilibrasi and Paul M. B. Vitanyi. The google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19:370–383, March 2007. ISSN 1041-4347. doi: 10.1109/TKDE.2007.48. URL <http://dl.acm.org/citation.cfm?id=1263132.1263333>.
- Dingyin Xia, Fei Wu, and Yueting Zhuang. Search-based automatic web image annotation using latent visual and semantic analysis. In *PCM*, pages 842–845, 2008.
- Scott C. Deerwester, Susan T Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science* 41(6), 41(6):391–407, 1990.
- Tao Mei, Yong Wang, Xian-Sheng Hua, Shaogang Gong, and Shipeng Li. Coherent image annotation by learning semantic distance. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587386.
- Stefanie Lindstaedt, Viktoria Pammer, Roland Mörzinger, Roman Kern, Helmut Mülner, and Claudia Wagner. Recommending tags for pictures based on text, visual content and user context. In *Proceedings of the 2008 Third International Conference on Internet and Web Applications and Services*, ICIW '08, pages 506–511, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3163-2. doi: 10.1109/ICIW.2008.26. URL <http://dx.doi.org/10.1109/ICIW.2008.26>.
- Hong-Ming Chen, Ming-Hsiu Chang, Ping-Chieh Chang, Ming-Chun Tien, Winston H. Hsu, and Ja-Ling Wu. Sheepdog: Group and tag recommendation for flickr photos by automatic search-based learning. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 737–740, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-303-7. doi: 10.1145/1459359.1459473. URL <http://doi.acm.org/10.1145/1459359.1459473>.
- Börkur Sigurbjörnsson and Roelof van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 327–336, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367542. URL <http://doi.acm.org/10.1145/1367497.1367542>.

- Ashton Anderson, K Ranghunathan, and Adam Vogel. Tagez: Flickr tag recommendation. *Association for the Advancement of Artificial Intelligence*, 2008.
- Jia Li and J.Z. Wang. Real-time computerized annotation of pictures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(6):985–1002, June 2008. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.70847.
- Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. Learning to tag. In *18th International World Wide Web Conference*, pages 361–361, April 2009. URL <http://www2009.eprints.org/37/>.
- Lei Wu, Mingjing Li, Zhiwei Li, Wei-Ying Ma, and Nenghai Yu. Visual language modeling for image classification. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, MIR '07*, pages 115–124, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-778-0. doi: 10.1145/1290082.1290101. URL <http://doi.acm.org/10.1145/1290082.1290101>.
- Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4:933–969, December 2003. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=945365.964285>.
- X. Li, C. G. M. Snoek, and M. Worring. Learning social tag relevance by neighbor voting. *IEEE Transactions on Multimedia*, 11(7):1310–1322, 2009b. URL <http://www.science.uva.nl/research/publications/2009/LiITM2009>.
- Xirong Li, Cees G.M. Snoek, and Marcel Worring. Learning tag relevance by neighbor voting for social image retrieval. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08*, pages 180–187, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-312-9. doi: 10.1145/1460096.1460126. URL <http://doi.acm.org/10.1145/1460096.1460126>.
- Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. Baselines for image annotation. *Int. J. Comput. Vision*, 90(1):88–105, October 2010. ISSN 0920-5691. doi: 10.1007/s11263-010-0338-6. URL <http://dx.doi.org/10.1007/s11263-010-0338-6>.
- M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 309–316, Sept 2009. doi: 10.1109/ICCV.2009.5459266.
- Jakob Verbeek, Matthieu Guillaumin, Thomas Mensink, and Cordelia Schmid. Image annotation with tagprop on the mirflickr set. In *Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10*, pages 537–546, New York,

- NY, USA, 2010. ACM. ISBN 978-1-60558-815-5. doi: 10.1145/1743384.1743476. URL <http://doi.acm.org/10.1145/1743384.1743476>.
- Thomas Mensink, Jakob J. Verbeek, and Gabriela Csurka. Trans media relevance feedback for image autoannotation. In *British Machine Vision Conference, BMVC 2010*, pages 1–12, 2010b.
- Junsong Yuan, Jiebo Luo, Henry Kautz, and Ying Wu. Mining gps traces and visual words for event classification. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08*, pages 2–9, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-312-9. doi: 10.1145/1460096.1460099. URL <http://doi.acm.org/10.1145/1460096.1460099>.
- Mor Naaman, Yee Jiun Song, Andreas Paepcke, and Hector Garcia-Molina. Automatic organization for digital photographs with geographic coordinates. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '04*, pages 53–62, New York, NY, USA, 2004. ACM. ISBN 1-58113-832-6. doi: 10.1145/996350.996366. URL <http://doi.acm.org/10.1145/996350.996366>.
- Till Quack, Bastian Leibe, and Luc Van Gool. World-scale mining of objects and events from community photo collections. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval, CIVR '08*, pages 47–56, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-070-8. doi: 10.1145/1386352.1386363. URL <http://doi.acm.org/10.1145/1386352.1386363>.
- Yan-Tao Zheng, Ming Zhao, Yang Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: Building a web-scale landmark recognition engine. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1085–1092, June 2009. doi: 10.1109/CVPR.2009.5206749.
- Lyndon S. Kennedy and Mor Naaman. Generating diverse and representative image search results for landmarks. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 297–306, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367539. URL <http://doi.acm.org/10.1145/1367497.1367539>.
- Lyndon Kennedy, Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: Context and content in community-contributed media collections. In *Proceedings of the 15th International Conference on Multimedia, MULTIMEDIA '07*, pages 631–640, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-702-5. doi: 10.1145/1291233.1291384. URL <http://doi.acm.org/10.1145/1291233.1291384>.

- Shane Ahern, Mor Naaman, Rahul Nair, and Jeannie Hui-I Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '07*, pages 1–10, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-644-8. doi: 10.1145/1255175.1255177. URL <http://doi.acm.org/10.1145/1255175.1255177>.
- Mor Naaman, Andreas Paepcke, and Hector Garcia-Molina. From where to what: Metadata sharing for digital photographs with geographic coordinates. In *11th International Conference on Cooperative Information Systems (COOPIS 2003)*, November 2003. URL <http://ilpubs.stanford.edu:8090/756/>.
- Mor Naaman and Rahul Nair. Zonetag’s collaborative tag suggestions: What is this person doing in my phone? *IEEE MultiMedia*, 15(3):34–40, July 2008. ISSN 1070-986X. doi: 10.1109/MMUL.2008.69. URL <http://dx.doi.org/10.1109/MMUL.2008.69>.
- An-Jung Cheng, Fang-Erh Lin, Yin-Hsi Kuo, and Winston H. Hsu. Gps, compass, or camera?: Investigating effective mobile sensors for automatic search-based image annotation. In *Proceedings of the International Conference on Multimedia, MM '10*, pages 815–818, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1874086. URL <http://doi.acm.org/10.1145/1873951.1874086>.
- Adrian Popescu and Pierre-Alain Moëllic. Monuanno: Automatic annotation of geo-referenced landmarks images. In *Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR '09*, pages 11:1–11:8, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-480-5. doi: 10.1145/1646396.1646412. URL <http://doi.acm.org/10.1145/1646396.1646412>.
- Rabeeh Abbasi, Marcin Grzegorzec, and Steffen Staab. Large scale tag recommendation using different image representations. In *Semantic Multimedia*, volume 5887 of *Lecture Notes in Computer Science*, pages 65–76. Springer Berlin / Heidelberg, 2009. ISBN 978-3-642-10542-5. doi: 10.1007/978-3-642-10543-2_8.
- Emily Moxley, Jim Kleban, and B. S. Manjunath. Spirittagger: A geo-aware tag suggestion tool mined from flickr. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, MIR '08*, pages 24–30, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-312-9. doi: 10.1145/1460096.1460102. URL <http://doi.acm.org/10.1145/1460096.1460102>.
- Ana Silva and Bruno Martins. Tag recommendation for georeferenced photos. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks, LBSN '11*, pages 57–64, New York, NY, USA, 2011. ACM. ISBN

- 978-1-4503-1033-8. doi: 10.1145/2063212.2063229. URL <http://doi.acm.org/10.1145/2063212.2063229>.
- Mădălina Mitran, Rada Mihalcea, Guillaume Cabanac, and Mohand Boughanem. Landmark image annotation using textual and geolocation metadata. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, OAIR '13*, pages 65–68, 2013. ISBN 978-2-905450-09-8. URL <http://dl.acm.org/citation.cfm?id=2491748.2491765>.
- Joseph A. Shaw, Edward A. Fox, Joseph A. Shaw, and Edward A. Fox. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)*, pages 243–252, 1994.
- Philip J McParlane and Joemon M Jose. Exploiting time in automatic image tagging. In *Advances in Information Retrieval*, pages 520–531. Springer, 2013.
- Ludovic Denoyer and Patrick Gallinari. A ranking based model for automatic image annotation in a social network. In *Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM-10)*. AAAI Press, May 2010.
- Neela Sawant, Ritendra Datta, Jia Li, and James Z. Wang. Quest for relevant tags using local interaction networks and visual content. In *Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10*, pages 231–240, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-815-5. doi: 10.1145/1743384.1743424. URL <http://doi.acm.org/10.1145/1743384.1743424>.
- Nikhil Garg and Ingmar Weber. Personalized tag suggestion for flickr. In *Proceedings of the 17th International Conference on World Wide Web, WWW '08*, pages 1063–1064, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367657. URL <http://doi.acm.org/10.1145/1367497.1367657>.
- Adam Rae, Börkur Sigurbjörnsson, and Roelof van Zwol. Improving tag recommendation using social networks. In *Adaptivity, Personalization and Fusion of Heterogeneous Information, RIAO '10*, pages 92–99, Paris, France, France, 2010. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE. URL <http://dl.acm.org/citation.cfm?id=1937055.1937077>.
- N. Elahi, R. Karlsen, and W. Younas. Semantic image annotation with social context. In *Internet Technology and Secured Transactions (ICITST), 2010 International Conference for*, pages 1–7, Nov 2010.
- Carsten Keßler, Patrick Maué, Jan T. Heuer, and Thomas Bartoschek. Bottom-up gazetteers: Learning from the implicit semantics of geotags. In *Proceedings of the*

- 3rd International Conference on GeoSpatial Semantics, GeoS '09*, pages 83–102, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-10435-0. doi: 10.1007/978-3-642-10436-7_6. URL http://dx.doi.org/10.1007/978-3-642-10436-7_6.
- J. Hays and A.A. Efros. Im2gps: estimating geographic information from a single image. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587784.
- Yannis Kalantidis, Giorgos Tolias, Yannis Avrithis, Marios Phinikettos, Evaggelos Spyrou, Phivos Mylonas, and Stefanos Kollias. Viral: Visual image retrieval and localization. *Multimedia Tools Appl.*, 51(2):555–592, January 2011. ISSN 1380-7501. doi: 10.1007/s11042-010-0651-7. URL <http://dx.doi.org/10.1007/s11042-010-0651-7>.
- G. Tolias and Y. Avrithis. Speeded-up, relaxed spatial matching. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1653–1660, Nov 2011. doi: 10.1109/ICCV.2011.6126427.
- Livia Hollenstein and Ross Purves. Exploring place through user-generated content: Using flickr tags to describe city cores. *Journal of Spatial Information Science*, 1(1): 21–48, 2013.
- Tobias Weyand, Jan Hosang, and Bastian Leibe. An evaluation of two automatic landmark building discovery algorithms for city reconstruction. In *Proceedings of the 11th European Conference on Trends and Topics in Computer Vision - Volume Part II, ECCV'10*, pages 310–323, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-35739-8. doi: 10.1007/978-3-642-35740-4_24. URL http://dx.doi.org/10.1007/978-3-642-35740-4_24.
- David J. Crandall, Lars Backstrom, Daniel Huttenlocher, and Jon Kleinberg. Mapping the world's photos. In *Proceedings of the 18th international conference on World wide web, WWW '09*, pages 761–770, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-487-4. doi: 10.1145/1526709.1526812. URL <http://doi.acm.org/10.1145/1526709.1526812>.
- ImageCLEF. Imageclef, 2014. URL <http://www.imageclef.org>. Accessed: 23/4/2014.
- Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43. ACM, 2008.
- Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of ACM Conference on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., 2009.

- MediaEval. Mediaeval benchmarking initiative for multimedia evaluation, 2014. URL <http://www.multimediaeval.org/about/>. Accessed: 23/2/14.
- Claudia Hauff, Bart Thomee, and Michele Trevisiol. Working notes for the placing task at mediaeval 2013. In *MediaEval*, 2013.
- del.icio.us, 2014. URL <https://delicious.com/>. Accessed: 23/4/2014.
- Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata, 2004. URL <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.
- Jonathan Gemmell, Maryam Ramezani, Thomas Schimoler, Laura Christiansen, and Bamshad Mobasher. The impact of ambiguity and redundancy on tag recommendation in folksonomies. In *Proceedings of the third ACM conference on Recommender systems, RecSys '09*, pages 45–52, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-435-5. doi: 10.1145/1639714.1639724. URL <http://doi.acm.org/10.1145/1639714.1639724>.
- Kilian Quirin Weinberger, Malcolm Slaney, and Roelof Van Zwol. Resolving tag ambiguity. In *Proceedings of the 16th ACM international conference on Multimedia, MM '08*, pages 111–120, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-303-7. doi: 10.1145/1459359.1459375. URL <http://doi.acm.org/10.1145/1459359.1459375>.
- Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd European Conference on The Semantic Web: Research and Applications, ESWC'06*, pages 411–426, Berlin, Heidelberg, 2006a. Springer-Verlag. ISBN 3-540-34544-2, 978-3-540-34544-2. doi: 10.1007/11762256_31. URL http://dx.doi.org/10.1007/11762256_31.
- Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Conference on The Semantic Web, ISWC '08*, pages 615–631, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-88563-4. doi: 10.1007/978-3-540-88564-1_39. URL http://dx.doi.org/10.1007/978-3-540-88564-1_39.
- Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Folkrank: A ranking algorithm for folksonomies. In *LWA*, pages 111–114, 2006b.
- Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *Proceedings of the 4th European conference on The Semantic Web: Research and Applications, ESWC '07*, pages 624–639, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-72666-1. doi: 10.1007/978-3-540-72667-8_44. URL http://dx.doi.org/10.1007/978-3-540-72667-8_44.

- Grigory Begelman, Philipp Keller, Frank Smadja, et al. Automated tag clustering: Improving search and exploration in the tag space. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, pages 15–33, 2006.
- Jonathan Gemmell, Andriy Shepitsen, Bamshad Mobasher, and Robin Burke. Personalization in folksonomies based on tag clustering. *Intelligent techniques for web personalization & recommender systems*, 12, 2008a.
- Jonathan Gemmell, Andriy Shepitsen, Bamshad Mobasher, and Robin Burke. Personalizing navigation in folksonomies using hierarchical tag clustering. In *Proceedings of the 10th international conference on Data Warehousing and Knowledge Discovery, DaWaK '08*, pages 196–205, Berlin, Heidelberg, 2008b. Springer. ISBN 978-3-540-85835-5. doi: 10.1007/978-3-540-85836-2_19. URL http://dx.doi.org/10.1007/978-3-540-85836-2_19.
- Edwin Simpson. Clustering Tags in Enterprise and Web Folksonomies. *HP Labs Technical Reports*, 2008. URL <http://www.hp1.hp.com/techreports/2008/HPL-2008-18.html>.
- Symeon Papadopoulos, Yiannis Kompatsiaris, and Athena Vakali. A graph-based clustering scheme for identifying related tags in folksonomies. In *Proceedings of the 12th international conference on Data warehousing and knowledge discovery, DaWaK'10*, pages 65–76, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15104-3, 978-3-642-15104-0. URL <http://dl.acm.org/citation.cfm?id=1881923.1881931>.
- Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- Yan Ke, Rahul Sukthankar, and Larry Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM International Conference on Multimedia*, pages 869–876, 2004.
- David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, 2006. ISBN 0-7695-2597-0.
- J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
- Jun Jie Foo and Ranjan Sinha. Pruning sift for scalable near-duplicate image matching. In *Proceedings of the eighteenth conference on Australasian database - Volume 63*, pages 63–71, 2007. ISBN 1-920-68244-9.

- P. Turcot and D.G. Lowe. Better matching with fewer features: The selection of useful features in large database recognition problems. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 2109–2116. IEEE, 2009.
- Wei Dong, Zhe Wang, Moses Charikar, and Kai Li. High-confidence near-duplicate image detection. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, pages 1:1–1:8. ACM, 2012. ISBN 978-1-4503-1329-2.
- L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, Nov 1998. ISSN 0162-8828. doi: 10.1109/34.730558.
- V. Pimenov. Fast image matching with visual attention and surf descriptors. In *Proceedings of the 19th International Conference on Computer Graphics and Vision*, pages 49–56, 2009.
- Fernando López-García, Xosé Ramón Fdez-Vidal, Xosé Manuel Pardo, and Raquel Dosil. *Scene Recognition through Visual Attention and Image Features: A Comparison between SIFT and SURF Approaches*, pages 185–198. InTech, 2011. ISBN 978-953-307-222-7.
- Shuo Chen, Cheng dong Wu, Xiao sheng Yu, and Dong yue Chen. Fast scene recognition based on saliency region and surf. In *Intelligent Control and Information Processing (ICICIP), 2011 2nd International Conference on*, volume 2, pages 863–866, july 2011.
- L. Juan and O. Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, 3(4), 2009.
- V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(9):1465–1479, sept. 2006. ISSN 0162-8828. doi: 10.1109/TPAMI.2006.188.
- M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(3):448–461, march 2010. ISSN 0162-8828. doi: 10.1109/TPAMI.2009.23.
- Peter Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semant.*, 5(1):5–15, March 2007. ISSN 1570-8268. doi: 10.1016/j.websem.2006.11.002. URL <http://dx.doi.org/10.1016/j.websem.2006.11.002>.
- Iván Cantador, Martin Szomszor, Harith Alani, Miriam Fernández, and Pablo Castells. Enriching ontological user profiles with tagging history for multi-domain recommendations. In *1st International Workshop on Collective Semantics: Collective Intelligence*

- E@ the Semantic Web (CISWeb 2008)*, June 2008. URL <http://eprints.soton.ac.uk/265451/>. Event Dates: 2 June, 2008.
- Andrés García-Silva, Asunción Gómez-Pérez, Mari Carmen Suárez-Figueroa, and Boris Villazón-Terrazas. A pattern based approach for re-engineering non-ontological resources into ontologies. In *Proceedings of the 3rd Asian Semantic Web Conference on The Semantic Web, ASWC '08*, pages 167–181, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-89703-3. doi: 10.1007/978-3-540-89704-0_12. URL http://dx.doi.org/10.1007/978-3-540-89704-0_12.
- Eirini Giannakidou, Vassiliki Koutsonikola, Athena Vakali, and Yiannis Kompatsiaris. Co-clustering tags and social data sources. In *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management, WAIM '08*, pages 317–324, Washington, DC, USA, 2008. IEEE Computer Society. ISBN 978-0-7695-3185-4. doi: 10.1109/WAIM.2008.61. URL <http://dx.doi.org/10.1109/WAIM.2008.61>.
- Yahoo. Yahoo, 2014. URL <http://www.yahoo.com/>. Accessed: 23/4/2014.
- Hanan Samet. The quadtree and related hierarchical data structures. *ACM Computing Surveys (CSUR)*, 16(2):187–260, 1984.
- Hanan Samet. *Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1990. ISBN 0-201-50300-X.
- Mark N. Gahegan. An efficient use of quadtrees in a geographical information system. *International journal of geographical information systems*, 3(3):201–214, 1989. doi: 10.1080/02693798908941508. URL <http://www.tandfonline.com/doi/abs/10.1080/02693798908941508>.
- Sudipto Guha, Rajeev Rastogi, and Kyuseok Shim. Cure: An efficient clustering algorithm for large databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, SIGMOD '98*, pages 73–84, New York, NY, USA, 1998. ACM. ISBN 0-89791-995-5. doi: 10.1145/276304.276312. URL <http://doi.acm.org/10.1145/276304.276312>.
- Roberto J Bayardo, Yiming Ma, and Ramakrishnan Srikant. Scaling up all pairs similarity search. *WWW*, 7:131–140, 2007.
- Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182, March 2003. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=944919.944968>.

- L.C. Molina, L. Belanche, and A. Nebot. Feature selection algorithms: a survey and experimental evaluation. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 306–313, 2002. doi: 10.1109/ICDM.2002.1183917.
- Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, June 2003. ISSN 0899-7667. doi: 10.1162/089976603321780317. URL <http://dx.doi.org/10.1162/089976603321780317>.
- Fan RK Chung. *Spectral Graph Theory*, volume 92. Amer Mathematical Society, 1997.
- Nikola Ljubešić, Damir Boras, Nikola Bakarić, and Jasmina Njavro. Comparing measures of semantic similarity. In *30th International Conference on Information Technology Interfaces, Cavtat, 2008*.
- Gokavarapu Srinivas, Niket Tandon, and Vasudeva Varma. A weighted tag similarity measure based on a collaborative weight model. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 79–86. ACM, 2010.
- Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, pages 641–650. ACM, 2009.
- Jia Li. A mutual semantic endorsement approach to image retrieval and context provision. In *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR '05*, pages 173–182, New York, NY, USA, 2005. ACM. ISBN 1-59593-244-5. doi: 10.1145/1101826.1101856. URL <http://doi.acm.org/10.1145/1101826.1101856>.
- Peter Kolb. DISCO: A Multilingual Database of Distributionally Similar Words. In Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner, editors, *KONVENS 2008 – Ergänzungsband: Textressourcen und lexikalisches Wissen*, pages 37–44, 2008. URL http://www.linguatools.de/disco/disco_en.html.
- Wikipedia, 2014. URL <http://en.wikipedia.org/gettyimages/>. Accessed: 23/4/2014.
- A. Marzal and E. Vidal. Computation of normalized edit distance and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):926–932, September 1993. ISSN 0162-8828. doi: 10.1109/34.232078. URL <http://dx.doi.org/10.1109/34.232078>.
- Gareth J. F. Jones, Daragh Byrne, Mark Hughes, Noel E. O’Connor, and Andrew Salway. Automated annotation of landmark images using community contributed

- datasets and web resources. In *SAMT*, volume 6725 of *Lecture Notes in Computer Science*, pages 111–126. Springer, 2010. ISBN 978-3-642-23016-5. URL <http://dblp.uni-trier.de/db/conf/samt/samt2010.html#JonesBHOS10>.
- Michael Calonder, Vincent Lepetit, and Pascal Fua. Keypoint signatures for fast learning and recognition. In *Proceedings of the 10th European Conference on Computer Vision: Part I, ECCV '08*. Springer-Verlag, 2008. ISBN 978-3-540-88681-5.
- F. Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI2000 Workshop on Imbalanced Data Sets*, 2000.
- William G Cochran. *Sampling techniques*. John Wiley & Sons, 2007.
- Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002. ISSN 1076-9757. URL <http://dl.acm.org/citation.cfm?id=1622407.1622416>.
- Tony Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, Norwell, MA, USA, 1994. ISBN 0792394186.
- Huan Liu and Hiroshi Motoda. *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 2007. ISBN 1584888784.
- Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7:11–32, 1991. ISSN 0920-5691.
- A.K. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern recognition*, 29(8):1233–1244, 1996.
- Savvas A. Chatzichristofis and Yiannis S. Boutalis. Cedd: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Proceedings of the 6th international conference on Computer vision systems, ICVS'08*, pages 312–322. Springer-Verlag, 2008a. ISBN 3-540-79546-4, 978-3-540-79546-9.
- S.A. Chatzichristofis and Y.S. Boutalis. Fcth: Fuzzy color and texture histogram - a low level feature for accurate image retrieval. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS '08. Ninth International Workshop on*, pages 191–196, may 2008b. doi: 10.1109/WIAMIS.2008.24.
- S.A. Chatzichristofis, Y.S. Boutalis, and M. Lux. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Signal Processing, Pattern Recognition and App.*, 2009.

- Leo Breiman. Random forests. *Mach. Learn.*, 45(1), October 2001. ISSN 0885-6125.
- Lior Rokach. Ensemble-based classifiers. *Artif. Intell. Rev.*, 33(1-2):1–39, February 2010. ISSN 0269-2821. doi: 10.1007/s10462-009-9124-7. URL <http://dx.doi.org/10.1007/s10462-009-9124-7>.
- Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, September 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411. URL <http://dx.doi.org/10.1023/A:1022627411411>.
- K-R Muller, Sebastian Mika, Gunnar Ratsch, Koji Tsuda, and Bernhard Scholkopf. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*, 12(2):181–201, 2001.
- L. Rokach and O. Maimon. Top-down induction of decision trees classifiers - a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 35(4), nov. 2005. ISSN 1094-6977. doi: 10.1109/TSMCC.2004.843247.
- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: an update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145.
- Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998. ISBN 0-201-89685-0.
- Y. Avrithis, G. Toulas, and Y. Kalantidis. Feature map hashing: Sub-linear indexing of appearance and global geometry. In *in Proceedings of ACM Multimedia (Full paper) (MM 2010)*, Firenze, Italy, October 2010.
- Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 9th edition, March 1990. ISBN 0471878766.
- Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 edition, July 2003. ISBN 158488388X.
- Mor Naaman, Shane Ahern, Rahul Nair, and Tye Rattenbury. How flickr helps us make sense of the world: context and content in community-contributed media collections. In *In Proceedings of the 15th International Conference on Multimedia (MM2007)*, pages 631–640. ACM, 2007.

- Andres Garcia, Martin Szomszor, Harith Alani, and Oscar Corcho. Preliminary results in tag disambiguation using dbpedia. In *Knowledge Capture (K-Cap'09) - First International Workshop on Collective Knowledge Capturing and Representation - CKCaR'09*, September 2009. URL <http://eprints.soton.ac.uk/267792/>.
- Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165, September 2009. ISSN 1570-8268. doi: 10.1016/j.websem.2009.07.002. URL <http://dx.doi.org/10.1016/j.websem.2009.07.002>.
- E. Kalogerakis, O. Vesselova, J. Hays, A.A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 253–260, 2009. doi: 10.1109/ICCV.2009.5459259.
- Michele Trevisiol, Hervé Jégou, Jonathan Delhumeau, and Guillaume Gravier. Retrieving geo-location of videos with a divide & conquer hierarchical multimodal approach. In *Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, ICMR '13*, pages 1–8, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2033-7. doi: 10.1145/2461466.2461468. URL <http://doi.acm.org/10.1145/2461466.2461468>.
- P. Cemerlang, Joo-Hwee Lim, Yilun You, Jun Zhang, and J.-P. Chevallet. Towards automatic mobile blogging. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 2033–2036, July 2006. doi: 10.1109/ICME.2006.262613.
- Hongzhi Li and Xian-Sheng Hua. Melog: Mobile experience sharing through automatic multimedia blogging. In *Proceedings of the 2010 ACM Multimedia Workshop on Mobile Cloud Media Computing, MCMC '10*, pages 19–24, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0168-8. doi: 10.1145/1877953.1877961. URL <http://doi.acm.org/10.1145/1877953.1877961>.
- Hiroshi Kori, Shun Hattori, Taro Tezuka, and Katsumi Tanaka. Automatic generation of multimedia tour guide from local blogs. In *Proceedings of the 13th International Conference on Multimedia Modeling - Volume Part I, MMM'07*, pages 690–699, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-69421-8, 978-3-540-69421-2. doi: 10.1007/978-3-540-69423-6_67. URL http://dx.doi.org/10.1007/978-3-540-69423-6_67.
- Xin Lu, Changhu Wang, Jiang-Ming Yang, Yanwei Pang, and Lei Zhang. Photo2trip: Generating travel routes from geo-tagged photos for trip planning. In *Proceedings of the International Conference on Multimedia, MM '10*, pages 143–152, New York, NY,

USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1873972. URL <http://doi.acm.org/10.1145/1873951.1873972>.