

Advanced Social Media Analysis

THÈSE N° 5713 (2013)

PRÉSENTÉE LE 17 MAI 2013
À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
GROUPE EBRAHIMI
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Ivan IVANOV

acceptée sur proposition du jury:

Dr J.-M. Vesin, président du jury
Prof. T. Ebrahimi, directeur de thèse
Dr F. Dufaux, rapporteur
Prof. S. Marchand-Maillet, rapporteur
Prof. S. Süssstrunk, rapporteur



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2013

“Success is the ability to go from one failure to another
with no loss of enthusiasm.”

— Sir Winston Churchill
*British orator, author and
politician (1874–1965)*

To my parents and my sister. . .

Abstract

In recent years, social media have spread around the world with remarkable speed and attracted a significant part of the online community. Social media are web sites as means of interactions among people in virtual communities (social networks) where people create, observe, share, exchange and comment content among them. One of the best aspects of social media is its potential to virally spread content as it goes from one person to a whole network of connected people. Information dissemination through content sharing has resulted in a continuously growing volume of publicly available information on social networks like Facebook, as well as on content sharing web sites, like YouTube or Instagram, and has created new challenges for access, search and retrieval.

This thesis deals with enhancing users' experience when consuming social media. Social media has three key components: users, content, and metadata. Therefore, the overall objective of the thesis is to explore existing and identify new techniques to efficiently enrich each of these three components.

For example, having more metadata in a social media web site, in particular tags, will improve "findability" of photos when performing search by keywords. Therefore, we propose to use similarity between image content (objects in images) and its context (associated tags) to automatically annotate non-tagged images — we refer to this challenge as metadata enrichment.

A significant subset of shared photos in social media is travel related and is associated with geotags. We adapt the previous approach to increase accessibility of non-geotagged photos for the purpose of exploring them when planning vacation or organizing them after coming back from vacation. However, wrong or spam tags can damage the reliability of the social media. Therefore, we investigate and propose new techniques to model users' reliability in tagging — we refer to this challenge as user enrichment.

Furthermore, to ensure metadata reusability and longevity when transferring photos from one content sharing web site to another one, we investigate methods for embedding metadata directly into image files and develop an advanced image management platform which supports this feature. Finally, social media consumption is more effective if it engages users through one of the most natural, pervasive, and gratifying activities of human beings: gaming. Therefore, we investigate and propose a solution on how social gaming can be used to address a challenge of photo album

Abstract

summarization — this challenge is referred in the thesis as content enrichment.

The effectiveness of the proposed methods is demonstrated through a set of experiments on databases of different content type and size.

Keywords: social media, social networks, photo-sharing web site, metadata, tagging, tags, geotags, visual search, image retrieval, image annotation, tag propagation, visual features, object duplicate detection, user reliability, user trust modeling, content trust modeling, spam detection, visual attention, saliency modeling, social gaming, photo album summarization, JPSearch standard.

Résumé

Ces dernières années, les médias sociaux se sont répandus à travers le monde à une vitesse impressionnante et ont attiré une grande partie de la communauté en ligne. Les médias sociaux sont des sites web qui permettent l'interaction entre individus de communautés virtuelles (réseaux sociaux) où les utilisateurs créent, observent, partagent, échangent et commentent du contenu entre eux. Un des meilleurs aspects des médias sociaux est son potentiel à diffuser de manière virale du contenu d'un individu à tout un réseau d'utilisateurs connectés. La diffusion de l'information par le partage de contenu a donné lieu à un volume sans cesse croissant d'informations disponibles publiquement sur les réseaux sociaux tels que Facebook, ainsi que sur les sites web de partage de contenu, tels que YouTube ou Instagram, et a créé de nouveaux défis en ce qui concerne l'accès, la recherche et l'extraction d'informations.

Cette thèse traite de l'amélioration de l'expérience utilisateur lors de la consommation de médias sociaux. Les médias sociaux comportent trois volets principaux : les utilisateurs, le contenu et les métadonnées. Par conséquent, l'objectif général de cette thèse consiste à explorer des techniques existantes et identifier de nouvelles techniques pour enrichir efficacement chacune de ces trois composantes.

Par exemple, utiliser plus de métadonnées dans un site web d'un média social, particulièrement en ce qui concerne les marqueurs, permettra de faciliter la recherche d'images. Par conséquent, nous proposons d'utiliser la similarité entre le contenu des images (objets dans des images) ainsi que son contexte (marqueurs associés) pour annoter automatiquement les images non marquées — nous qualifions ce défi “enrichissement de métadonnées”.

En cas de balises de géolocalisation, nous proposons d'utiliser l'approche précédente pour augmenter l'accessibilité de photos qui ne possèdent pas de balise de géolocalisation dans le but de les examiner lors de la planification de vacances ou pour les organiser lors de son retour de vacances. Cependant, les marqueurs erronés ou spam peuvent altérer la fiabilité des médias sociaux. Par conséquent, nous étudions et proposons de nouvelles techniques pour modéliser la fiabilité du processus de marquage effectué par les utilisateurs — nous qualifions ce défi “enrichissement de l'utilisateur”.

En outre, afin d'assurer la réutilisation des métadonnées et la longévité des photos lors du transfert d'un site web de partage de contenu à un autre, nous étudions des méthodes pour intégrer les

Résumé

métadonnées directement dans les fichiers image et développons une plate-forme de gestion d'image avancée qui prend en charge cette fonctionnalité. La consommation de médias sociaux est plus efficace si elle engage les utilisateurs à travers une des activités les plus naturelles, omniprésente et gratifiante pour les êtres humains : le jeu. Par conséquent, nous étudions et proposons une solution sur la façon dont le jeu social peut être utilisé pour relever le défi de résumer un album photo — nous qualifions ce défi “enrichissement du contenu”.

L'efficacité des méthodes proposées est démontrée par une série d'expériences sur des bases de données de différents types de contenus et de différentes tailles.

Mots-cles : médias sociaux, réseaux sociaux, sites web de partage de photos, métadonnées, marquage, marqueurs, balises de géolocalisation, recherche visuelle, récupération d'image, annotation d'image, propagation de marqueurs, caractéristiques visuelles, détection de doublons, fiabilité des utilisateurs, modélisation de la confiance de l'utilisateur, modélisation de la confiance du contenu, détection de spam, attention visuelle, modélisation de la saillance, jeu social, résumé d'album photo, standard JPSearch.

Acknowledgements

You are now reading a book that took me four years to write. It was a very challenging and at the same time interesting experience on this long way of finding ideas, developing tools, obtaining results and, finally, presenting them. Without help, inspiration, encouragement and assistance from many wonderful people mentioned below, the final result of this thesis would never have been a reality.

First of all, I would like to express my sincere gratitude to my supervisor Prof. Touradj Ebrahimi for giving me the opportunity to work in his group on this challenging and interesting research topic. I appreciate the freedom he gave me in my research and thank him for diverse skills I have developed during my PhD by learning from him.

Then, I would like to thank Prof. Sabine Süssstrunk, Prof. Stéphane Marchand-Maillet and Dr Frédéric Dufaux for accepting to be part of the thesis committee, for the careful reading of my thesis, and for the valuable comments that helped to improve the quality of this thesis. Dr Jean-Marc Vesin was a great president of the jury, and made sure everything went smoothly during the oral exam.

I thank the following institutions and projects for the financial support I have received during the last few years: the Swiss National Foundation for Scientific Research in the framework of NCCR Interactive Multimodal Information Management, and the European Network of Excellence PetaMedia.

My appreciation goes to my co-authors for fruitful collaborations and for contributing considerably to this thesis: Dr Péter Vajda, Dr Jong-Seok Lee, Dr Lutz Goldmann and Dr Pavel Korshunov.

This work would not have been the same without the help and commitment of master students under my supervision: Leila Mirmohamadsadeghi, Denis Filimonov, Keishi Nishida and Sasan Yazdani. I am grateful to them for their excellent work that is referenced within this thesis.

Special thanks go to all other former and current group members: Francesca, Eleni, Ashkan, Martin, Philippe, Gelareh, Sabina, Daniele, and others, for the many activities that we have done together inside and outside the office. At the various Signal Processing Labs I thank all people

Acknowledgements

working there for the great ski weekends, hiking tours, movie sessions, evenings in Satellite, etc.

I am indebted to my Serbian friends for making me feel at home and for spending many unforgettable moments together in Switzerland: Tamara, Marija, Aleksandra, Danica, Nikola, Mihailo, Žarko, Deki, Igor, and many others. My appreciation goes to my flatmates Mirjana, Dušica, Vladimir, and Aleksandar, for being respectful and tolerant. I also want to pay tribute to my friends with whom I shared many pleasant moments at home in Serbia and around the world: Mića, Rade, Nataša, Ksenija, Tijana, Milena, Ćipri, Nikola, and others. I thank them for the great friendship regardless of the distance.

Finally, my most important acknowledgments are towards my parents Svetlana and Krsta, as well as my sister Ivana, for their love, encouragement, advice, and unconditional support throughout my whole life. I thank them for easing my life immensely during all these years. This dissertation is dedicated to them.

Mama, tata, Ivana, hvala vam puno na svemu! Ova doktorska disertacija je vama posvećena.

Contents

Abstract	v
Résumé	vii
Acknowledgements	ix
Contents	xi
List of Figures	xvii
List of Tables	xxiii
List of Abbreviations	xxv
1 Introduction	1
1.1 Motivation and Objectives	1
1.2 Social Media	4
1.2.1 Definition	4
1.2.2 Modeling	7
1.2.3 Examples	8
1.2.4 Research Challenges	12
1.2.4.1 Users	12
1.2.4.2 Content	13
1.2.4.3 Metadata	14
1.3 Contributions	16
1.4 Organization	18
I Metadata Enrichment	21
2 Object-based Tag Propagation for Semi-Automatic Annotation of Images	23
2.1 Introduction	25
2.2 Related Work	26
2.3 System Overview	29
2.3.1 Offline Part	29

2.3.1.1	Feature Extraction	30
2.3.1.2	Clustering	31
2.3.2	Online Part	34
2.3.2.1	Object Selection	34
2.3.2.2	Image Matching	34
2.3.2.3	Object Duplicate Detection	36
2.3.2.4	Tag propagation	39
2.3.3	Implementation Details	39
2.4	Experiments and Results	41
2.4.1	Database	41
2.4.2	Evaluation	42
2.4.3	Results and Analysis	45
2.5	Conclusion	49
3	Saliency-Driven Automatic Extraction of Informative Image Content	53
3.1	Introduction	55
3.2	Related Work	56
3.3	Bottom-up Visual Attention Models	57
3.3.1	A Model Motivated by Cognitive Concepts in Human Vision	58
3.3.2	A Frequency-Tuned Saliency Model	60
3.3.3	A Graph-Based Saliency Model	61
3.4	Application Scenarios	61
3.4.1	Salient Object Detection	62
3.4.2	Visual Search	63
3.5	Experiments	69
3.5.1	Dataset	69
3.5.2	Evaluation	70
3.6	Results and Analysis	72
3.6.1	Results of Salient Object Detection	72
3.6.2	Results of Visual Search	80
3.7	Conclusion	83
II	User Enrichment	87
4	User Trust Modeling for Automatic Landmark Tagging	89
4.1	Introduction	91
4.2	Related Work	94
4.3	Trust Modeling in Geotagging Applications	96
4.3.1	User Reliability Based Model	98
4.3.2	A Coincidence-based Model	99
4.3.3	A Wisdom of Crowds Model	100
4.3.4	An “Authority” Model Based on Goodness of Tags	101

4.3.5	A Co-occurrence Model	101
4.4	System for Automated Landmark Tagging	102
4.4.1	Object Duplicate Detection	102
4.4.2	Tag Propagation	104
4.4.3	User Trust Modeling	106
4.5	Experiments	106
4.5.1	Database	107
4.5.2	Scenarios	107
4.5.3	Evaluation	108
4.6	Results	110
4.6.1	Results of the Tag Propagation Scenario	110
4.6.2	Results of the User Trust Scenario	112
4.7	Conclusion	119
5	Fighting Spammers in Social Tagging Systems	121
5.1	Introduction	123
5.2	Related Work	124
5.3	Distinct Features	126
5.3.1	LegitTags/SpamTags	127
5.3.2	Tags Popularity Based Features	128
5.3.3	User Activity Based Features	128
5.4	Evaluation	130
5.4.1	Database	130
5.4.2	Classification Metrics	131
5.5	Results and Analysis	132
5.6	Conclusion	137
III	Content Enrichment	139
6	“Epitome” – A Social Game for Photo Album Summarization	141
6.1	Introduction	143
6.2	Related Work	144
6.2.1	Automatic Photo Album Summarization	145
6.2.2	Crowdsourcing Through Games	146
6.3	Algorithms for Photo Album Summarization	146
6.3.1	Social Game “Epitome”	146
6.3.2	Automatic Photo Album Summarization	152
6.4	Experiments and Results	153
6.4.1	Performance Evaluation	153
6.4.1.1	Database	153
6.4.1.2	Evaluation Methodology and Results	153
6.4.2	Usability Evaluation	156

Contents

6.4.2.1	Motivation to Play the Game	157
6.4.2.2	Platform	158
6.4.2.3	Privacy Issues	159
6.4.3	Implementation Challenges	160
6.4.4	Summarizing Complete Facebook with the Game	164
6.4.5	Statistics of the Game	164
6.4.6	Advantages and Disadvantages	164
6.5	Conclusion	166
IV	Conclusions	167
7	Conclusions and Future Prospects	169
7.1	Summary of Achievements	169
7.1.1	Metadata Enrichment	169
7.1.2	User Enrichment	170
7.1.3	Content Enrichment	171
7.2	Future Prospects	172
	Appendices	175
A	Database Overview	177
A.1	Introduction	177
A.2	General Purpose Image Database	179
A.3	MIRFLICKR-1M Image Database	182
A.4	Personal Image Database	182
A.5	Image Database of Famous Landmarks	184
A.6	HP Challenge 2010 Image Database	186
A.7	ECML PKDD Discovery Challenge 2008 Bookmarks Database	186
B	Portability of Metadata Across Image Repositories – JPSearch Standard	189
B.1	Introduction	189
B.2	Related Work	190
B.3	JPSearch Overview	191
B.4	“Cheese” – JPSearch - Part 4 Complaint Platform	193
B.4.1	System Overview	193
B.4.2	Object-based Visual Search for Tag Propagation	194
B.4.3	Portability of Metadata	195
B.4.4	Implementation Details	197
B.5	Conclusion	199

C A User Study of the Social Game “Epitome”	201
C.1 Questionnaire	201
C.2 Results of the Questionnaire	207
Bibliography	211
Curriculum Vitae	229

List of Figures

1.1	Ranking the key social media players in terms of number of users.	3
1.2	General model of a social tagging system is represented as a tripartite graph structure which includes three kinds of nodes (objects): <i>users</i> , <i>content</i> and <i>tags</i>	7
1.3	Screenshot from Flickr.	9
1.4	Screenshot from a user profile page in Facebook.	10
1.5	Screenshot from Panoramio.	11
1.6	Screenshot from BibSonomy.	12
1.7	Tag clouds in Flickr showing all time the most popular community tags.	15
2.1	Overview of the system for semi-automatic annotation of objects in images.	30
2.2	An illustration of the process of building a vocabulary tree with branch factor three and only two levels.	33
2.3	Screenshot of the object selection and tagging process in the image management platform “Cheese”.	35
2.4	The object duplicate detection step.	38
2.5	Screenshot of the tag propagation process in the image management platform “Cheese”.	40
2.6	Some example images from the controlled database used for the performance evaluation of the tag propagation method.	41
2.7	Some example images from the distractor database used for the performance evaluation of the tag propagation method.	42
2.8	Performance of the image matching and object duplicate detection parts of the proposed system measured as precision vs. recall curve averaged over all the classes. The evaluation database has 3200 images.	46
2.9	Performance of the image matching and object duplicate detection parts of the proposed system measured as average F-measure vs. object duplicate detection threshold T_O . The evaluation database has 3200 images.	46
2.10	Performance of the image matching and the object duplicate detection parts of the proposed system measured as F-measure across the different classes. The evaluation database has 3200 images.	47

List of Figures

2.11	Performance of the image matching part of the proposed system measured as recall across the different classes. Recall values are shown on the $N = 1000$ predicted images in the image matching step. The evaluation database has more than 1 million images.	47
2.12	Performance of the image matching part of the proposed system measured as recall across the different classes. Recall values are shown for the various numbers of the predicted images in the image matching step. The evaluation database has more than 1 million images.	48
2.13	Performance of the image matching and the object duplicate detection parts of the proposed system measured as: (a) rank factor, and (b) squared rank factor. Results are shown across different object classes. The evaluation database has more than 1 million images.	50
2.14	Performance of the object duplicate detection part of the proposed system measured as F-measure across different object classes. Object duplicate detection is seen as a re-ranking method for validating the object's geometry.	51
3.1	Original image of a dog and its saliency map (the whiter regions indicate the more salient parts).	55
3.2	An overview of a visual attention method by Itti <i>et al.</i>	59
3.3	An overview of a visual attention method by Achanta <i>et al.</i>	60
3.4	The architecture of the system which exploits visual focus of attention for automatic object detection and visual search.	62
3.5	Three examples of features used for visual search: (a) the source image, (b) global color histogram, (c) local SIFT features, and (d) local HOG features. . .	65
3.6	Sample images from the 160 objects within the database of images used for evaluation of visual attention models.	70
3.7	An example of visual comparison between <i>acceptance</i> and <i>accuracy</i> measures.	71
3.8	The distribution of the images with respect to the <i>acceptance</i> values obtained by applying considered visual attention models on the evaluation database of images.	75
3.9	The distribution of the images with respect to the <i>accuracy</i> values obtained by applying considered visual attention models on the evaluation database of images.	76
3.10	The average <i>acceptance</i> values for each of the object classes obtained by applying considered visual attention models on the evaluation database of images.	77
3.11	The distribution of the relative object sizes in the evaluation database of images.	78
3.12	The average <i>acceptance</i> values for each of the object classes obtained by applying considered visual attention models on the evaluation database of images.	79
3.13	The average calculation time for each of the considered visual attention models applied on the evaluation database of images.	81
3.14	Performance evaluation of different features for visual search across object classes: (a) books, (b) buildings, (c) cars and (d) gadgets.	84
3.15	Performance evaluation of different features for visual search across object classes: (a) newspapers, (b) shoes, (c) text and (d) trademarks.	85

4.1	Examples of imprecise or spam tags and incorrect geotags in Flickr.	93
4.2	Overview of the system for geotag propagation in images.	103
4.3	The closed and the open set problems.	105
4.4	Some example images from the database of famous landmarks used for the performance evaluation of the system for automated landmark tagging.	107
4.5	The recognition rate for all landmarks.	111
4.6	The recognition rate across the different landmark categories in the closed set problem.	111
4.7	Precision versus recall curves for the open set problem across the different landmark groups.	113
4.8	F-measure versus detection threshold \hat{S} across different landmark groups. . . .	113
4.9	The distribution of the normalized trust values for different user trust models. .	114
4.10	The distribution of the normalized trust values for all users by different trust models.	115
4.11	The average normalized trust values for different groups of users.	115
4.12	The scatter plots between the normalized trust values from different trust models.	117
4.13	The recognition rate of the geotag propagation system versus the number of the propagated tags.	118
4.14	The recognition rate of the geotag propagation system and the percentage of the propagated tags versus the threshold \hat{T} for the user trust modeling in our socially-driven approach.	118
5.1	An example of spam content on a popular social tagging system: spam bookmarks in Delicious.	124
5.2	The performance of each proposed feature plotted as accuracy, AUC ROC and F-measure.	133
5.3	Discrimination power of the feature <i>LegitTags</i> to separate two types of users, when: (a) used alone, (b) combined with the feature <i>SpamTags</i>	134
5.4	Chi-squared ranking for all tags popularity based features.	135
5.5	Enhancement in the classification performance by aggregating: (a) all tag popularity based features, and (b) all user activity based features.	136
6.1	Screenshots from “Epitome” game: (a) “Select the Best!” game, and (b) “Split it!” game.	147
6.2	System architecture of the “Epitome” game as a Facebook application.	149
6.3	An example of selecting the three most representative photos within one Facebook album through “Epitome” game.	151
6.4	Screenshot from “Epitome” game: “My collage!” page.	151
6.5	Some example photos from the database used for performance evaluation of the “Epitome” game.	153
6.6	Comparison between different visual and time features.	155
6.7	The distribution of the participants’ performance in the “Epitome” game. . . .	156

List of Figures

6.8	The comparison of normalized performance in summarizing photo albums performed by “Epitome” game, automatic photo selection using color histogram and by users who participated in creating the ground truth data.	157
6.9	The most representative photos selected by the proposed method and by making use of color histogram.	158
6.10	Average rank of different motivations to play “Epitome” game.	159
6.11	Different permission pages used in our study: (a) default Facebook permission page, (b) <i>user_photos</i> permission page, (c) <i>user_photo_video_tags</i> permission page, and (d) <i>friends_photos</i> permission page.	161
6.12	Acceptance rate for the default Facebook and permission pages used in our study.	162
6.13	Comparison of the computational time of the initialization phase in “Epitome” game for two approaches.	163
6.14	The distribution of players’ score in “Epitome” game.	165
6.15	The distribution of photos’ scores in “Epitome” game.	165
6.16	The number of photos changed in collages over time for the first two months after “Epitome” game was launched.	166
A.1	Sample images from the 160 objects within the general purpose image database.	180
A.2	Selected objects for 3 different objects from the general purpose image database: Merrell Moab hiking shoe, Golden Gate Bridge (San Francisco), and Starbucks trademark.	180
A.3	Samples of salient objects from the general purpose image database along with their locations shown with white bounding boxes.	181
A.4	Sample images from the MIRFLICKR-1M database.	183
A.5	Sample images from the personal image database.	183
A.6	Sample landmarks for each of the 22 cities within the image database of famous landmarks.	185
A.7	Images for 3 selected landmarks: Berlin (Reichstag), San Francisco (Golden Gate Bridge) and Paris (Eiffel Tower).	185
A.8	Some example images for each of six albums in the HP Challenge 2010 image database.	187
B.1	The global architecture of the JPSearch system.	193
B.2	“Cheese” platform is JPSearch - Part 4 compliant.	196
B.3	An example of JPSearch files embedded in (a) JPEG file format, with the potential to have multiple JPSearch files carried by one image file. (b) Each JPSearch file can contain on its turn multiple JPSearch Core Metadata schema and registered schema.	197
B.4	An example of (a) an annotated object in an image, (b) the corresponding JPSearch file embedded in the JPEG image file, and (c) the corresponding XML representation (JPSearch - Part 4 compliant) of the metadata.	198
C.1	Results for the survey questions 1–5.	207

C.1	Results for the survey questions 6–12.	208
C.1	Results for the survey questions 13–20.	209

List of Tables

1.1	Common social media categories based on their scope and functionality.	6
2.1	Statistical overview of the SIFT and SURF local descriptors extracted from the evaluation database of 3200 images.	31
3.1	Summary of three bottom-up visual attention models used to obtain saliency maps.	58
3.2	Results of salient object detection scenario by making use of considered visual attention models.	73
3.3	Statistical overview of the SIFT and SURF local descriptors extracted from the evaluation database.	82
4.1	Summary of representative recent techniques that combine geographical context and visual content for automatic geotagging of images.	94
4.2	Summary of five trust modeling techniques used for combatting noise and spam in social tagging systems.	97
4.3	Notation used in this chapter.	97
5.1	Summary of tags popularity based features.	128
5.2	Summary of user activity based features.	129
5.3	Statistics of the original database (“ECML PKDD Discovery Challenge 2008” bookmarks database) and a reduced database used for evaluation.	130
5.4	Top classifiers created in Weka.	135
A.1	Summary of the databases used throughout this work.	178
A.2	Summary of the classes and some example objects of the general purpose image database.	179

List of Abbreviations

ACM	Association for Computing Machinery
Ajax	Asynchronous JavaScript and XML
API	Application programming interface
APP	Application marker
AUC ROC	Area under receiver operating characteristic
BBF	Best-bin-first
BoW	Bag-of-words
BRIEF	Binary robust independent elementary features
CAPTCHA	Completely Automatic Public Turing Test to Tell Computers and Humans Apart
CD	Compact Disc
CHOG	Compressed histogram of gradients
CIE	Commission Internationale de l'Éclairage
colHSV	Color in HSV space
colLab	Color in CIE <i>Lab</i> space
CPU	Central processing unit
DOG	Difference of Gaussians
DVD	Digital Versatile Disk
EOH	Edge orientation histogram
EOI	End of Image
EXIF	Exchangeable Image File Format
GBVS	Graph-based visual saliency
GHT	Generalised Hough transform
GPS	Global Positioning System
GWAP	Game with a purpose
HITS	Hyperlink-Induced Topic Search

List of Abbreviations

HOG	Histogram of oriented gradients
HSV	Hue-Saturation-Value
HTD	Homogenous texture descriptor
HTTP	Hypertext Transfer Protocol
HVS	Human visual system
IDF	Inverse document frequency
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
IPTC	International Press Telecommunications Council
ISO	International Organization for Standardization
JDOM	Java Document Object Model
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
JTC	Joint Technical Committee
LBP	Local binary patterns
MAP	Mean average precision
MCC	Matthews correlation coefficient
MPEG	Moving Picture Experts Group
NCCR	National Centres of Competence in Research
NLP	Natural language processing
ODD	Object duplicate detection
OS	Operating system
P2P	Peer-to-peer
PHP	Hypertext Preprocessor
PR	Precision-recall
RAM	Random-access memory
RDF	Resource Description Framework
RGB	Red-Green-Blue
ROI	Region of interest
SIFT	Scale-invariant feature transform
SOI	Start of Image
SQL	Structured Query Language
SURF	Speeded up robust features
SVM	Support vector machine

TF	Term frequency
URL	Uniform resource locator
WG	Working Group
XML	Extensible Markup Language

1 Introduction

1.1 Motivation and Objectives

Back in 2004 and before, “‘Facebook’ didn’t exist; ‘Twitter’ was a sound; the ‘cloud’ was in the sky; ‘4G’ was a parking place; ‘LinkedIn’ was a prison; ‘applications’ were what you sent to college; and ‘Skype’, for most people, was a typo”, as Thomas Friedman, an American journalist and author, put it [1]. Email dominated supreme as the king of communication channels online, and the word ‘friend’ was just beginning its metamorphosis from the rigid uni-dimensional noun it was to the ubiquitous, transformative verb it has become.

Since then, the online landscape has changed immensely. With the widespread adoption of the Internet around the world and the technology development, in particular Web 2.0 and mobile devices, came the rise of social media as a global phenomenon. The best way to define social media is to break it down. Media is an instrument of communication, like a newspaper or a radio, and social media is a social instrument of communication. *Social media* are social networking web sites which allow online users to create and share online information or content they have created, and to build relationships with others online. User-generated content can range from textual content, such as web blogs and encyclopedia articles, to different forms of multimedia content, such as photos, video clips and music. Social media do not only give online users information, but interacts with users while giving them that information. This interaction can be as simple as asking for users’ comments or letting them vote on an article or a photo, or it can be as complex as YouTube’s video recommendation to users based on the ratings of other users with similar interests. If we think of a regular media as a one-way street where users can read a newspaper or listen to a report on television, but they have very limited ability to give their thoughts on the matter [2], then social media is a two-way street that gives users the ability to communicate, too. Some of the most popular social media web sites are Facebook, YouTube, Twitter, Pinterest, Google+, LinkedIn, Flickr and Wikipedia.

Social media have gained considerable popularity in global scale. For example, a Pew Internet survey found that 69 % of Internet users in the U.S. used social networking sites in 2012 [3], up

from 65 % percent in 2011 [4]. Figure 1.1 shows the ranking of the key social media players in terms of number of users. Social media, such as Twitter, Facebook and Instagram, have changed the way information is disseminated. Most people are not clustering around a TV or radio, waiting to see updates. They are taking to their mobile devices and computers, and looking on social media web sites for the most recent news. We saw it during the earthquake in Haiti and the revolutions in Egypt, Tunisia, and Syria, and recently we were seeing it again with the reactions to Hurricane Sandy. According to Mashable, Instagram users posted 10 Hurricane Sandy photos per second [5], and this is just one of many social media sites. Social media has become a medium as big, if not bigger, as traditional media like TV networks and newspapers, with the advantage of nearly instantaneous propagation of information and little to zero entry barriers for content creation. Also, social media have influenced and forever altered the lives of individuals, communities and societies all over the world. For example, consumers are 71 % more likely to make a purchase based on social media referrals [6], and 36 % of consumers use social media to research locations, hotels, and airlines before booking a holiday [7]. These are only a few reasons why it is important to consider social media and to perform an advanced social media analysis from multiple perspectives. This thesis provides a significant contribution in that direction.

Once people have access to the Internet, they tend to use it for producing and consuming content on social media web sites. Content is no longer created and published by individuals, but instead is continuously modified by all users in a participatory and collaborative fashion. For example, users upload their personal photos and share them through online communities, letting other people comment or rate them. In these social environments, photos are usually accompanied with metadata, such as tags, comments, ratings, information about users and their social network. Information dissemination through content and metadata sharing help users to express themselves and to socialize. This trend has resulted in a continuously growing volume of publicly available information (photos, tags, user profiles, etc.) on social networks like Facebook, as well as on content sharing web sites, like Flickr or Instagram, and have created new challenges for access, search and retrieval. For example, Facebook Graph Search was introduced recently and allows users to search through their networks for information about people, photos, places and even local businesses [9]. This semantic search engine is based on both the content of the user and their friends' profiles (more specifically, their likes, expressed interests, visited places), and the relationships between the user and their friends.

This thesis deals with enhancing users' experience when consuming social media. Social media has three key components: users, shared content, and metadata, as we will explain in more details in Section 1.2. Therefore, the overall objective of the thesis is to explore existing and identify new ways to efficiently enrich each of these three components. For example, having more metadata in a social media web site, in particular tags, will improve "findability" of photos when performing search by keywords. Therefore, we propose to use similarity between image content (objects in images) and its context (associated tags) to automatically annotate non-tagged images, and at the same time, to reduce time-consuming manual annotation — we refer to this challenge as metadata enrichment. A significant subset of shared photos in social media

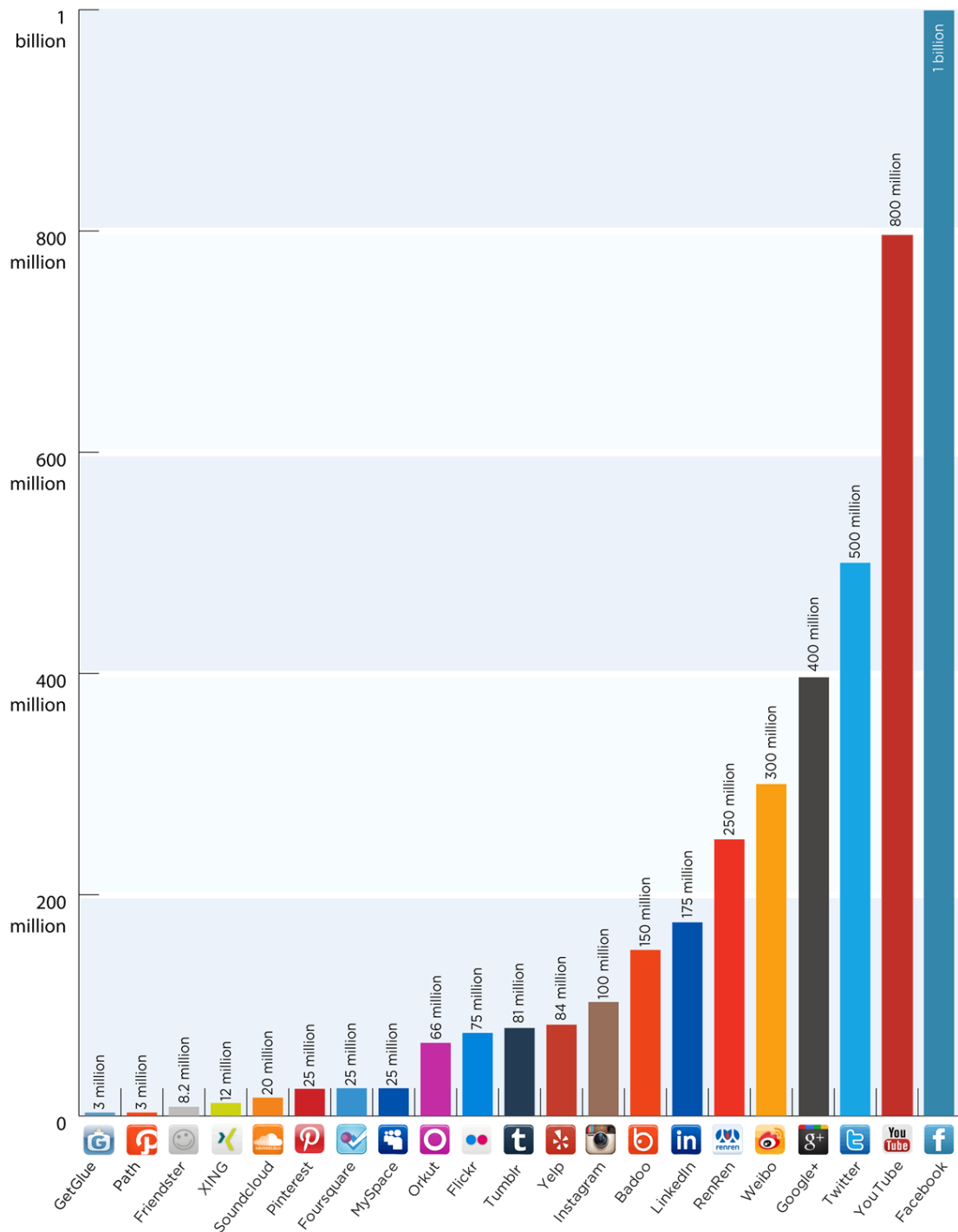


Figure 1.1: Ranking the key social media players in terms of number of users. This is non-exhaustive list of social media web sites. Sites are ranked by the number of total registered users. In a few select cases, sites are valuated in terms of active users (e.g., Facebook) or unique visitors (e.g., YouTube). Data as of November 2012 (image source: [8]).

is travel related and is associated with geotags. We adapt the previous approach to increase accessibility of non-geotagged photos for the purpose of exploring them when planning vacation or organizing them after coming back from vacation. However, wrong or spam tags can damage the reliability of social media. Therefore, we investigate and propose new techniques to model users' reliability in tagging — we refer to this challenge as user enrichment. Furthermore, to ensure metadata reusability and longevity when transferring photos from one content sharing web site to another one, we investigate methods for embedding metadata directly into image files and develop an advanced image management platform which supports this feature. Finally, social media consumption is more effective if it engages users through one of the most natural, pervasive, and gratifying activities of human beings: gaming. Therefore, we investigate and propose a solution on how social gaming can be used to address a challenge of photo album summarization — this challenge is referred in the thesis as content enrichment. The effectiveness of the proposed methods is demonstrated through a set of experiments on databases of different content type and size.

For the benefit of social media users, we have addressed the challenges mentioned before and brought significant contributions to research community through several publications. First, we define social media and provide various research challenges around it in Section 1.2. Then, main contributions of this thesis are summarized in Section 1.3. Finally, the organization of the thesis is presented in Section 1.4.

1.2 Social Media

This section provides an introduction to the basic concepts of social media, which will be widely used throughout this thesis. We define the term social media, introduce a mathematical model to represent social media, and present and discuss various research challenges around social media.

1.2.1 Definition

There have been many efforts to define social media [10]. Our definition of social media refers to web sites as means of interactions among people in virtual communities (social networks) where they create, observe, share, exchange and comment content among them. These web sites apply Web 2.0 technologies [11] to allow users to interact and collaborate with each other as creators of user-generated content, in contrast to other web sites where people are limited to the passive viewing of content. In social media, people contribute user-generated content, which can range from textual content (e.g., web blogs, encyclopedia articles) to different forms of multimedia content (e.g., video clips, photos, music). Content sharing has been made possible during the last years when digital cameras, camera phones and broadband connections became widely available and created new kind of opportunities for communication. All this content is collaboratively indexed by the community and qualitative feedback can be given to either the contributor or the content itself. The system weaves all these contributions into an easily accessible web by

providing cross links wherever possible.

Social media has emerged from different perspectives. Primarily two main classes can be identified. First, systems that start from the notion of a social network and stimulate the interaction between the users by allowing them to share content. Examples include social networks like Facebook, Google+, LinkedIn and Twitter. Second, networks that were primarily created for the distribution or management of content, and use the social network as an overlay that stimulates this content distribution. Examples include content sharing web sites such as YouTube, Instagram, Flickr, Delicious, Pinterest, BibSonomy and CiteULike. Both types of systems have emerged simultaneously over recent years, and both developments have showed that only when both the social and content features are effectively implemented, systems have been able to satisfy a large user community for a long period.

Non exhaustive list of social media categories based on the their scope and functionality are shown in Table 1.1. Of course, there is a cross between many of these categories. For example, Twitter is placed under micro-blogging, but it also belongs to social networking, or Facebook is placed under social networking, but it is also the largest photo-sharing web site.

The popularity of combined content sharing and social features can be explained by various arguments. As social media can be used to share activities or status updates, people can be easily updated on the current well-being of their friends. Hereby, the overall community interaction is stimulated, and it has become easier to maintain solid social relationships with many people. Clever matching algorithms make sure everyone is informed about the most relevant information and even stimulate the exploration of new content or social groups. Every single click on a web site generates a page full of information about people, objects or events. In this way, social media provide answers to and simultaneously stimulate the curiosity inherently present in the human race. As of 2012, social media has become one of the most powerful sources for news updates through platforms such as Facebook, Twitter, Pinterest, Google+, and Tumblr¹, and Internet users continue to spend more time in social media than any other web site [12].

Our definition of social media is built on three key elements: users, content and metadata. Content refers to user created content which may be of very different types, such as textual documents, web pages, images, videos, and music files. One of the more interesting aspects of social media today, apart from the sheer amount of content available online, is the even richer metadata. Users create, upload, and then annotate content, be it through tags, regions of interest, or comments. Annotation is just one of the explicit ways through which metadata is created – and the most visible, but many other actions also lead to the generation of metadata: rating, adding photos or videos to favorite lists, organizing content in sets or collections.

The social media that will be studied in this thesis is about users who share photos and annotate them through tag assignments of textual tags and geotags, and binary ratings. Examples are social

¹ <http://www.tumblr.com>

Chapter 1. Introduction

Table 1.1: Common social media categories based on their scope and functionality.

Category	Social media examples
Blogs	Blogger ³ , LiveJournal ⁴ , WordPress ⁵
Micro-blogs	Twitter ⁶ , Google Buzz ⁷ , Yammer ⁸
Opinion mining	Epinions.com ⁹ , Yelp.com ¹⁰
Photo, video and audio sharing	YouTube ¹¹ , Instagram ¹² , Flickr ¹³ , Pinterest ¹⁴ , Photo-bucket ¹⁵ , Last.fm ¹⁶ , Picasa ¹⁷
Social bookmarking	Delicious ¹⁸ , BibSonomy ¹⁹ , CiteULike ²⁰ , StumbleUpon ²¹
Social networking sites	Facebook ²² , LinkedIn ²³ , Google+ ²⁴ , MySpace ²⁵
Social news	Digg ²⁶ , Reddit ²⁷ , Google Reader ²⁸ , Slashdot ²⁹
Wikis	Wikipedia ³⁰ , Scholarpedia ³¹ , wikiHow ³²

tagging systems in Facebook, Flickr, and Panoramio². The informational value of the relations created by the annotations will be studied by making use of the content of the individual photos. Therefore, many of the presented results will extend to other platforms that employ similar photo-sharing and annotation methods for content description. We have also made valuable contribution in analyzing data from BibSonomy by considering only annotations.

² <http://www.panoramio.com>

³ <http://www.blogger.com>

⁴ <http://www.livejournal.com>

⁵ <http://www.wordpress.com>

⁶ <http://www.twitter.com>

⁷ <http://www.google.com/buzz>

⁸ <http://www.yammer.com>

⁹ <http://www.epinions.com>

¹⁰ <http://www.yelp.com>

¹¹ <http://www.youtube.com>

¹² <http://www.instagram.com>

¹³ <http://www.flickr.com>

¹⁴ <http://www.pinterest.com>

¹⁵ <http://www.photobucket.com>

¹⁶ <http://www.last.fm>

¹⁷ <http://picasa.google.com>

¹⁸ <http://www.delicious.com>

¹⁹ <http://www.bibsonomy.org>

²⁰ <http://www.citeulike.org>

²¹ <http://www.stumbleupon.com>

²² <http://www.facebook.com>

²³ <http://www.linkedin.com>

²⁴ <http://plus.google.com>

²⁵ <http://www.myspace.com>

²⁶ <http://www.digg.com>

²⁷ <http://www.reddit.com>

²⁸ <http://www.google.com/reader>

²⁹ <http://www.slashdot.org>

³⁰ <http://www.wikipedia.org>

³¹ <http://www.scholarpedia.org>

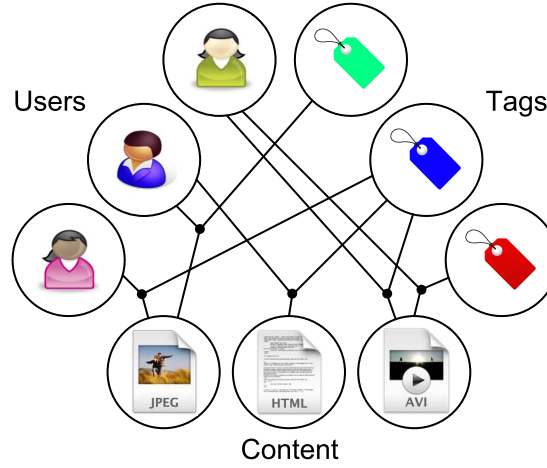


Figure 1.2: General model of a social tagging system is represented as a tripartite graph structure which includes three kinds of nodes (objects): *users*, *content* and *tags*. An edge linking a user, a tag and a content represents a tag assignment.

1.2.2 Modeling

In this section, we introduce the model of a social tagging system that will be used throughout this thesis.

The entities (or objects) that make up the model of a social tagging system [13] are shown in Figure 1.2. The model consists of *users* who interact with the system and among themselves, *content* (resources or documents), which is a piece of information such as photo, video, textual document, or web pages; and *tags*, the descriptions assigned to the piece of the content by users. The action of associating a tag to a content by a user is usually referred to as *tag assignment* [14]. Depending on the system under consideration, a user can assign one or several tags to each content.

A mathematical model of a social tagging system is represented as a hyper-graph structure called folksonomy (in practice also known as collaborative tagging and social indexing) [15, 16], where the set of nodes consists of three kinds of objects: users, content and tags, and hyper-edges connect these objects based on their relations [17]. The hyper-graph G can be defined as a quaternary structure $G = (U, T, R, P)$:

$$U = \{u_1, u_2, \dots, u_k\}, \quad (1.1)$$

$$T = \{t_1, t_2, \dots, t_l\}, \quad (1.2)$$

$$R = \{r_1, r_2, \dots, r_m\}, \quad (1.3)$$

³² <http://www.wikihow.com>

where U represents the set of users u in the system, T is the set of tags t posted by users, R shows the set of resources r and P defines the relation existing between users, resources and tags. A relation linking a user, a tag and a resource represents a post. A post p in a social system can be represented with a triple $p = (u, r, T_u)$ which relates a user u who associated a resource r with a set of n tags $T_u = \{t_1, t_2, \dots, t_n\}$, $n \leq l$. An example of a social tagging system with 3 users, 4 tags and 3 resources is presented in Figure 1.2.

Due to diverse topics of social media analysis covered in this thesis, it would be difficult to stick to unique mathematical symbols throughout the whole thesis. Therefore, each chapter has its own symbols which are well defined around the place where they are introduced.

1.2.3 Examples

The analyses and evaluations in this thesis are conducted on a selection of real-world social media databases. These social media are briefly presented in the following, while further details about the databases are given in Appendix A.

Flickr – An image management and sharing web site where registered users can upload their photos and annotate them with tags and geotags. Social networking is also possible by maintaining a group of contacts and tagging or commenting other's photos. There are also groups to join, forums in which to discuss photographs, and competitions in which to participate. Screenshot from Flickr is shown in Figure 1.3.

Facebook – A social networking service, where registered users may create a personal profile, add other users as friends, join common-interest user groups and exchange messages, photos, videos or music files. Every user's homepage gets updated with their friends' news feed, which highlights information including profile changes, exchanged news and upcoming events. Screenshot from Facebook is shown in Figure 1.4.

Panoramio – A geolocation-oriented photo-sharing web site where uploaded photos are accessed as a layer in Google Earth and Google Maps³³. Registered users organize images using tags, which allow searchers to find images concerning a certain topic such as place name or subject matter. The main purpose of the web site is to allow users to learn more about a given area by viewing the photos that other users have taken at that place. Screenshot from Panoramio is shown in Figure 1.5.

BibSonomy – A website for categorization and sharing of literature references and web site bookmarks. It offers users the ability to store and organize and exchange their bookmarks and publication entries. Both bookmarks and publications can be tagged to help structure and re-find information. As the descriptive terms can be freely chosen, the assignment of tags from different users creates a spontaneous, uncontrolled vocabulary. Screenshot from BibSonomy is shown in Figure 1.6.

³³ <http://maps.google.com>

The screenshot shows a Flickr page for a photo of surfers. The photo is titled "surfers" and has a description: "there's no photoshop at all Canon350D + Lensbabies 3G". The photo is by user "lasgalletas" (Dmitry) and was taken on August 22, 2009, in Buleleng Kabupaten, Bali, ID. The photo has 140 views, 3 favorites, and 4 comments. The photo is associated with several groups: "Bali", "Places", "Days out and about", "Lensbaby", "People Portraits", "Travel Photography", and "The World Through My Eyes". The photo also appears in a "photostream" of 672 photos. The photo is tagged with "bali", "indonesia", "surfers", "surf", "lensbabies 3g", "Travel", "путешествие", "туризм", "tourism", "tourist", "trip", "Бали", "Индонезия", "океан", "серфинг", "серфер", "серферы", and "surfer". The photo is licensed under "Some rights reserved" and is visible to everyone. The page also shows a comment section with a "POST COMMENT" button.

Figure 1.3: Screenshot from Flickr (retrieved in January 2013). A typical image is associated with comments, tags, ratings (favorites) and groups (lists) this image belongs to.

Chapter 1. Introduction



Figure 1.4: Screenshot from a user profile page in Facebook (retrieved in January 2013). A typical user profile page provides basic personal information, lists a few friends, recent photos and any other activity.


Panoramio Sign in SIGN UP

Explore - Community - Upload

World Map > Germany > Sachsen > Dresden


Dresdner Altstadt

See in Google Earth Share on:



by **Juri Kowski**
Selected for Google Earth [?] - ID: 68425013
on Wikimedia Commons

More photos by Juri Kowski



Comments (13)

Georgy K, on April 16, 2012, said:
Beautiful view!
LIKE!
Greetings, Georgy
[Translate](#)

Juri Kowski, on April 18, 2012, said:
Hello Georgy, many thanks for your visit, kind comment and like. Best wishes, Juri.
[Translate](#)

Nenad Obr, on April 28, 2012, said:
Very nice photo. **Like**, Greeting, **Nenad**
[Translate](#)

Dušan Le, on May 3, 2012, said:
L4!
[Translate](#)

SEIMA, on May 13, 2012, said:
Beautiful night shot!! **like**
Greetings from Japan, SEIMA
[Translate](#)

Juri Kowski, on May 18, 2012, said:
Nenad, Dušan and Seima, thanks a lot for your visits, comments and likes. Kind regards from Germany, Juri.
[Translate](#)

Y. Megel, on May 19, 2012, said:
Excellent series of night panoramas(Like!). Beautiful view of night Dresden. Congratulation and good luck. Yuri
[Translate](#)

Juri Kowski, on June 10, 2012, said:
Hello Yuri, I am very happy and proud to receive your appreciations. All the best, Juri.
[Translate](#)

Romulus/Anghel, on June 14, 2012, said:
great night shot!
[Translate](#)

chortis, on August 20, said:
Excellent shot, and reflections on the water are awesome, Greetings from Argentina, Jorge
[Translate](#)

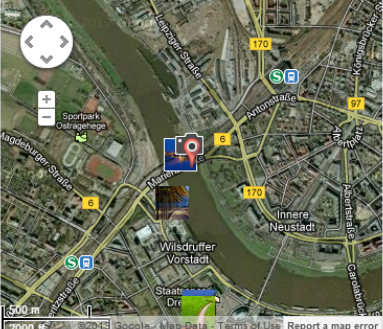
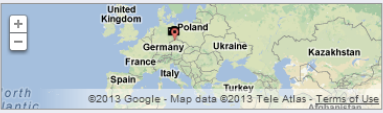


Photo taken in Palaisgarten, 01097 Dresden, Germany
51° 3' 40.14" N 13° 44' 0.07" E
[Misplaced? Suggest new location](#)



Flag photo:
[Report inappropriate or offensive](#)

Tags in this photo:
2012
Altstadt
Cathedral
Deutschland
Dresden
Kathedrale Sanctissimae Trinitatis
Night

Photo stats: [?]
1618 views 1 favorite
10 likes

Groups:
 HOLY PLACES !
 At - Cathedrales - Churc...
 ****City Photos****
 1 My Country Colourful
[« Previous](#) [Next »](#)

Photo details:
Uploaded on March 13, 2012
 Attribution-Share Alike
by Juri Kowski
Extra information
Camera: Panasonic DMC-TZ8
Taken on 2012/03/09 21:30:07
Exposure: 10.000s
Focal Length: 12.80mm
F/Stop: f/6.300
ISO Speed: ISO80
Exposure Bias: 0.00 EV
No flash

Figure 1.5: Screenshot from Panoramio (retrieved in January 2013). A typical image is associated with comments, tags, geotags, ratings (favorites and likes) and groups this image belongs to.

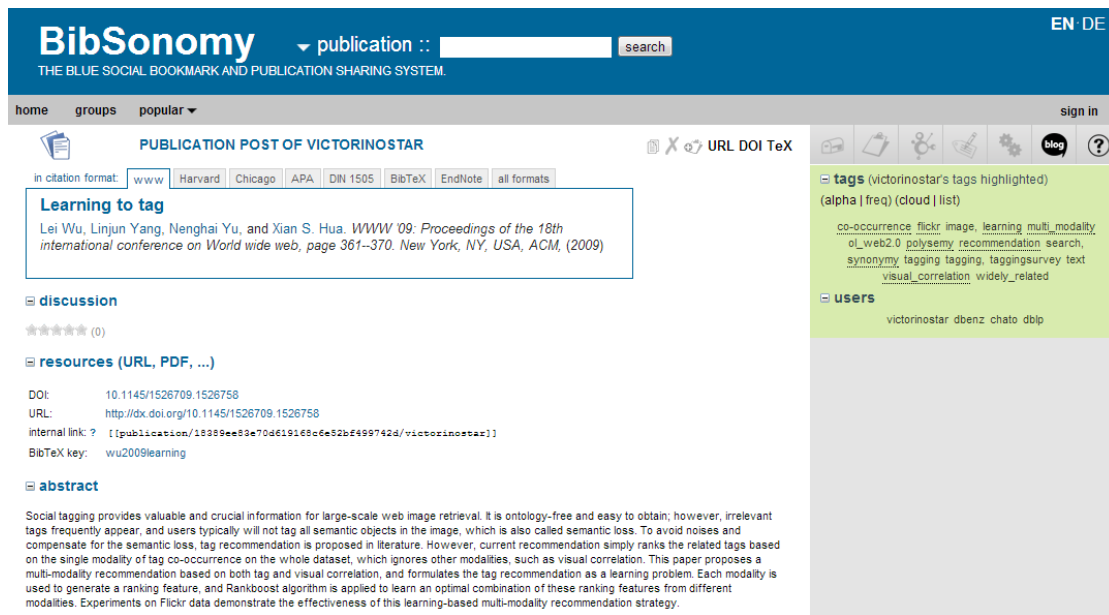


Figure 1.6: Screenshot from BibSonomy (retrieved in January 2013). A typical publication is associated with comments, tags, ratings (stars) and hyperlinks to publication file.

1.2.4 Research Challenges

In this section, we present and discuss diverse research challenges around social media, and more specifically, around each of its involving elements: users, content and metadata. Although these elements equally contribute to the development of social media and are rarely considered solely in analysis, we find it more suitable to present challenges around each of these elements separately. It should be noticed that the purpose of this section is not to give an exhaustive and thorough list of research challenges, but rather to highlight the areas that have been selected as the most important ones.

1.2.4.1 Users

Social media are used to maintain social relationships with many people. Naturally, people have just a handful of intimate *friends* and several hundreds of *acquaintances*. While most people just focus on their closest friends, some users try to connect to as many people as possible and manage to obtain thousands of social relationships. The resulting network is known to have a small-world structure [18, 19] (also known as “six degrees of separation” [20]), which is characterized by high clustering and short path length between any two selected people in the network. This means that anyone in the world can easily be contacted through the friends of his/her friends [21]. Not only it has become easier to maintain relationships within someone’s *social circle*, but social media actively stimulates the discovery of new relations. For example, this feature can be very useful to get introduced to people or companies when searching for a new job. By modeling interaction activity (e.g., communication, tagging) and user similarity (e.g., interest

into heterogeneous topics), new friend relationships can be recommended to users (e.g., [22, 23]) and strength of existing relationships can be estimated (e.g., [24]). Users are forming social graphs with their relations (e.g., friends) and this is the ideal source of fresh data to apply graph analysis (e.g., [25]) and data mining (e.g., [26]) to discover patterns in the behavior of the social media users. For instance, based on the users behavior in tagging, we will present and analyze techniques to estimate the tagging reliability of users in Part II.

Easy distribution and access to each other's self generated content has even made some of the social media popular platforms to display interests (e.g., photography, travelling) or skills (e.g., artistic, technical skills). The open community stimulates the recognition of these interests, and social media have become a new doorway to instant fame. At the same time, social media improve social experience by connecting users with common interests or around specific topic into *groups* [27], and thus, social media have many dedicated members. For example, Facebook is ranked as one of the most visited sites in the world, with over 1 billion monthly active users as of October 2012 [28]. On top of that, there are dozens of other social media with vibrant communities, such as YouTube with more than 800 million users showing up every month [29], Twitter with more than 200 million monthly active users [30], and Google+, which has about 135 million active users [31]. Not all social networks are oriented to non-professional users. For example, LinkedIn, with over 187 million monthly active users as of January 2013 [32], is mostly oriented in establishing professional connections between their users and initiate potential business collaborations.

1.2.4.2 Content

Recent years have seen a transformation in the type of content available on the web. During the first decade of the web's prominence – from the early 1990s onwards – most online content resembled traditional published material: the majority of web users were consumers of content, created by a relatively small amount of publishers. From the early 2000s when social media emerged, user-generated content has become increasingly popular on the web: more and more users participate in content creation and sharing, rather than just in consuming. Social media enables a form of self-expression for users and help them to socialize and share content with other users. Most of user-generated content consists of *text* (blogs, journals, scientific articles, etc.), *images* (holiday photos, creative photography, photojournalism, social activism, etc.), and *video* (musical performances, photojournalism, video blogs, comic shows, etc.). As this content is widely shared on social network web sites, security and privacy are emerging as important and crucial research topics (e.g., [33]). Social media recommendation is one of the most important services to recommend personalized content to users in social networks. Series of discrete characteristics of content are utilized in content-based filtering and ranking approaches to recommend additional content with similar properties (e.g., [34]). Low-level content features play an important role in search and retrieval of similar content given a query example (e.g., [35]). We have also exploited content-based similarity to automatically tag photos shared online, which will be described in Part I, and used social gaming for content summarization in photo-sharing

platforms, which will be presented in Part III.

In the scale that most of the prime social networks operate, even the most common operations are not trivial. The most powerful example is Facebook that has to handle more than 219 billion photos [36] and its servers should serve about 300 million photos shared per day [37]. Instagram users share around 5 million photos per day [37], while Flickr hosts around 1.6 million new photos per day [37]. More than 72 hours of video are uploaded to YouTube every minute [38]. For such volumes of content, management becomes a very crucial issue, which includes high performance data storage and indexing (e.g., [39], [40]). Here we refer to some technologies and tools that most of the social networks use in order to survive the torrents of queries (e.g., clicks, uploads, search queries): BigTable (in use at Google) [41], Apache Cassandra (in use at Facebook, Twitter, Digg) [42], Apache Hadoop (in use at Google, Yahoo!, LinkedIn, eBay) [43].

1.2.4.3 Metadata

Metadata is the information that describes content and facilitates search for a particular content. Associating metadata (annotations) to online content became popular with the launch of a variety of content sharing platforms, like Delicious, Flickr and YouTube. With the introduction of annotations in social media, content indexing has shifted from objective statistical methods to a more subjective categorization. Everybody contributes to the description and organization of the content. Actively or passively, everyone leaves traces that can be used to improve the index of the content. Some of these traces are subjective and therefore inherently related to the individual user, others objectively say something about the content.

Social tagging has emerged as one of the best ways of associating metadata of various types with online content, and it has become a trend now. *Tags* are keywords that the user considers representative of the topic of the content. It allows for multiple, overlapping associations, rather than rigid categories. For example, a Flickr photo of a puppy might be tagged with both tags “puppy” and “cute”, allowing for retrieval based on these two keywords. These keywords can either consist of free text or be selected from a limited vocabulary in the system. As an example, Figure 1.7 displays Flickr’s all time most popular tags in a form of tag cloud [44]. Research topics like why people tag, what influences the choice of tags, kinds of tags, how to model the tagging process, different power laws observed in tagging domain, and how tags are created, are some of the challenges [45]. Opinions about the quality of content can be expressed through *ratings*. Different interface elements let users express a rating in various ways, ranging from scales with five stars to binary judgements in the form of “digg” or “like” buttons. A rating can also be derived from a mouse click or an actual purchase of the item. Ratings create a relation between user and content where the value of the rating determines the strength of the relation. A metadata can also refer to other types of descriptions, like length, author, video format. An increasingly popular annotation is the geographical location of the item. Mostly photographers use these *geotags* to indicate that the photo was not only created at that location, but indirectly show that they have visited that place as well.



Figure 1.7: Tag clouds in Flickr showing all time the most popular community tags, the most popular tags in the last week, as well as the popular tags in the last 24 hours (retrieved in January 2013). Popularity is emphasized by dynamic font sizes. Selecting a tag lists the most recent photos this tag was assigned to.

A few challenges have been identified in research community as important in tagging of social media, namely tag recommendation, tag propagation and tag relevance. For example, tag recommendation approaches suggest appropriate tags to resources (e.g., videos) in order to make it easy for users to search and access information in social systems [46]. To speed up the time-consuming manual tagging process, tags can be automatically assigned to images by making use of tag propagation techniques based on the similarity between image content (e.g., famous landmarks) and its context (e.g., associated geotags), as we will discuss in Part I. Since user-contributed tags are known to be uncontrolled, ambiguous and personalized, one of the fundamental issues in tagging is how to reliably determine the relevance of a tag with respect to the content it is describing [47]. The fact that tags are user-contributed enables spammers to pollute social systems with irrelevant or wrong information (spam) to mislead other users, and to damage the integrity and reliability of social systems. In general, spam on the Internet is created to trick search engines by giving the spam content higher rank in the search results

for advertisement or self-promotion purposes. Various techniques have been proposed in the literature for combatting spam, for example, Google’s PageRank [48] and TrustRank [49]. We will also address this issue in Part II.

1.3 Contributions

The overall objective of this thesis is to explore existing and identify new techniques to efficiently enrich each of the three key components in social media, namely metadata, users, and shared content. According to the target of analysis in social media, we have clustered our contributions into three groups: metadata, user, and content enrichment. Although we have made this clustering, we note here that there is an overlap between each of these groups. For example, our work on embedding metadata in image files to ensure metadata portability can be seen both as metadata and content enrichment, however we find it more suitable to be presented as content enrichment.

Compared to state-of-the-art in social media analysis, we make in this work several contributions, more specifically:

- Research, design and development of a novel interactive online platform that is capable of performing semi-automatic image annotation (tag propagation) for an extensive online database. When a user marks and tags a specific object in a query image, the system performs an object duplicate detection and propagates initial tags to retrieved images in order to automatically annotate images containing similar objects. We demonstrate the effectiveness of the proposed system through a set of experiments considering various classes of objects. Evaluation and thorough analysis is performed on a large-scale database of more than 1 million images, and the results show that the system performs better for text based objects (e.g., books) than for shiny objects (e.g., cars). This work belongs to metadata enrichment and we published it in [50].
- Research and design of a system which exploits visual focus of attention for automatic object detection and visual search. Visual attention approaches are applied for automatic detection of informative content (objects) in images. Assuming that people spontaneously tag the most informative objects in shared images, we further explore the performance of the visual search system for object-based tag propagation which relies solely on automatic object detection. The thorough analysis showed that for achieving the best performance in the object-based tag propagation scenario, additional user input, e.g., adjusting the borders of the bounding box around the predicted object, is necessary. This work belongs to metadata enrichment.
- Research and design of a system for automatic geotag propagation, which adopts user trust modeling for reliable geotags propagation. Tag propagation in images is performed by making use of the similarity between image content (famous landmarks) and its context (associated geotags). In such scenario, however, a wrong or a spam tag can damage

the integrity and reliability of the automated propagation system. Therefore, we suggest adopting user trust models based on a social feedback from the users of the photo-sharing system. Results of the experiments conducted on an image database containing various landmarks show that by propagating tags based on the trust modeling relying on users' tagging behavior, the larger number of tags (more than twice) can be propagated with the same accuracy compared to using other trust models that simply rely on the user contributed tags or if using no trust modeling at all. This work belongs to user enrichment and we published it in [51, 52, 53, 54].

- Research and design of a set of distinct features that can efficiently distinguish between legitimate users and spammers in a social bookmarking system. The proposed features address various properties of social spam and users activities in the system, and provide a helpful signal to identify spam users in the system. The effectiveness of the proposed features is demonstrated through a set of experiments on a database of social bookmarks. We show that aggregation of features leads to the improvement in the classification performance. The performance of different classifiers shows promising results in achieving high accuracy, while keeping false positive rate relatively low. This work belongs to user enrichment and we published it in [52, 55].
- Research, design and development of an approach for photo album summarization through a novel social game “Epitome” as a Facebook application. This game represents an efficient summarization tool which can help people to receive a quick overview of an album containing large number of photos. The proof of concept of the game is demonstrated and validated through a set of experiments on several photo albums. The results are promising and show that the summarization game outperforms automatic visual summarization methods. The usability of this game is validated by making use of a questionnaire. The results of our user study showed that the main motivation for a player of the game is to watch his/her friends' photos and obtain his/her album summarization. This work belongs to content enrichment and we published it in [56, 57, 58].
- Research, design and development of an advanced image management platform for online use, called “Cheese”. For improved interoperability between different image repositories and applications, the platform supports the export and import of image files with embedded metadata in JPSearch - Part 4 compliant format. To the best of our knowledge, “Cheese” is the world's first platform to be JPSearch - Part 4 compliant, which gives the users confidence in the longevity and portability of their annotations. We summarize the main points of the JPSearch standard, and put more focus on how the use and reuse of metadata is established through the JPSearch - Part 4 standard. This work belongs to content enrichment and we published it in [50, 59, 60].
- Design and collection of a general purpose image database containing 3200 images of 8 object classes, such as books, buildings, cars, gadgets, newspapers, shoes, text, and trademarks. All images from the database are associated with human annotated ground truth in terms of tags that define class and object name. In addition, the most salient object

in every image is located and its borders are outlined with a bounding box (rectangle). The database is used to assess the robustness of the tag propagation method with respect to different object classes (results are reported in [50]), as well as, to examine the performance of different visual attention models. The database is also suitable for image retrieval applications, and object detection or recognition.

- Design and collection of a database of images depicting famous landmarks. The database consists of 1320 images in total from 22 cities and 66 geographically unique landmarks. All images from the database are associated with human annotated ground truth in terms of tags that define city and landmark, and several other tags describing landmarks depicted in images. Within this work, the image database is used to demonstrate the effectiveness of our method for modeling the user trust (reliability) in geotagging, and results are reported in [51]. The database is also suitable for image retrieval applications.

The contributions presented in this thesis have been disseminated in the research community through several publications by the author and collaborators. These publications are mentioned previously for every contribution and the complete list of the publications is presented in the author's Curriculum Vitae at the end of this thesis. Furthermore, in the beginning of every chapter, after abstract, the list of the publications that address the content of that particular chapter is given.

1.4 Organization

The organization of this thesis is closely fit to the aforementioned contributions. Since the thesis covers diverse areas that belong to social media analysis, it is written in a way that each chapter is self contained and can be read independently from the rest.

The rest of this thesis is organized into several chapters. Each chapter (except this one, Chapter 7 and Appendices) starts with a short abstract, an introduction and motivation to address one of the challenges, and it continues with discussion of the related work. Then, proposed technique(s) is (are) presented in details, evaluation methodology is described and finally, results and thorough analysis are given. A chapter concludes with summary of achievements. At the end of the thesis, Appendices provide more details on some important topics. Every chapter and appendix is shortly summarized below.

The thesis is divided into three parts according to the objective of analysis in social media, i.e., metadata, user and content enrichment.

Part I addresses the challenge of efficient enrichment of metadata in social media.

In Chapter 2, we propose a semi-automatic approach for image annotation (tag propagation) and evaluate it on an extensive online database of more than 1 million images. Furthermore, we present an interactive online platform, called "Cheese", which applies developed approach to

minimize the users' tedious and time-consuming manual annotation process.

In Chapter 3, we go one step further and explore visual attention approaches to automatically detect informative content (objects) in images. Moreover, we evaluate the performance of the visual search system for object-based image annotation which relies solely on automatic object detection.

Part II focuses on improving information about users, namely, distinguishing between reliable (trustworthy) and non-reliable users in tagging.

In Chapter 4, we propose a system for automatic geotag propagation in images based on the similarity between image content (famous landmarks) and its context (associated geotags). For reliable geotags propagation, we suggest adopting user trust model based on a social feedback from the users of the photo-sharing system. The effectiveness of the proposed model is then evaluated by comparing it with different user trust models.

Chapter 5 focuses on applying machine learning approach to facilitate the process of identifying legitimate users and spammers in a social tagging system. We propose and analyze a set of distinct features based on user behavior in tagging and tags popularity, and demonstrate their effectiveness through a set of experiments on a database of social bookmarks.

Part III shows how content (photos) can be efficiently enriched to increase users' experience when consuming social media.

Chapter 6 presents an approach for photo album summarization through a novel social game "Epitome" as a Facebook application. The performance and usability of the game are evaluated and promising results are presented.

Chapter 7 concludes the thesis with the summary of the main achievements, discussion of the presented approaches, some open issues regarding this work and prospects for future research.

Appendix A presents in details the databases used for evaluation throughout this thesis, with emphasis on the databases created on our own.

Appendix B addresses an issue of interoperability between different image repositories and applications. We present an important feature of the advanced image management platform "Cheese". This feature complies with JPSearch - Part 4 standard and ensures the longevity and portability of metadata by embedding them in image files themselves.

In Appendix C, we first list all questions from the questionnaire used for usability evaluation of the social game "Epitome", and then present results for each of the questions.

Metadata Enrichment **Part I**

2 Object-based Tag Propagation for Semi-Automatic Annotation of Images

Over the last few years, social network systems have greatly increased users' involvement in online content creation and annotation. Since such systems usually need to deal with a large amount of multimedia data, it becomes desirable to realize an interactive service that minimizes tedious and time-consuming manual annotation. In this chapter, we propose an interactive online platform that is capable of performing semi-automatic image annotation for an extensive online database. First, when the user marks a specific object in a query image, the system performs an object duplicate detection and returns the search results with images containing similar objects. Then, when the user enters his/her tags for the selected object, they are propagated to retrieved images in order to automatically annotate images containing similar objects. The user has the possibility to select a subset of matched images to which he/she wants to propagate initial tags. Different techniques to speed-up the process of indexing and retrieval of the large number of images are presented in this chapter and their effectiveness is demonstrated through a set of experiments considering various classes of objects.

Portions of this chapter are published in:

I. Ivanov, P. Vajda, L. Goldmann, J.-S. Lee, and T. Ebrahimi, “Object-based tag propagation for semi-automatic annotation of images,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 497–506, Mar. 2010

2.1 Introduction

In recent years, social networks, digital photography and web-based personal image collections have gained popularity. A social network service typically focuses on building online communities of people who share interests and activities, or desire exploring the interests and activities of others. At the same time, they have become a popular way to share and to disseminate information, creating new challenges for access, search and retrieval. For example, users upload their personal photos and share them through online communities, letting other people comment or rate them. This trend has resulted in a continuously growing volume of publicly available photos on content sharing web sites like Flickr, or Instagram, as well as social networks like Facebook. For instance, 219 billion photos have been uploaded on Facebook since fall 2005 [36], as already pointed out in Chapter 1.

In these environments, photos are usually accompanied with metadata, such as comments, ratings, information about users and their social network. Moreover, a recent trend is also to “tag” them. Tags are short textual annotations used to describe photos in order to provide meaningful information about them. The most popular tags in photo sharing web sites such as Flickr are usually related to the location where the photo was taken (e.g., San Francisco or France), the objects/persons appearing in the photo (e.g., baby, car or house), or the event/time when the photo was taken (e.g., wedding or summer) [61, 62].

Annotations and their association with images provide a powerful cue for their grouping and indexing. This cue is also essential for image retrieval systems to work in practice. The current state-of-the-art in content-based image retrieval systems has not yet delivered widely accepted solutions, except for some very narrow application domains, mainly because of the semantic gap problem, i.e., it is hard to extract semantically meaningful information using just low-level features [63]. In social networks, the success of Flickr and Facebook proves that users are willing to provide this semantic context (“subjective” users’ impressions) through manual annotations [62], which can help to bridge the semantic gap, and therefore improve the results of visual content search engines. Different users who annotate the same photo can provide different annotations, which enrich information about that photo.

However, manual tagging of a large number of photos is a time-consuming task. Users typically tag a small number of shared photos only, leaving most of them with incomplete metadata. This lack of metadata seriously impairs search, as photos without proper annotations are typically much harder to find. Therefore, robust and efficient algorithms for automatic or semi-automatic tagging (or tag propagation) are desirable to help people organize and browse large collections of personal photos in an efficient way.

The main novelty of the research work presented in this chapter comes from the application which realizes an interactive service that minimizes the users’ tedious and time-consuming manual annotation process, and the evaluation of the object duplicate detection part of the system. We propose an interactive online platform which is capable of performing semi-automatic image

annotation for an extensive online database of images containing various object classes. Since the most salient regions in images usually correspond to specific objects, we consider object-based tagging within the system. First, when the user marks a specific object in a query image, the system performs an object duplicate detection and returns the search results with images containing similar objects. Then, when the user enters his/her tag (or tags) for the selected object, it is (they are) propagated to retrieved images in order to automatically annotate images containing similar objects. The user has the possibility to select a subset of matched images to which he/she wants to propagate initial tag (or tags). Similarly, the system can perform tag recommendation to suggest tags for the object in the query image by showing the corresponding tags of all matched objects within the retrieved images and the user can then select appropriate tags among them. The presented tag propagation approach is implemented in an advanced image management platform for online use³⁴ called “Cheese”, which is previously described in [64]. Beside standard features such as image upload, tagging and keyword based search, the platform offers the user visual similarity based search, object-based tagging and semi-automatic tag propagation.

The remaining sections of this chapter are organized as follows. We introduce related work in Section 2.2. Section 2.3 describes our approach for the interactive online platform and discusses the tag propagation scenario. Experiments and results are discussed in Section 2.4. Finally, Section 2.5 concludes the chapter with a summary and some perspectives for future work.

2.2 Related Work

The proposed system is related to different research fields including visual content analysis, social networking and tagging. The goal of this section is to review the most relevant works on visual search in images, human tagging and various approaches for tag propagation.

Tagging images is a very time consuming process and tagging objects within images even more. Therefore, it is necessary to understand and increase the motivation of users to annotate images. Ames and Naaman [65] have explored different factors that motivate people to tag photos in mobile and online environments. One way is to decrease the complexity of the tagging process through tag recommendation which derives a set of possible tags from which the user can select suitable ones. Another way is to provide incentives for the user in form of entertainment or rewards. The most famous examples are games with a purpose (GWAPs), such as ESP Game [66] and Peekaboom [67], developed for collecting information about image content. More information about these games will be given in Section 6.2.2. Another example is LabelMe, a web-based tool that allows easy image annotation and sharing of such annotations [68]. Using this tool, a large variety of annotations is collected spanning many object categories (cars, people, buildings, animals, tools, etc.). Another method is shown in the TagCaptcha image annotation system, where the authors have proposed to obtain image annotations as side product of the challenging CAPTCHA process [69].

³⁴ <http://cheese.epfl.ch>

Since manually tagging a large number of photos is still a tedious and time-consuming task, automatic image annotation has received a lot of attention recently. Automatic image annotation is a challenging task which has not been solved in a satisfactory fashion for real-world applications. The most recent techniques for automatic image annotation and semantic interpretation using social images and tags are surveyed in [70]. Most of the solutions are developed for a specific application and usually consider only one tag type, e.g., faces, locations, objects or events. Some of the representative approaches are discussed below.

Berg *et al.* [71] propose an approach to label *people*, i.e. assign names to faces within newspapers (images with captions). They cluster face images visually in appropriate discriminant coordinates and apply natural language processing techniques to the caption to check whether the person whose name appears in a caption is depicted in the associated image or not. Picasa also provides a service for name tagging which automatically finds similar faces in a photo collection.

Annotating images with geographical information such as *landmarks* and *locations* is a topic which has recently gained increasing attention. Ahern *et al.* [72] have developed a mobile application called ZoneTag³⁵ which enables the upload of context-aware photos from mobile phones equipped with a camera to Flickr. In addition to automatically supplying the location metadata for each photo (provided by a GPS device or mobile phone), ZoneTag supports context-based tag suggestions in which tags are provided from different sources including past tags from the user, the user's social network, and external geo-referenced data sources like Yahoo! Local³⁶ and Upcoming³⁷. More details on recent techniques that combine geographical context and visual content for automatic geotagging of images are provided in Section 4.2.

Another application that combines textual and visual techniques for objects and events tagging has been proposed by Quack *et al.* [73]. The authors developed a system that crawls photo collections on the internet to identify clusters of images referring to a common object (physical items on fixed locations), and events (special social occasions taking place at certain times). The clusters are created based on the pair-wise visual similarities between the images, and the metadata of the clustered photos is used to derive a label for the clusters. Finally, Wikipedia articles are attached to the images and the validity of these associations is checked. Gammeter *et al.* [74] extends this idea towards object-based auto-annotation of holiday photos in a large database that includes landmark buildings, statues, scenes, pieces of art, with help of external resources such as Wikipedia. In both papers, [74] and [73], GPS coordinates are used to pre-cluster objects which limit classes of objects to landmarks and buildings. Another limitation is that GPS coordinates may not be always available. In contrast, our work considers extensive online database of various object classes, such as landmark buildings, cars, cover or text pages of newspapers, shoes, trademarks and different gadgets like mobile phones, cameras, watches.

Lindstaedt *et al.* [75] developed a tag recommendation system called tagr, which deals with

³⁵ <http://zonetag.research.yahoo.com>

³⁶ <http://local.yahoo.com>

³⁷ <http://upcoming.yahoo.com>

pictures depicting fruits and vegetables. They combine three types of information: visual content, text and user context. At first, they group annotated images into classes using global color and texture features. The user defined annotations are then linked with the images. The resulting set of tags for visually similar images is then extended with synonyms derived from WordNet. When the user uploads an untagged image, it is assigned to one of the classes and corresponding tags are recommended to the user. In addition, this system analyzes tags which the user assigns to the images and returns the profiles of users with similar tagging preferences. This method has been proven to be effective to recommend tags for a set of selected fruits and vegetables, but it cannot be applied to other classes of objects, which limits its applicability. Other approaches for automatic image annotation consider only the context. Sigurbjörnsson and van Zwol [62] developed a system which recommends a set of tags based on collective knowledge extracted from Flickr. Given a photo with user-defined tags, a new list of candidate tags is derived for each of the user-defined tags, based on tag co-occurrence. The lists of candidate tags are aggregated, tags are ranked, and a new set of recommended tags is provided to the user. Liu *et al.* [76] provided a comprehensive survey of the technical achievements in the research area of tag assignment, tag ranking and tag refinement for social images.

Some examples of the online applications that provide visual search services are described in the following.

TinEye³⁸ is a service for near-duplicate image search. One can submit an image to TinEye to find out where it comes from, how it is being used, if modified versions of the image exist, or to find its higher resolution versions. It was claimed in [77] to be the first image search engine on the web to use image identification technology rather than keywords or metadata. When a user submits an image to be searched, TinEye creates a distinct and compact digital signature of the image using a perceptual hash derived from color features. Then, it compares this signature to stored signatures of indexed web images to retrieve matches. TinEye does not typically find similar images (i.e. a different image with the same object in it), but rather finds exact and altered copies of the submitted image including those that have been cropped, color adjusted, resized or slightly rotated. However, it is sensitive to rotation and heavy scaling. TinEye constantly crawls the web and updates its image database regularly. As of October 2012, it has indexed around 2.2 billion images from the web.

INRIA has developed an image search platform, called Bigimbaz³⁹. Users can upload their own photos or select already stored photos to query for similar images in a database of more than 10 million images. It uses binary representation of local binary patterns descriptors extracted from both query and database images, and searches the database organized as an inverted file to speed up this process. The search is then refined using the Hamming embedding and burstiness algorithms, and finally, the most similar images are matched with the query using weak geometric consistency constraints and returned to the user as results. A comparison with the state-of-the-art shows the interest of this approach when high accuracy is needed [78].

³⁸ <http://www.tineye.com>

³⁹ <http://bigimbaz.inrialpes.fr>

Since recently, a class of commercial applications and services enables visual search for mobile phones. These applications, such as SnapTell⁴⁰, Kooaba⁴¹ or Google Goggles⁴², can be used for identifying products, comparison shopping, finding information about movies, CDs or products of the visual arts. They use the camera phone to initiate search queries about an object in visual proximity to the user. A set of image feature descriptors is used to assess the similarity between the query photo and each database photo. Kooaba is based on SURF features and detects specific objects, such as posters, CDs, DVDs, books, and game covers. Snaptell detects objects through local features and accumulated signed gradient matching. Google Goggles is the most recent commercial application from Google. It can detect logos, book covers, artworks, places and wines using visual and GPS information. All of these commercial applications use their own database and do not allow users to create and annotate objects.

The tag propagation and tag recommendation are nowadays very important in environment such as social network, since they provide efficient information for grouping or retrieving images. The system proposed in this chapter provides these functionalities in an interactive way. The novelty is that image annotation is performed at the object level by making use of content based processing. It does not consider context, such as text or GPS coordinates, which may limit its applicability. This approach is suitable for all kinds of objects, such as trademarks, books, newspapers, and not just buildings or landmarks. One of the most suitable applications based on this approach is when several people who went on the same trip want to share their own photos and explore photos of the other people that were taken on that trip. The proposed approach can be used to easily explore similar photos of the same scene or objects captured by different people.

2.3 System Overview

In this section, we present our method for object-based tag propagation. Since the manual annotation of all the instances of an object within a large set of images is very time-consuming, the system offers tag propagation of marked and tagged objects. Image annotation is performed at the object level, outlining the object with a bounding box. The system architecture is illustrated in Figure 2.1.

2.3.1 Offline Part

The goal of the offline processing is to preprocess uploaded images in order to allow efficient and interactive object tagging. It starts by describing each image with a set of sparse local features. In order to speed up the feature matching, the features of all images are grouped hierarchically into a tree representation.

⁴⁰ <http://www.snaptell.com>

⁴¹ <http://www.kooaba.com/>

⁴² <http://www.google.com/mobile/goggles>

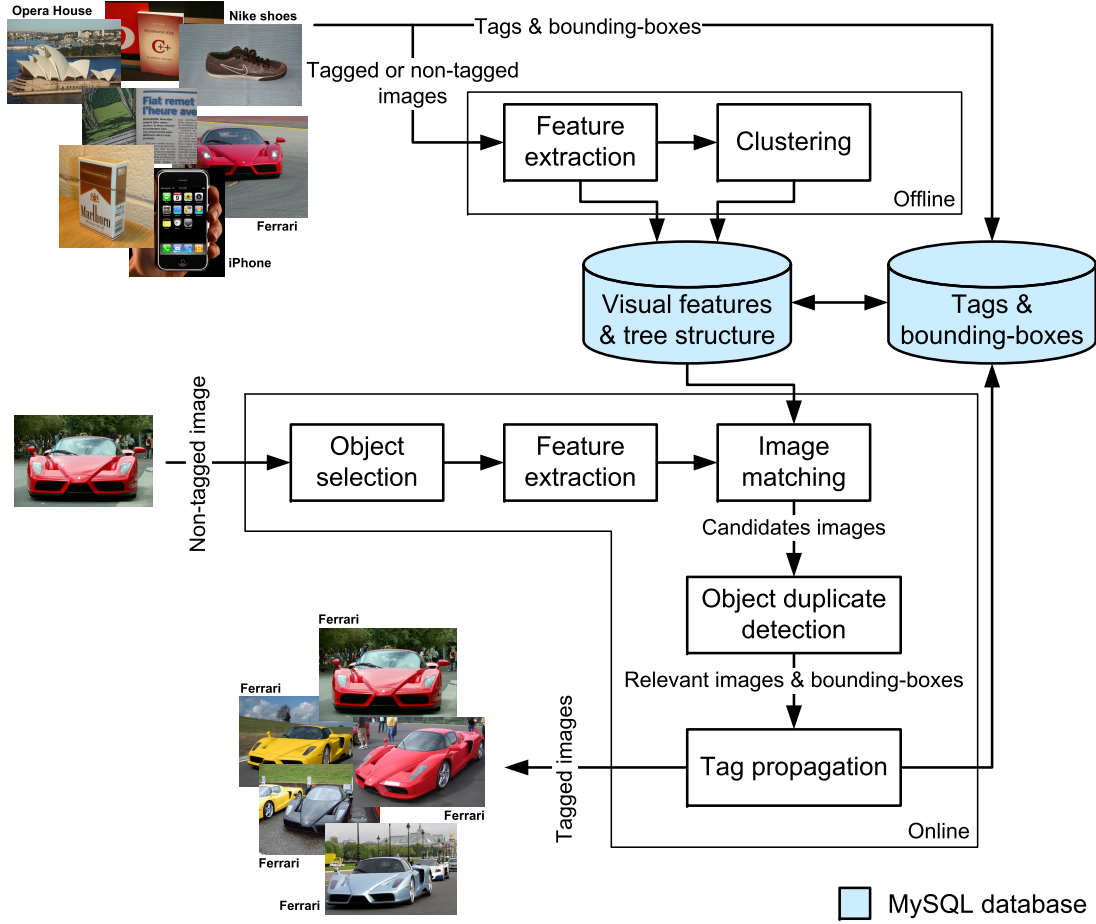


Figure 2.1: Overview of the system for semi-automatic annotation of objects in images.

2.3.1.1 Feature Extraction

For a robust and efficient object localization, sparse local features are adopted to describe the image content. Salient regions (interest or key points) are detected using the Fast-Hessian detector [79], which is based on approximation of Hessian matrix detector. The position and scale are computed for each of the regions and will be used for the object duplicate detection (described in Section 2.3.2.3). As the focus of this chapter is on providing efficient technique for indexing and matching large number of images, we here summarize feature extraction process, and provide detailed description in Section 3.4.2.

The detected regions are described using speeded up robust features (SURF) [79], which can be extracted very efficiently and are robust to arbitrary changes in viewpoints. The goal of SURF is to approximate the popular and robust features based on the scale-invariant feature transform (SIFT) [80].

We selected combination of Fast-Hessian detector and SURF descriptor as an efficient couple in

Table 2.1: Statistical overview of the SIFT (of dimensionality 128) and SURF (of dimensionality 64) local descriptors extracted from the evaluation database of 3200 images. The resolution of every image is 500×500 pixels at maximum. The average image size is 27.5 kB. All timings were obtained using Intel Core2 Duo P9400 CPU running at 2.4 GHz with 4 GB of RAM memory. The average, minimum and maximum number of features per image, as well as, feature extraction time and size is given.

Metric	Feature	Measurement
Average number of features per image	SIFT	1035.1
	SURF	548.3
Minimum number of features per image	SIFT	9
	SURF	21
Maximum number of features per image	SIFT	4781
	SURF	2273
Average time of feature extraction per image [s]	SIFT	0.46
	SURF	0.09
Average feature size per image [kB]	SIFT	121.5
	SURF	17.1

terms of computational time and storage space, as well as, good performance. The Fast-Hessian detector finds in average 550 interest points per image, of which each will be associated with a feature vector as calculated by the SURF descriptor. We specifically use this interest point detector because other detectors often find many more. For instance, the Hessian affine detector (that is usually combined with SIFT descriptor [81]) typically detects 1000s of interest points per image, resulting in a descriptor that can easily take 1 MB of space. Representing images with as few, but still sufficiently strong, interest points as possible is very important to reduce required storage space as well as reduce time needed to compare images with each other. Comparison of SIFT (detected by the Hessian affine detector) and SURF features (detected by the Fast-Hessian detector) extracted from the evaluation database is presented in Table 2.1 with respect to the number of features per image, as well as, extraction time and feature size. SURF features take less time and storage space compared to SIFT features.

2.3.1.2 Clustering

For the object tagging, features of a selected object have to be matched against all the images in the database. Therefore a fast matching algorithm is required to ensure interactivity of the application. Hierarchical clustering is applied to group the features according to their similarity. This improves the efficiency of the feature matching since a fast approximation of the nearest neighbor search can be used.

Hierarchical k-means clustering applied on a set of image feature descriptors (or visual words) is

used to derive the vocabulary tree, similar to the one described in [82]. If the vocabulary tree has L levels excluding the root node and each interior (intermediate) node has C children (or branches), then a fully balanced vocabulary tree contains $K = C^L$ leaf nodes. Figure 2.2 shows an example of vocabulary tree with $L = 2$, $C = 3$, and $K = 9$. The vocabulary tree for a particular database of images is constructed by performing hierarchical k-means clustering on a set of image feature descriptors, as shown in Figure 2.2. Initially, C large clusters are generated from all the descriptors by ordinary k-means with an appropriate distance function, like L2 norm or symmetric KL divergence [83]. Then, for each large cluster, k-means clustering is applied to the descriptors assigned to that cluster to generate C smaller clusters. This recursive division of the descriptor space is repeated until there are enough bins to ensure good classification performance. Within the tree, parent nodes correspond to the cluster centers derived from the features of all its children nodes and leaf nodes correspond to the real features within the images. The clustering leads to a balanced tree with a similar depth for all the leaves. Typically, $L = 6$ and $C = 10$ are selected [82], in which case the vocabulary tree has $K = 10^6$ leaf nodes.

Since the importance of the individual visual words (i.e., nodes in the tree structure) may differ among the images in the database, weights w_i are assigned to each of the corresponding nodes i . These weights are equivalent to the inverse document frequency (IDF) commonly used in text retrieval, which is defined as:

$$w_i = \log \left(\frac{N}{N_i} \right), \quad (2.1)$$

where N is the number of images in the database and N_i is the number of images which have features in the subtree, if the i -th node is considered as a root of this subtree. The basic idea behind IDF is that the importance of a visual word is higher if it is contained in only a few images. Furthermore, the importance of a visual word i in relation to an individual image j is considered using the term frequency (TF), which is defined as:

$$m_{ij} = \frac{N_{ij}}{\sum_k N_{kj}}, \quad (2.2)$$

where N_{ij} is the number of occurrences of a visual word i within an image j and the denominator is the number of occurrences of all features within image j .

Given this TF-IDF weighting scheme the overall weight d_{ij} for a visual word i within an image j is given as:

$$d_{ij} = m_{ij} \cdot w_i, \quad (2.3)$$

which can be combined into a vector \mathbf{d}_j . Vectors \mathbf{d}_j are associated to leaf nodes in the vocabulary tree. This vector will be matched to the one extracted from the query image to compute the similarity within the image matching step described in Section 2.3.2.2.

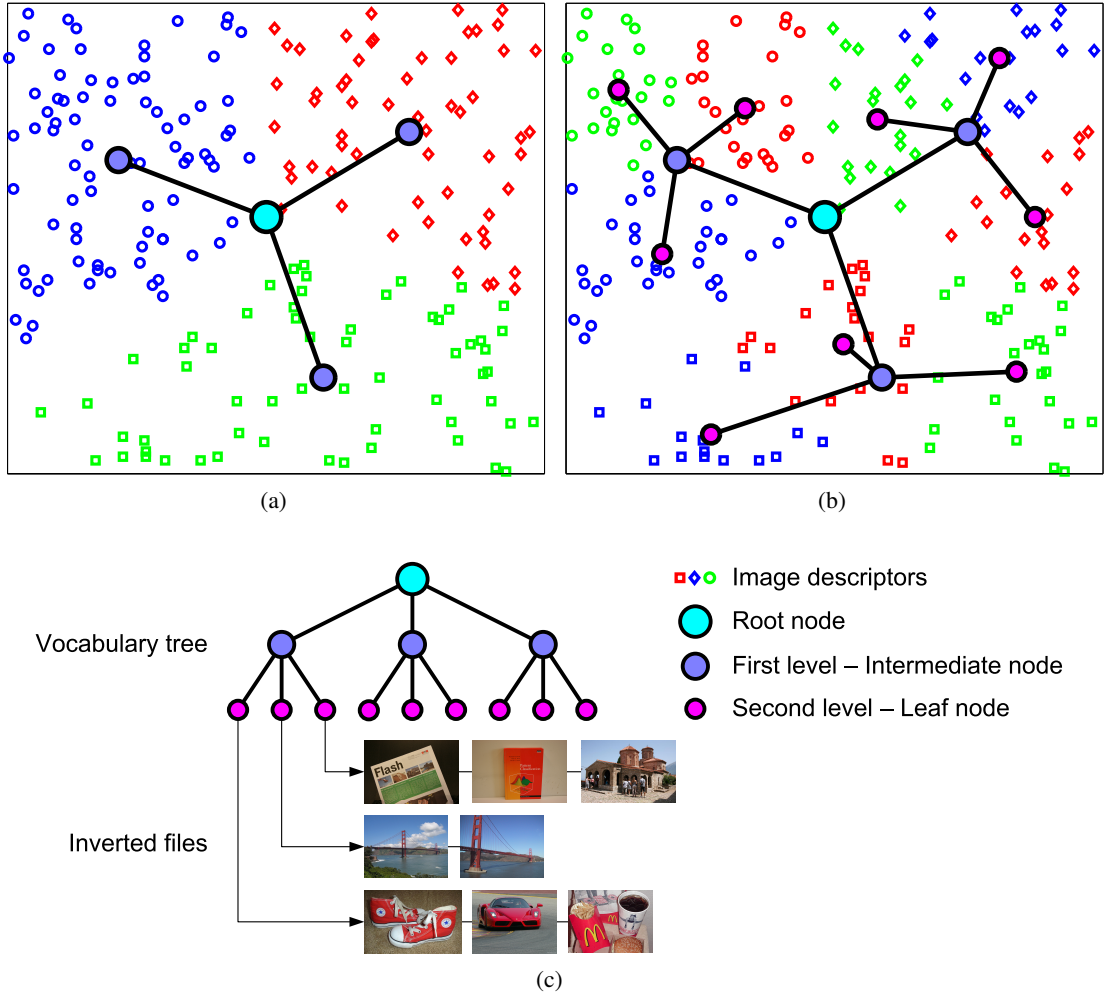


Figure 2.2: An illustration of the process of building a vocabulary tree with branch factor three and only two levels. Image feature descriptors are quantized by hierarchical k-means clustering. The hierarchical quantization is defined at each level by three cluster centers and their Voronoi regions. (a) In the first level of hierarchy, the descriptor is assigned to the closest of the three blue centers. (b) In the second level, the descriptor is assigned to the closest of the three violet descendants to the blue center. (c) With each leaf node in the vocabulary tree there is an associated inverted file with references to the images containing an instance of that node.

The computational complexity of finding a nearest neighbor for a given feature in this tree structure is significantly less than an exhaustive nearest neighbor search. However, it is important to mention that this approximation may occasionally cause erroneous matches, which are discarded by the further validation process. The computational complexity of the complete offline phase is $O((n \cdot N \cdot \log(n))^2)$, where n is the size of the image and N is the number of images, since the clustering, which is the most time consuming part of this method, uses limited number of iterations.

2.3.2 Online Part

The goal of the online processing is to automatically propagate a tag (or tags) entered for an object in a given image by a user to other images containing the same object. The user marks a desired object in the query image by selecting a bounding box around it and enters tag (or tags) for the object. Search for similar objects is done through two-level detection approach: image matching and geometric validation by object duplicate detection. The system performs image matching by making use of local features and selects a reduced set of candidate images which are most likely to contain the target object. The object duplicate detection is applied to detect and to localize the target object within the reduced set of images. All images containing matched objects are shown to the user who can then select appropriate among them to which he/she wants to propagate tag (or tags). Once an object has been tagged and target objects are detected, the user can ask the system to automatically propagate the initial tag (or tags) to other images within the database containing target objects.

2.3.2.1 Object Selection

The user can annotate any photo in the database, which is either uploaded by himself/herself or by any other user. Images are annotated on the object level. The database used in this work covers a wide range of different classes, which will be described in more details in Section 2.4.1. Once the user chooses a photo which he/she wants to annotate, the user is free to label as many objects depicted in the image as he/she chooses. The user interface used in this work is shown in Figure 2.3. By clicking on the button “Add a note”, the user marks an object by selecting object’s boundaries as a bounding box. This process is commonly used in many photo sharing services, such as Facebook or Flickr. When a user enters the page with particular image from the database, tags which are previously entered by other users accompanied with the corresponding bounding boxes, will already appear on the image. If there is a mistake in the annotation (either the outline or the text of the label is not correct), the user may either edit the label by renaming the object or redrawing along the object’s boundary. Once the desired object is marked, the tag propagation process can start.

2.3.2.2 Image Matching

In order to speed up the object duplicate detection process, the image matching is used to select a reduced set of candidate images which are most likely to contain the target object. Since the more complex object duplicate detection is only applied to this reduced set, the overall speed is considerably increased. By making use of the local features, target images can be distinguished from non target images even if the target object is just a small part of it.

Given the local features within the selected region in the query image and the vocabulary tree, a weighting vector \mathbf{q} is computed in the same way as the weighting vector \mathbf{d}_j for image j , described in Section 2.3.1.2.

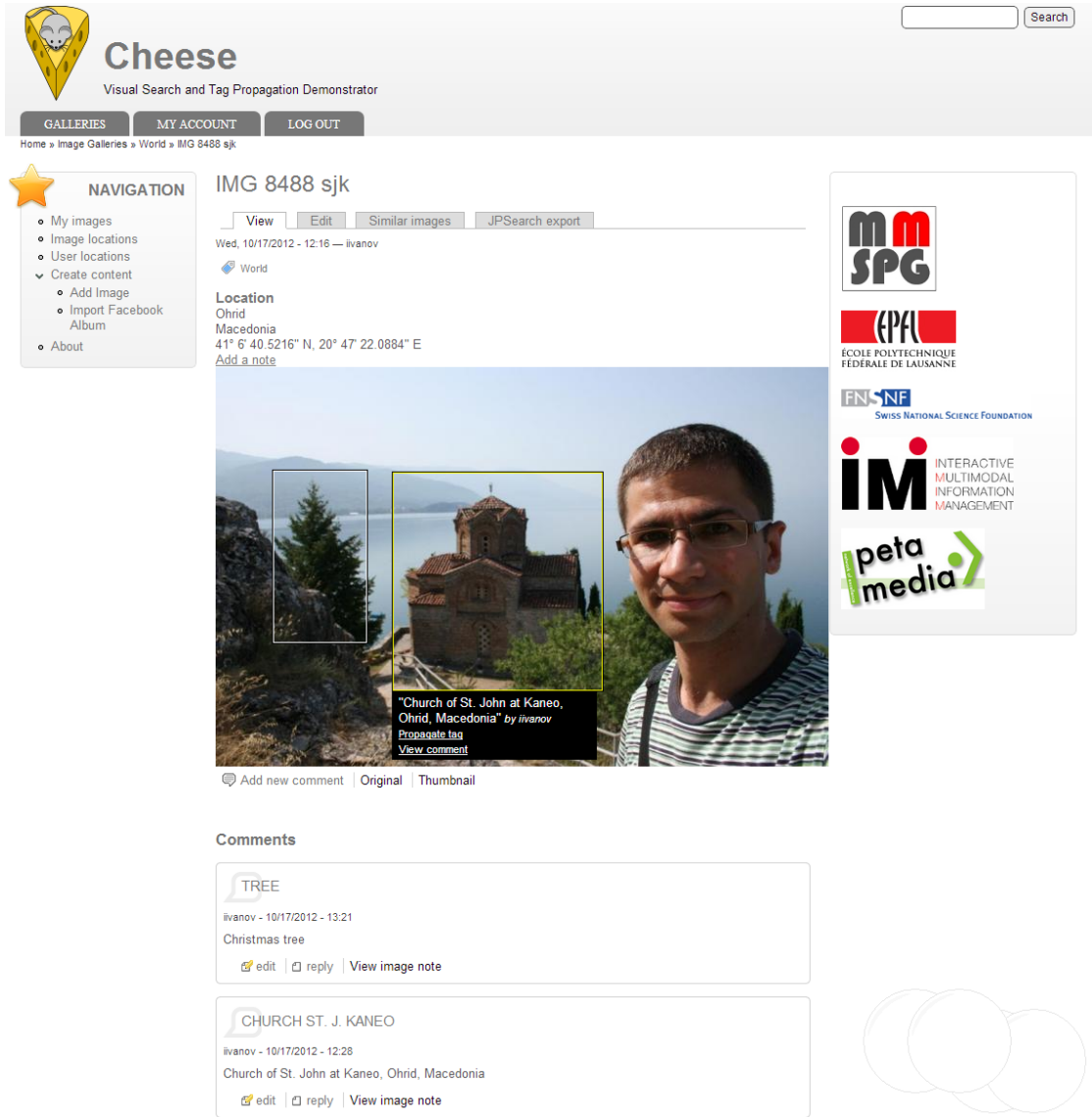


Figure 2.3: Screenshot of the object selection and tagging process in the image management platform “Cheese” (retrieved in January 2013). Two objects are selected and tagged: “tree” and “church St. J. Kaneo”. Based on the selected object, the system automatically propagates tags to other appropriate images within the database.

Based on these weighting vectors, the query image is matched to all the images j in the database and the individual matching (similarity) scores s_j are computed in the same way as in [82], using L2 norm:

$$s_j = \|\mathbf{q} - \mathbf{d}_j\| = 2 - 2 \cdot \sum_{\forall i: q_i \neq 0 \cap d_{ij} \neq 0} q_i \cdot d_{ij} . \quad (2.4)$$

In other words, when the query image is submitted for image matching, a database of N total

images can be quickly scored by traversing only the nodes visited by the descriptors from the selected region of the query image. Let s_{ij} be the matching score between the query image and image j , when the query descriptor visits node i of the vocabulary tree. Initially, prior to visiting any node, s_{ij} is set to zero. Suppose node i is visited by the query descriptors a total of N_{iq} times. Then, all images j in the inverted list under the node i will have their scores incremented according to:

$$s_{ij} = s_{ij} + q_i \cdot d_{ij} = s_{ij} + w_i^2 \cdot \frac{N_{ij}}{\sum_k N_{kj}} \cdot \frac{N_{iq}}{\sum_k N_{kq}}, \quad (2.5)$$

where N_{kq} is the number of occurrences of a visual word k within the query image q . Scores for images at the other nodes visited by the query image descriptors are updated similarly. At the end, individual scores s_{ij} are summed for all visual words (nodes) i to calculate the scores s_j .

Images j whose score s_j does not exceed a predefined threshold T_l are discarded and will not be considered for the object duplicate detection. The database images attaining the highest scores s_j are judged to be the best matching candidates and kept in a short list for further geometric verification.

The complexity of the search step for similar images is $O(n \cdot \log(n))$, where n is the size of the query image, as the feature extraction creates $O(n \cdot \log(n))$ features by making use of pyramids for detection of scale-invariant features [79].

2.3.2.3 Object Duplicate Detection

The goal of the object duplicate detection step is to detect and to localize the target object within the reduced set of images returned from the image matching step. The outcome of this step is a set of predicted objects described through their bounding boxes for each of the images.

Local features are used for object duplicate detection in [80]. Then, generalised Hough transform (GHT) is applied for object localization. Our object duplicates detection method is based on this algorithm and the detection accuracy is improved by using inverse document frequency. Inverse document frequency has been used for the similar purpose in [84]. In [84], descriptors are extracted from local affine-invariant regions and quantized into visual words, reducing the noise sensitivity of the matching. Inverted files are then used to match the video frames to a query object and retrieve those which are likely to contain the same object. Other techniques for the object duplicate detection have been proposed in the literature. For example, Vajda *et al.* [85] proposed to use sparse features which are robust to arbitrary changes in viewpoints. Spatial graph model matching is then applied to improve the accuracy of the detection, which considers the scale, orientation, position and neighborhood of the features. We used this approach for efficient automatic geotagging, which is described in Section 4.4.1. Philbin *et al.* [86] applied the bag-of-words method for detecting buildings in a large database. To resolve the problem of

large database, they use a forest of eight randomized k-d trees as a data structure for storing and searching features.

The detection and localization in our approach starts by matching the features within the selected region of the query image to the features with the candidate (test) image. Again, the hierarchical vocabulary tree is used to speed up the nearest neighbor search. Matches whose distance is larger than a predefined threshold T_F are discarded. Figure 2.4 (b) shows an example of accepted (valid) matches.

In order to detect and to localize target objects based on these matched features, the generalised Hough transform [87] is applied. The GHT is a model-based method for object localization utilizing a voting process to identify the most probable position of the object of interest. To this end, the image space is mapped to a transformation parameter space, known as Hough space, which depicts possible locations of the object. Assuming that the object in the test image has undergone some rotation θ and uniform scaling p from its position in the query image, coordinates of the features from the query image space (x', y') are mapped onto coordinates in the test image space (x'', y'') as:

$$x'' = (x' \cos(\theta) - y' \sin(\theta)) p, \quad (2.6)$$

$$y'' = (y' \sin(\theta) + x' \cos(\theta)) p. \quad (2.7)$$

Each matched feature within the candidate image votes for the position (center) and the scale of a bounding box based on the position and scale of the corresponding feature within the query image. Given the center of the object (x'_C, y'_C) in the query image space, the predicted center of the object (x''_C, y''_C) in the candidate image space is calculated as:

$$x''_C = x'' + ((x'_C - x') \cos(\theta) - (y'_C - y') \sin(\theta)) p, \quad (2.8)$$

$$y''_C = y'' + ((x'_C - x') \sin(\theta) + (y'_C - y') \cos(\theta)) p, \quad (2.9)$$

where $\theta \in [\theta_{min}, \theta_{max}]$ and $p = p''/p'$, p' and p'' are scales of the matched features in the query and the test image spaces, respectively. Equations (2.8) and (2.9) are repeated for every pair of the matched features. Since unique features may provide a more reliable estimate of the bounding box, the vote of a feature for the center of the bounding box is equal to its inverse document frequency value w_i , which is already described in Section 2.3.1.2. The vote of a feature for the scale of the bounding box is equal to p . This leads to a 3-dimensional histogram that describes the distribution of the votes across the bounding box parameters (position, and scale), as shown in Figures 2.4 (c) and (d). To obtain the set of predicted objects the local maxima of the histogram are searched and thresholded with T_O .

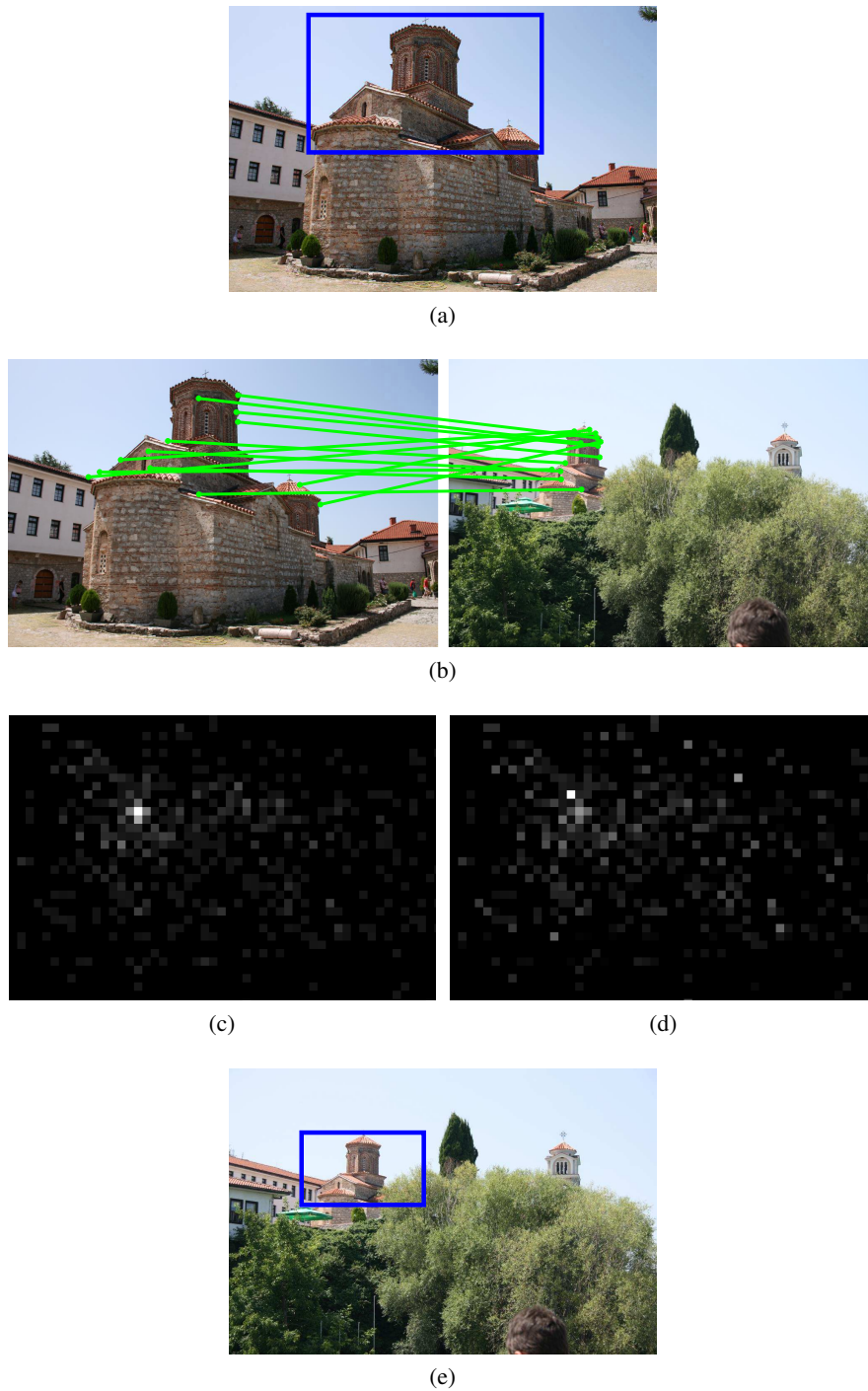


Figure 2.4: The object duplicate detection step: (a) desired object selected in a query image from the buildings object class, (b) typical matching result using SURF features between the query image and test image with view point change (green lines are valid matches; for visibility purposes, only subset of matched features is shown), (c) distribution of object center positions in the test image obtained by GHT (whiter areas have higher votes), (d) distribution of object scales in the test image obtained by GHT (whiter areas have higher votes), and (e) predicted object in the test image outlined with the bounding box.

The complexity of our method for object duplicate detection is $O(n \cdot \log(n))$, where n is the size of the query image, since the SURF feature extraction uses image pyramids for detection of scale-invariant features [79] and the generalised Hough transform has the same computational complexity, since we do not consider rotated objects within the database.

2.3.2.4 Tag propagation

Since the manual annotation of all the instances of an object within a large set of images is very time-consuming, the system offers tag propagation of marked and tagged objects as shown in Figure 2.5. Thereby duplicates of the tagged object are detected within the database and the result is shown to the user.

Once an object has been marked and tagged, a user can ask the system to propagate it automatically to the other images in the database by pressing the “Propagate tag” button. The system performs object duplicate detection in the way explained in previous sections, and returns images containing object duplicates. Considering matches between propagated and already tagged objects one has to distinguish two cases:

- If an object duplicate does not match any already tagged object both its bounding box and tag can be automatically propagated to the corresponding image. However since the object duplicate detection may return a few non-relevant objects the user can verify the propagated tags.
- If an object duplicate matches an already tagged object the two bounding boxes and sets of tags have to be merged. The system can either ask the user to resolve the conflict and merge the two objects, or this can be done automatically using some heuristics. Since manually tagged objects are usually more reliable than automatically propagated ones, the bounding box of the object duplicate will be discarded but the tags will be combined.

In the current implementation of the tag propagation approach in the “Cheese” platform, we use the former approach, namely, bounding box and associated tag (or tags) are always automatically propagated after verification done by the user. The tag propagation is performed regardless of the existence of the annotation in the matched image.

2.3.3 Implementation Details

The implementation of the modules inside the presented interactive system for tag propagation to extract features, make the database index of images and features, match and retrieve them, are done in C++. Some modules use external libraries:

Chapter 2. Object-based Tag Propagation for Semi-Automatic Annotation of Images

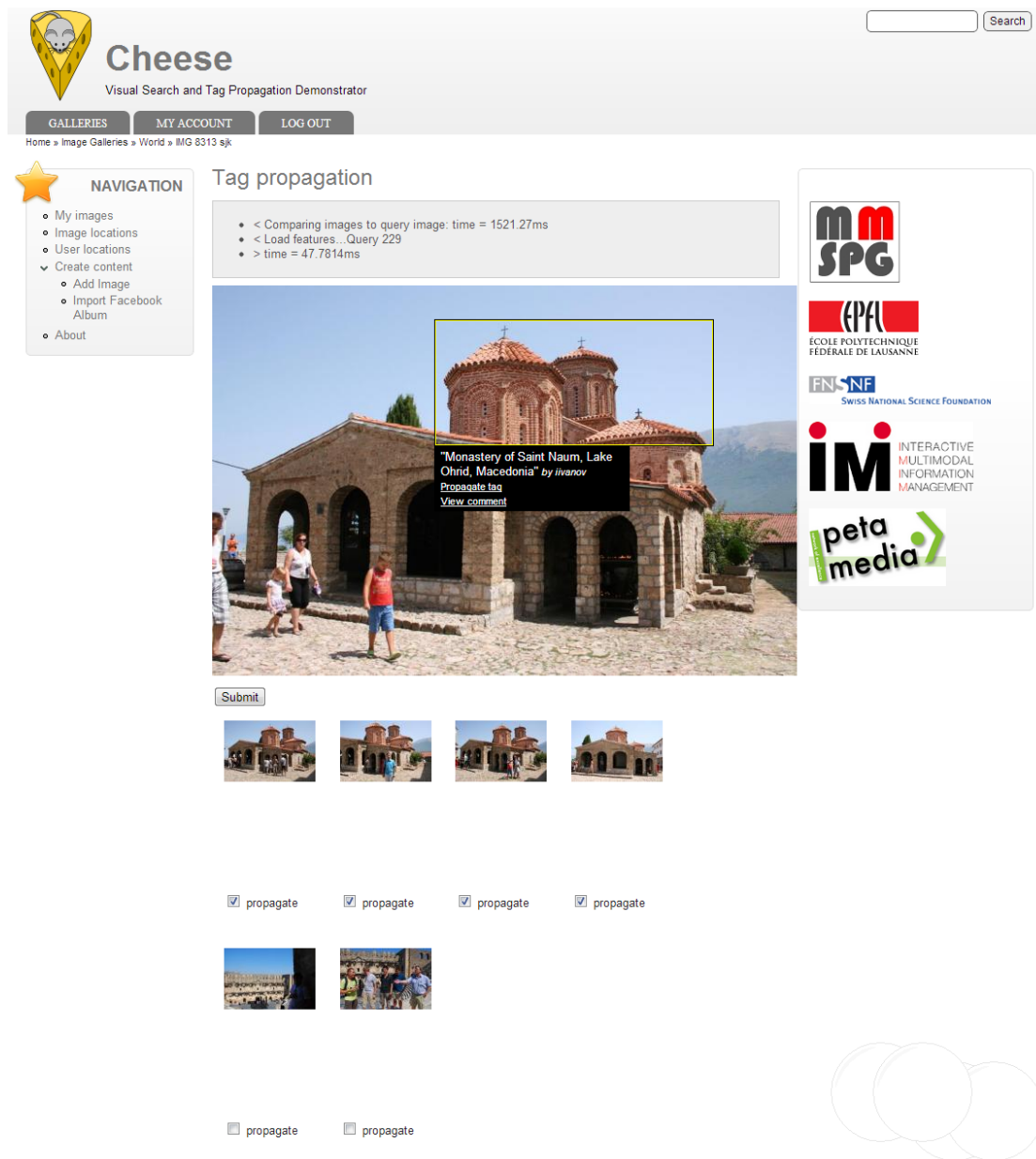


Figure 2.5: Screenshot of the tag propagation process in the image management platform “Cheese” (retrieved in January 2013). Tags of the selected object are automatically propagated to other images within the database which contain similar objects and which are confirmed by the user.

- (1) SURF detector and descriptor in OpenCV⁴³. SURF descriptor extraction has a lot of parameters which can influence the final detection performance. We used the default values for all parameters from the original work of Bay *et al.* [79]. Feature matching inside object duplicate detection uses the same library.

⁴³ <http://www.opencv.org>



Figure 2.6: Some example images from the controlled database used for the performance evaluation of the tag propagation method. More sample images of this database are provided in Appendix A.2.

- (2) LIBPMK library to create vocabulary tree (the database index of images and features), calculate TF-IDF weighting scheme, and perform image matching. This library is available online as LIBPMK – A Pyramid Match Toolkit⁴⁴. Parameters of the vocabulary tree are discussed in Section 2.4.2

Generalised Hough transform is implemented on our own. We assumed that objects in test images are scaled, and not rotated ($\theta_{min} = \theta_{max} = 0$). Therefore, only scaling in GHT is implemented. The predicted position of the object center is quantized with bin size set to 1/50 of the image size.

2.4 Experiments and Results

In this section, the performance of the proposed tag propagation method is evaluated and analyzed in two scenarios. In the first scenario, the proposed method is evaluated on the controlled database (with ground truth data) of 3200 images, while in the second scenario, more than 1 million distractor images are added and the performance of the method in a large-scale environment is evaluated. The considered databases are described in Section 2.4.1. In Section 2.4.2 the evaluation is presented, and finally the results are discussed for both scenarios in Section 2.4.3.

2.4.1 Database

We used three different databases for the performance evaluation of the proposed tag propagation approach.

The controlled database of images is created on our own and reported in [50] as a general purpose database. It consists of 3200 images: 8 classes of objects, 20 objects for each class, and 20 sample images of each object. Object classes include books, buildings, cars, gadgets, newspapers, shoes, text, and trademarks. All sample images are associated with the ground truth data. Some example photos are shown in Figure 2.6. This database is described in details in Appendix A.2.

To test scalability of the proposed method we collected two distractor databases with large number of images. We used a MIRFLICKR-1M image database containing 1 million images selected

⁴⁴ <http://people.csail.mit.edu/jjl/libpmk>



Figure 2.7: Some example images from the distractor database used for the performance evaluation of the tag propagation method. More sample images of this database are provided in Appendix A.3 and A.4.

based on their high interestingness rating according to Flickr. More details about this database are provided in Appendix A.3. We collected on our own another database of personal images from a few researchers who agreed to make publicly available images selected from their personal collections taken while traveling around the world. This database has around 16 thousand photos. More details are available in Appendix A.4. Sample images from the distractor database are shown in Figure 2.7.

In summary, the evaluation database consists of 3200 images associated with ground truth data and 1016018 distractor images. In order to make our approach more computationally feasible, all images are downsized to maximum dimension of 500×500 pixels and JPEG compressed before further processing.

2.4.2 Evaluation

The goal of the evaluation is to assess the performance of a method in order to make it comparable to other methods [88]. This section provides instructions for the particular evaluation methodology performed in our experiments.

In the first scenario, the proposed tag propagation method is evaluated only on the controlled database of 3200 images with ground truth data. The controlled database is split into training and test subsets. Training images are chosen carefully from the database so that they provide a frontal wide angle view of the objects depicted in images. Objects are selected using bounding boxes, in a way explained in Appendix A.2. One sample image from each object is chosen as a training image. All other images from the controlled database are test images. From each training and test image, the SURF features are extracted in a way described in Section 2.3.1.1. Features extracted from training images of the controlled database are used to create the vocabulary tree with $L = 5$ levels excluding the root node and $C = 10$ branches under each interior node, which leads to $K = C^L = 100000$ leaf nodes. Features extracted from the test images are matched with the features in the vocabulary tree and objects in test images are predicted.

Since tag propagation relies on the performance of the image matching and object duplicate detection parts, only these parts of the whole system are assessed. It can be evaluated as a typical detection problem where the set of predicted objects is compared against a set of ground truth

objects. Objects are matched against each other based on the overlap of their bounding boxes. If the ratio between the overlapping area and the overall area exceeds 50 %, it is considered as a match [89]. Based on that, a confusion matrix is computed, which contains the number of true positives (TP), false positives (FP) and false negatives (FN). For the evaluation precision-recall (PR) curves can be derived from this confusion matrix. PR curves plot the recall (R) versus the precision (P), defined as:

$$P = \frac{TP}{TP + FP}, \quad (2.10)$$

$$R = \frac{TP}{TP + FN}. \quad (2.11)$$

The F-measure is calculated to determine the optimum thresholds for the object duplicate detection. It can be computed as the harmonic mean of P and R values:

$$F = \frac{2 \cdot P \cdot R}{P + R}. \quad (2.12)$$

Thus, it considers precision and recall equally weighted.

In the second scenario, the proposed tag propagation approach is evaluated as a large-scale visual search method, where the goal is to search through large number of images in a limited time and evaluate retrieved results as a ranking problem. More than 1 million distractor images are added to the controlled database to form the evaluation database of images. Here we assume that objects in images from the controlled database are not present in the distractor database. We performed leave-one-out cross-validation, meaning that one image from the controlled database is used as the test data, and the remaining images from the entire evaluation database are the training data. This evaluation is repeated such that each image from the controlled database is used exactly once as the test data. Again, objects from the controlled database are selected using bounding boxes, as described in Appendix A.2. The SURF features are extracted from both training and test images. Features extracted from training images are used to create the vocabulary tree with $L = 6$ levels excluding the root node and $C = 10$ branches under each interior node, which leads to $K = 10^6$ leaf nodes. For each test image, candidate objects are predicted from training images by matching features extracted from the test image and features in the vocabulary tree.

Following the first scenario, here we also evaluate the performance of the image matching and object duplicate detection parts. As it is common in the field of large-scale visual search, recall value is calculated to predict a certain number of images, generated by image ranking method. This list is later re-ranked by object duplicate detection through geometry validation. The number of predicted images may vary from $N = 20$ till $N = 1000$, depending on the computational complexity of the geometry validation and the further re-ranking method. In this chapter, we describe three evaluation methodologies for large-scale visual search.

Chapter 2. Object-based Tag Propagation for Semi-Automatic Annotation of Images

First, recall value is calculated with Equation 2.11. This measure does not consider the order of the results.

Second, weighted performance metric called rank factor is used for evaluation [90]. It measures the impact of the query object in a ranking list. Assuming the scenario in which the rank method returns a ranked list of N objects, o_i for $i \in [1 \dots N]$, for a query q , rank factor is defined as:

$$RankFactor(N, q) = \frac{\sum_{i=1}^N \omega(o_i, q) \cdot \frac{1}{i}}{\sum_{i=1}^N \frac{1}{i}}, \quad (2.13)$$

where

$$\omega(o_i, q) = \begin{cases} 1, & \text{if } q \text{ is the same object as } o_i, \\ 0, & \text{if } q \text{ is not the same object as } o_i. \end{cases} \quad (2.14)$$

Rank factor is normalized between 0 and 1. It takes into consideration both the number of similar objects and their position in the result list. The higher the position of a similar content in the ranking list, the greater contribution of the content to rank factor is. A higher rank factor represents existence of more similar objects or higher ranks of similar objects in the result list.

Third, squared rank factor is proposed in this chapter. Rank factor is not convergent by N , therefore the results can be sensitive to the selection of N . We propose a convergent rank performance metric defined as:

$$SquaredRankFactor(N, q) = \frac{\sum_{i=1}^N \omega(o_i, q) \cdot \frac{1}{i^2}}{\sum_{i=1}^{\infty} \frac{1}{i^2}}, \quad (2.15)$$

where

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}. \quad (2.16)$$

Squared rank factor is normalized between 0 and 1. Because of its convergence, the maximum error can be calculated considering all objects in the database, but evaluating just the first N

results, as follows:

$$MaxError(N) = 1 - \frac{\sum_{i=1}^N \frac{1}{i^2}}{\sum_{i=1}^{\infty} \frac{1}{i^2}}. \quad (2.17)$$

Rank based results for visual search can be also evaluated as detection performance metric by F-measure or mean average precision (MAP), however it does not consider the whole ranking list as the listed methods from above.

2.4.3 Results and Analysis

First, we present evaluation results obtained by applying the proposed method on 3200 images from the controlled database.

Figure 2.8 shows the performance of the image matching and object duplicate detection parts in form of the average PR curve computed over all the classes within the controlled database. It provides a good visualization of the opposing effects (high precision versus high recall) which are inherent to any detection task. The results show that if both effects are considered with equal importance, the optimum is achieved at $R = 0.4$ and $P = 0.6$. However, the precision can be greatly increased if $R = 0.2$ is considered enough for the tag propagation.

In order to determine the optimal threshold T_O for the object duplicate detection, the F-measures across the different thresholds has been computed. Figure 2.9 shows the threshold versus F-measure curve. The optimal threshold of 50 is chosen for the maximum F-measure of 0.49 and shown in all related figures by green markers. The final F-measure averaged over the whole database is 0.48.

In order to compare the different classes with each other, the F-measure is computed for each of the object classes as shown in Figure 2.10. Trademarks perform best, thanks to the large number of salient regions. In the case of text or cover pages of newspapers, books or gadgets, the proposed tag propagation approach performs worse because the objects do not have enough discriminative features. Shiny or rotated objects, such as cars, shoes or buildings, are hard to detect due to the changing reflections and varying viewpoint.

Then, we extend the evaluation to include a large-scale database of more than 1 million images. The controlled database is part of this large database. Again, the system is evaluated by measuring performance of the image matching and object duplicate detection parts.

Figure 2.11 shows the performance of the image matching part in the form of the average recall computed over all the sample objects for each class within the controlled database. In the evaluation, 1000 candidate (predicted) images were selected for further geometry validation.

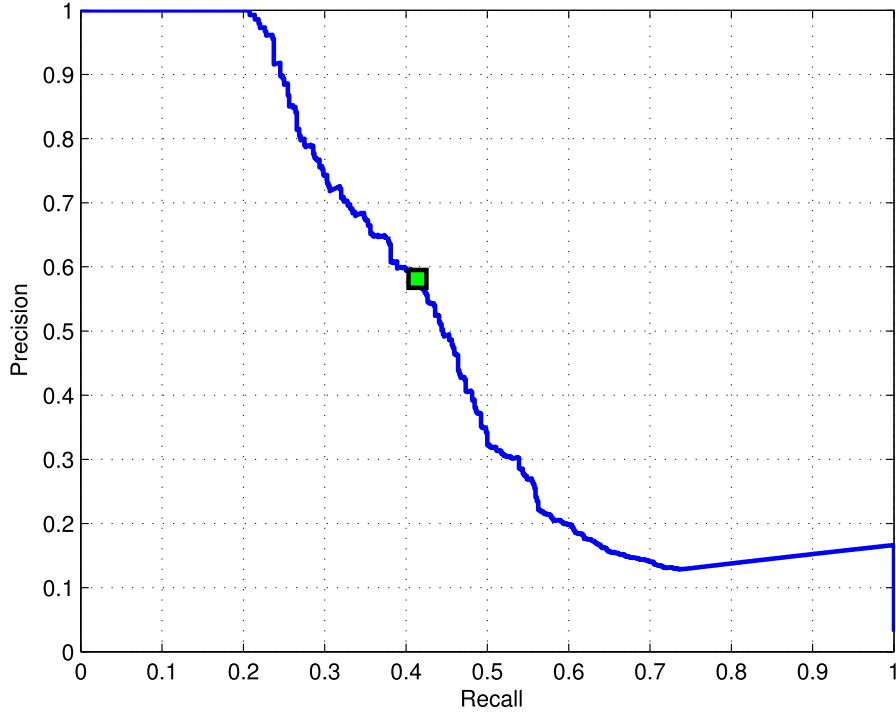


Figure 2.8: Performance of the image matching and object duplicate detection parts of the proposed system measured as precision vs. recall curve averaged over all the classes. The evaluation database has 3200 images.

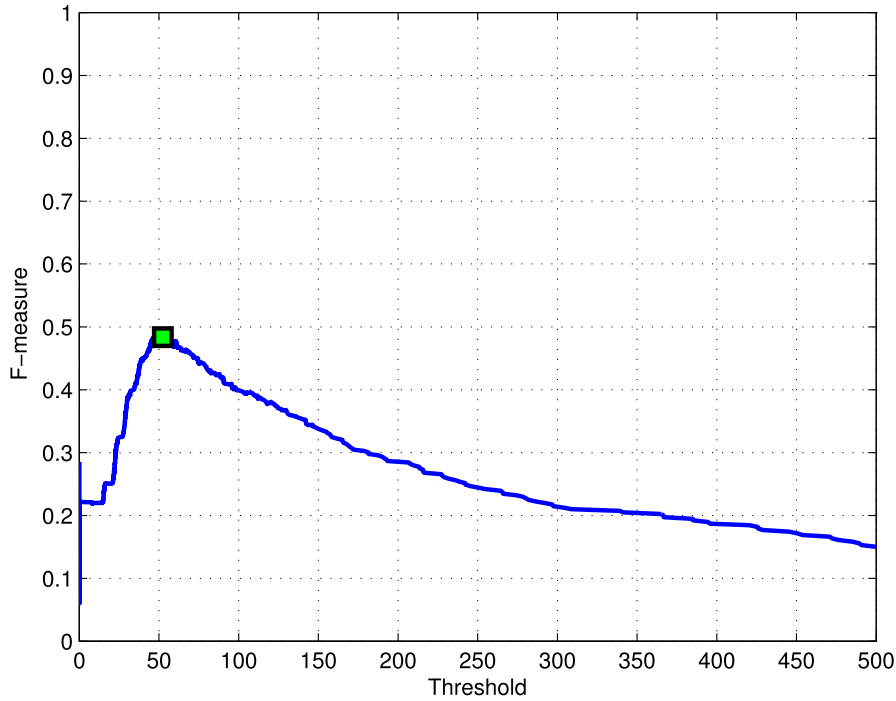


Figure 2.9: Performance of the image matching and object duplicate detection parts of the proposed system measured as average F-measure vs. object duplicate detection threshold T_O . The evaluation database has 3200 images.

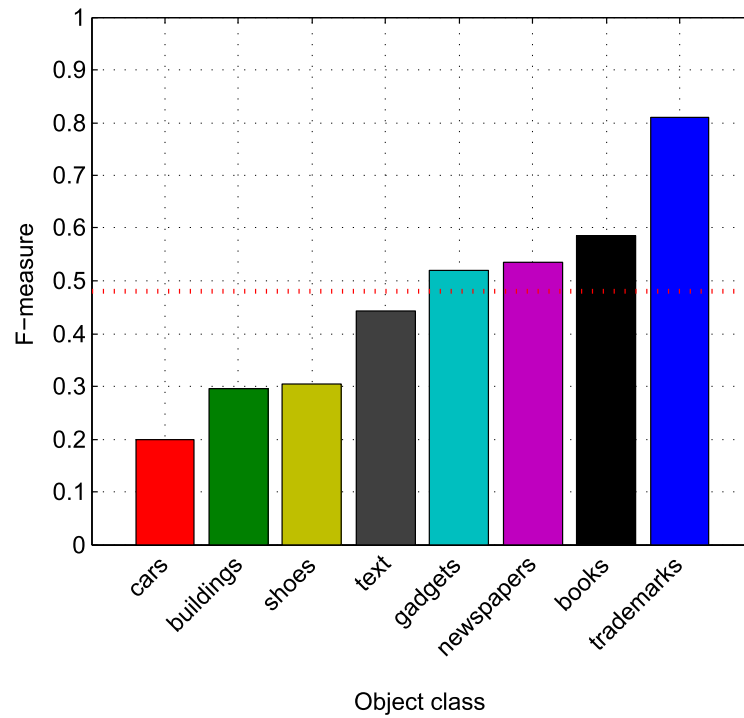


Figure 2.10: Performance of the image matching and the object duplicate detection parts of the proposed system measured as F-measure across the different classes. The F-measure averaged over the whole database is 0.48. The evaluation database has 3200 images.

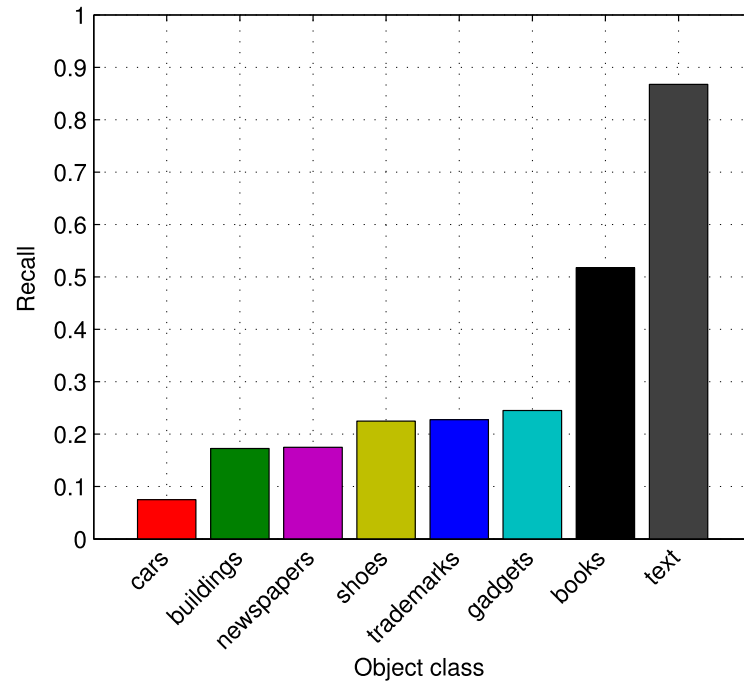


Figure 2.11: Performance of the image matching part of the proposed system measured as recall across the different classes. Recall values are shown on the $N = 1000$ predicted images in the image matching step. The evaluation database has more than 1 million images.

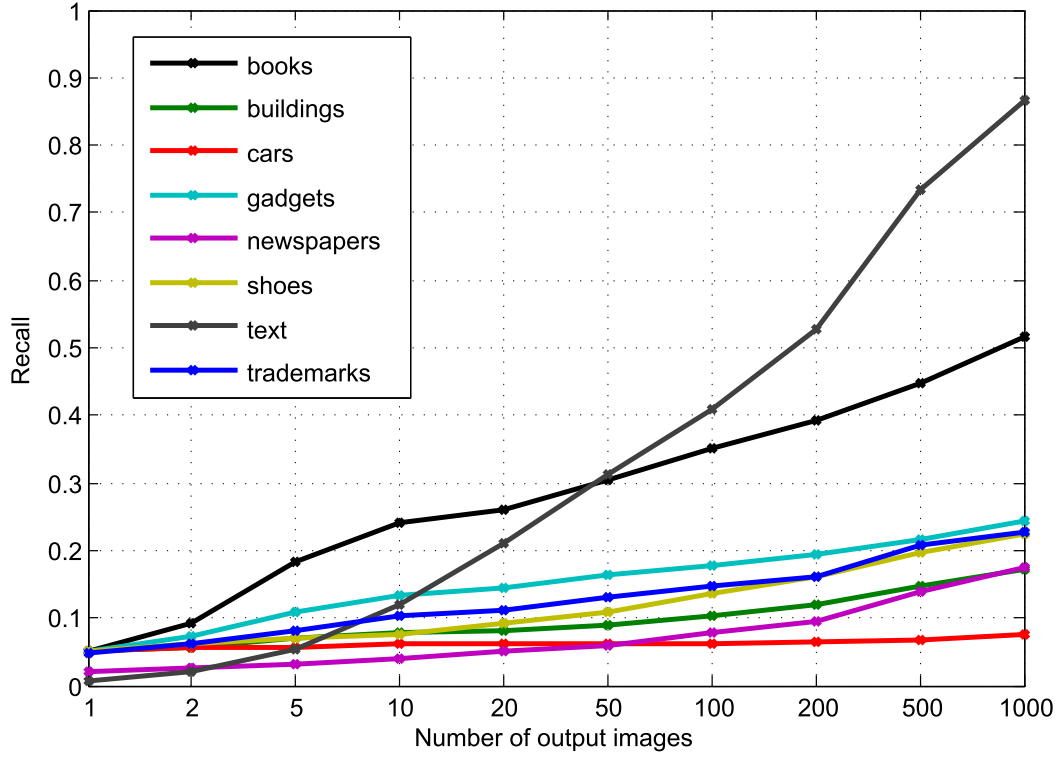


Figure 2.12: Performance of the image matching part of the proposed system measured as recall across the different classes. Recall values are shown for the various numbers of the predicted images in the image matching step. The evaluation database has more than 1 million images.

Results show that texts and books perform well. Images from cars perform the worst due to the few features and their shiny surface, since it is difficult to handle the light condition changes by local features, such as SURF.

Considering various number of predicted images, the evaluation of the image matching part is shown in Figure 2.12 for each of the object classes. Most of the visual search systems use 100 images for geometric validation. In this case maximum 40 % of the objects belonging to the text class can be detected. However, there are faster re-ranking methods with more levels of processing [91], which are able to check the geometry of 1000 images in a second. In the case of 1000 predicted images, around 90 % of recall can be reached for the text objects. The results show that for the majority of object classes, the recall value is kept very low, due to the large and challenging database. We have to mention that most of the visual search methods are evaluated on images of scenes or objects taken by one photo camera even without changing the background, light condition and the object itself. The database used in the evaluation of our method is mainly collected from various sources from the Internet, which makes the database very challenging as the results show.

Performance of the system is also measured in terms of the rank factor and squared rank factor. The results are presented in Figure 2.13 (a) and (b). These evaluation metrics take

into consideration both the number of similar objects and their position in the result list. The higher the position of a similar object in the ranking list, the greater contribution of the object to rank factor is. The evaluation results are shown for the two cases: (1) when only image matching part is considered, and (2) when image matching and object duplicate detection parts are evaluated together. General conclusion is that the system performs better when geometric verification through object duplicate detection is included in the tag propagation process. This mainly comes from the fact that, usually, there are no similar objects in the first couple of images in the ranking list. The results also show significant difference between the performance of the system with and without object duplicate detection for the text class of objects. This is reasonable and expected, due to the importance of spatial information spread over the object. Similar trends are noticed for both the rank factor and squared rank factor. However, squared rank factor is convergent by N , therefore we can calculate the maximum error of the results, which is less than 0.001 for $N = 1000$.

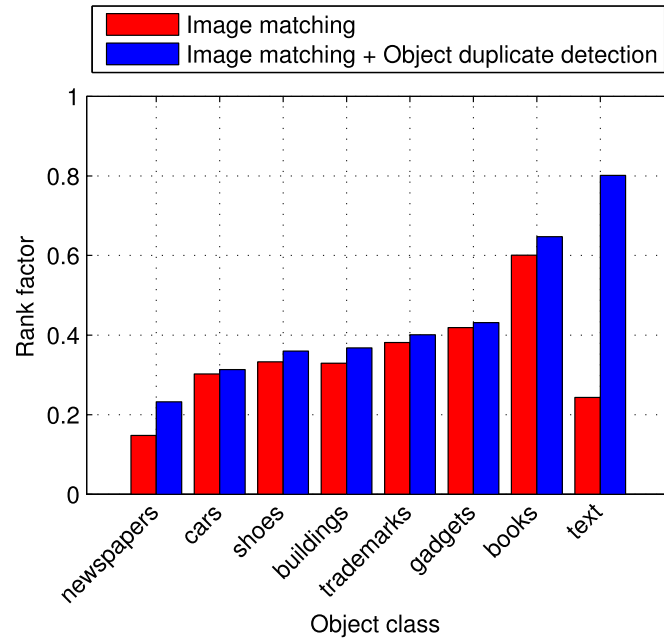
Finally, we evaluate separately the object duplicate detection step, which is based on the generalised Hough transform, as it is previously described in Section 2.3.2.3. The results are shown in Figure 2.14 as the average F-measure for each object class. Since the evaluation data in the object duplicate detection step is the output of the ranking method inside image matching part of the system, the number of objects differs per class. This unbalanced data is the reason why the object duplicate detection algorithm works well on cars, however it shows that if the ranking method finds the targeted car, the image with that particular car will be re-ranked with higher position.

2.5 Conclusion

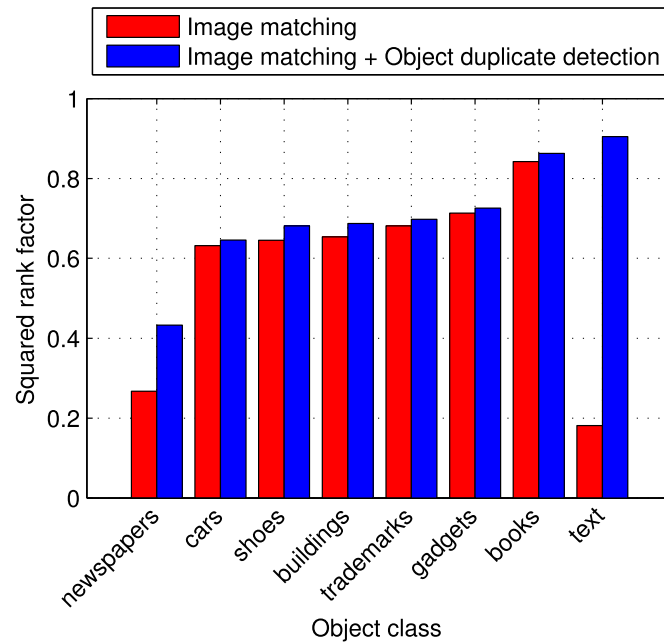
Social networks are gaining popularity for sharing interests and information. Especially photo sharing and tagging is becoming increasingly popular. Among others, tags of people, locations, and objects provide efficient information for grouping or retrieving images. Since the manual annotation of these tags is quite time-consuming, automatic tag propagation based on visual similarity offers a very interesting solution.

In this work we have developed an efficient system for semi-automatic object tagging in images. After marking desired object in an image, the system performs object duplicate detection in the whole database and returns the search results with images containing similar objects. Then, the annotation can be performed through a tag propagation process, when the user enters his/her tag for the object and it is propagated to the images in the search results.

The performance of the system has been assessed by evaluating the performance of the image matching and object duplicated detection steps, since tag propagation relies on their outcomes. First, we evaluated the proposed system on a database of 3200 images associated with ground truth, and showed that the detection works reliably for salient objects such as trademarks, books, newspapers, and gadgets. Then, we extended the evaluation to include a large-scale database of more than 1 million images. The results show that there are types of objects, where the result is



(a)



(b)

Figure 2.13: Performance of the image matching and the object duplicate detection parts of the proposed system measured as: (a) rank factor, and (b) squared rank factor. Results are shown across different object classes. The evaluation database has more than 1 million images.

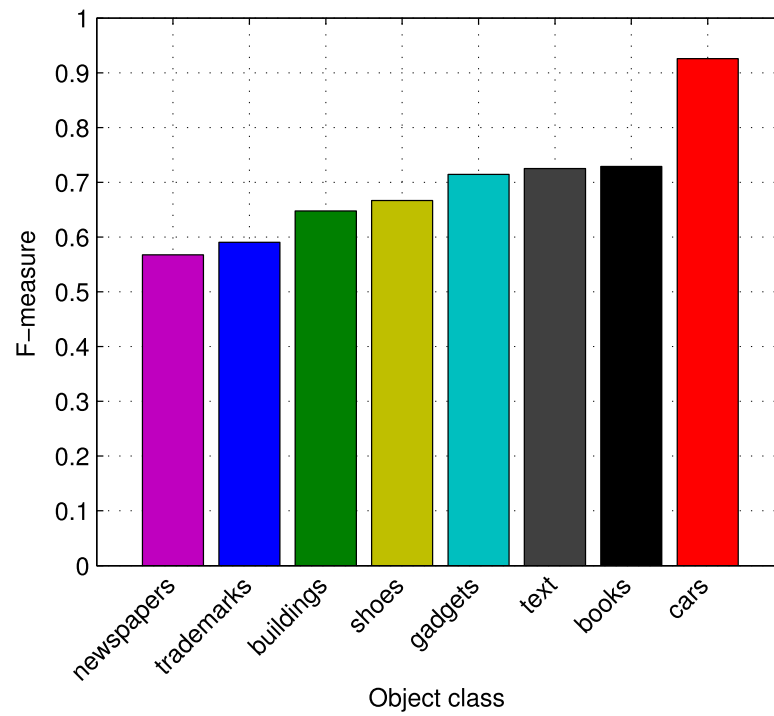


Figure 2.14: Performance of the object duplicate detection part of the proposed system measured as F-measure across different object classes. Object duplicate detection is seen as a re-ranking method for validating the object's geometry.

satisfying. Usually text based objects work better. However, for objects, which have few number of features or they are shiny such as cars, the algorithm performs worse.

Our semi-automatic tag propagation system has the potential to be improved in many ways. As a future study, one could extend it to support other classes of objects and consider evaluation of the system in the view point of the database size and latency in the system because it is important for the system to be interactive. The system can be extended to support tag recommendation. The current system relies on users' contributions in selecting the object of interest in an image which he/she wants to tag. However, it is interesting to see if any automatic approach can replace users' manual work in this matter. We deal with this challenge in Chapter 3. Since our interactive system supports interaction between users, future work can also focus on modeling users' trust in such a manner that only tags from trusted users are finally propagated. We address the challenge of modelling users' trust in social tagging systems in Chapters 4 and 5.

3 Saliency-Driven Automatic Extraction of Informative Image Content

Technique which simulates human visual attention is suited to quickly find regions of interest in images and is an interesting preprocessing method for a variety of applications. Although this technique is considered as a very simple approximation of attention, it has been found to be quite successful in computer vision, where it has been modeled by the saliency map highlighting regions which “catch the eye” in the sense of low level image properties (color, intensity, orientation). In this chapter, we study three visual attention approaches for extracting saliency in images. These methods are applied in automatic detection of informative content (objects) in images. We then perform an objective comparison of the accuracy of the saliency maps for the three state-of-the-art methods using a database of images depicting different object classes. One can assume that people spontaneously tag the most informative objects in shared images. We further explore the performance of the visual search system for object-based tag propagation which relies solely on automatic object detection.

3.1 Introduction

Human beings have the powerful ability to rapidly understand complex scenes and to recognize patterns in images or videos [92]. When an observer (e.g., a user in a photo sharing website) looks at an image, he/she routinely and effortlessly samples in detail the most relevant features of a scene, judges the importance of image parts composing the scene, and focuses attention only on important parts of the scene [93]. The important parts of the scenes in images or videos, which could be objects or regions, appear to an observer to stand out relative to its neighboring parts and are called salient regions. These salient regions can help human beings to recognize a semantic object in an image, such as a cat or a dog, as shown in Figure 3.1.

The ability of human beings to focus on specific parts of an image which carry most of the useful information needed for scene interpretation is known as visual attention. Visual attention is a neurobiological conception having the ability to concentrate mental powers upon an object for close observation. As stated in [94, 95], many physiological experiments have proved that human vision system only processes part of incoming information in full detail, whereas the rest is left nearly unprocessed. Visual attention is responsible for deciding what visual information is fully processed [95]. Visual attention is subjective, however attention from different observers are often consistent – observers tend to look at the same locations [96].

Identifying the most salient regions in images or videos is sufficient to represent the semantic meanings in most cases and consequently play an important role in many applications of image and video analysis and processing. For example, different regions in the scene in images or videos can be encoded with different quality in such a way that the most salient regions are encoded with higher quality than the other less salient regions, which leads to compression efficiency without significant impact on perceived quality [97]. The other example is object recognition where the recognition performance in highly cluttered scenes can be improved dramatically by including saliency-based bottom-up attention models [98]. Furthermore, visual attention helps reducing the need for detailed calibration for scene recognition and object localization for autonomous mobile

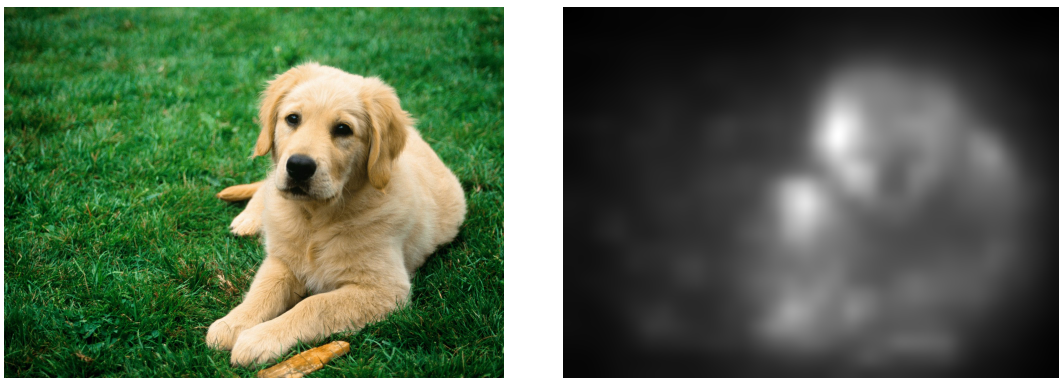


Figure 3.1: Original image of a dog and its saliency map (the whiter regions indicate the more salient parts).

robots operating in an outdoor environment [99].

Visual saliency well accords with human visual perception and can be used as one sort of selection mechanisms of the important (informative or interesting) content. Saliency-based technique is proposed recently as an alternative for object extraction [100]. The focus of this chapter is the automatic extraction of informative image content using visual saliency. We do not intend to develop new algorithms for salient region detection, but rather to compare a few existing visual attention approaches for images and examine their usefulness and performance in applications such as object detection and visual search for tag propagation. We provide an objective comparison of the accuracy of the saliency maps for three state-of-the-art methods using a ground truth of 3200 images depicting different object classes.

In this chapter, we consider two application scenarios for which visual saliency is evaluated: object detection and visual search for tag propagation. In social network systems where people share images, we assume that people spontaneously tag the most informative objects in images and therefore, tagging process can be speeded-up by making use of automatic object detection based on saliency [50]. When a user uploads an image, a system applies visual attention modeling and extracts a corresponding saliency map. Obtained saliency map is post-processed and salient region that roughly contains an interesting object is determined and outlined with a bounding box to be visible for a user in a photo sharing web site. Then, the user can adjust the borders of the extracted interesting object, provide initial tags for the object and finally apply tag propagation to automatically annotate other images containing visually similar objects in the database, for the reasons already discussed in Chapter 2.

This chapter is organized as follows. The relevant state-of-the-art approaches in salient region detection are presented in Section 3.2. Selected models for detection of salient regions based only on image content analysis are briefly described in Section 3.3 and their performance is analyzed in applications such as object detection and visual search in Sections 3.4 and 3.6. Finally, in Section 3.7 conclusions are presented.

3.2 Related Work

Visual attention is closely related to how we perceive and process visual stimuli. Therefore, it has been investigated by multiple disciplines including cognitive psychology (e.g., [101]) and neurobiology (e.g., [102]). The ability to predict, given an image (or video), where a human being might focus in a free viewing scenario, has been also of interest in the computer vision and image processing community (e.g., [92]), and that domain will be the focus of this chapter. The goal of this section is to provide an overview of the visual attention models for image processing applications, while the next section presents in more details the three selected and highly cited models for visual attention.

In the domain of computer vision, many visual attention methods have been proposed to automat-

ically extract interesting objects in images. Researchers have described two different aspects of how our brains focus on some interesting regions in images, and classified attention models into bottom-up and top-down approaches. Bottom-up attention models are mainly based on characteristics of a visual scene (stimulus-driven, or data-driven), whereas top-down models involve some cognitive phenomena like knowledge, expectations, reward, and current goals (task-specific knowledge-driven, or goal-driven) [103]. We will focus on bottom-up visual attention models as they rely solely on image content analysis, and are less computationally complex and thus faster than top-down models. According to the other classification strategy, attention models can be biologically motivated, or purely computational, or a combination of both aspects [104]. A systematic review of the major visual attention models applied to arbitrary images, with respect to previously mentioned classes, is presented by Borji and Itti in [103].

Bottom-up visual attention models utilize image content analysis to determine the contrast of image regions to their surroundings, using one or more low-level image features such as intensity, color, and orientation [92, 105]. For example, Itti *et al.* [92] determine center-surround contrast using a difference of Gaussians (DOG) approach. Frinot *et al.* [106] extend the previous approach and compute center-surround differences with square filters and use integral images to speed up the calculations. Ma and Zhang [107] generated a contrast-based saliency map and extracted objects by fuzzy growing. Achanta *et al.* [104] output a frequency-tuned saliency map and binarized it with an adaptive threshold. Hou and Zhang [108] constructed the saliency map by analyzing the log-spectrum of the image and used a simple threshold to detect objects. Harel *et al.* [109] create feature maps using Itti's method, but perform their normalization using a graph based approach. This graph-based method generates low resolution saliency maps for computational efficiency. In the following, we will focus on three selected approaches, namely Itti *et al.* [92], Achanta *et al.* [104] and Harel *et al.* [109].

3.3 Bottom-up Visual Attention Models

In this section, three highly cited bottom-up visual attention models in still images are explained in more details. We have selected models that belong to different categories based on mechanism to obtain saliency, namely biological model by Itti *et al.* [92], computational model by Achanta *et al.* [104], and combined biological and computational model by Harel *et al.* [109]. These models are summarized in Table 3.1. In the rest of this chapter, we will focus only on these models, which implementations are publicly available. All models take as input an arbitrary color image and return a corresponding saliency map as a grayscale image. Starting from obtained saliency map, salient regions that roughly contain interesting objects can be determined and outline objects to be visible for a user in a photo sharing website. Then, the user can adjust the borders of the extracted interesting objects, provide initial tags for the objects and finally apply tag propagation for automatic image annotation.

Chapter 3. Saliency-Driven Automatic Extraction of Informative Image Content

Table 3.1: Summary of three bottom-up visual attention models used to obtain saliency maps. Models are ordered according to their type.

Reference	Type	Method	Application
Itti <i>et al.</i> [92]	biological	a saliency model inspired by cognitive concepts in human vision, which combines multiscale features into a single topographical saliency map and adopts a dynamical neural network to select salient regions	salient object detection
Achanta <i>et al.</i> [104]	computational	an efficient frequency-tuned saliency model, which subtracts the mean value of the original image from the Gaussian band-pass blurred image to generate the saliency map	salient object detection and segmentation
Harel <i>et al.</i> [109]	combined (biological & computational)	a visual saliency model, which adopts ideas from graph theory to concentrate mass on certain feature channels and to highlight conspicuity	prediction of human fixation

3.3.1 A Model Motivated by Cognitive Concepts in Human Vision

Itti *et al.* [92] presented a saliency-based biologically inspired model for visual attention. Its principal idea is based on the human vision characteristic that objects with distinctive features to their surroundings are often perceived. An overview of this approach is shown in Figure 3.2.

First, an input image is subsampled into a Gaussian pyramid and each pyramid level $\sigma \in [0 \dots 8]$ is decomposed into channels for contrast-based image features such as color, intensity, and orientation. In other words, red (R), green (G), blue (B), yellow (Y), intensity (I), and local orientations (O) channels are extracted. From these channels, center-surround “feature maps” f_l for different features l are constructed as differences between “center” fine scales c and “surround” coarser scales s , and normalized. In each channel, maps are summed across scales and normalized again:

$$f_l = \mathcal{N} \left(\sum_{c=2}^4 \sum_{s=c+3}^{c+4} f_{l,c,s} \right), \quad \forall l \in L_I \cup L_C \cup L_O, \quad (3.1)$$

where $L_I = \{I\}$, $L_C = \{RG, BY\}$ and $L_O = \{O\}$. A map normalization $\mathcal{N}(\cdot)$ is performed to globally promote maps in which a small number of strong peaks of activity (conspicuous locations) is present, and to globally suppress maps which contain numerous comparable peak responses [92]. Feature maps are linearly summed and normalized once more to yield the “conspicuity

3.3. Bottom-up Visual Attention Models

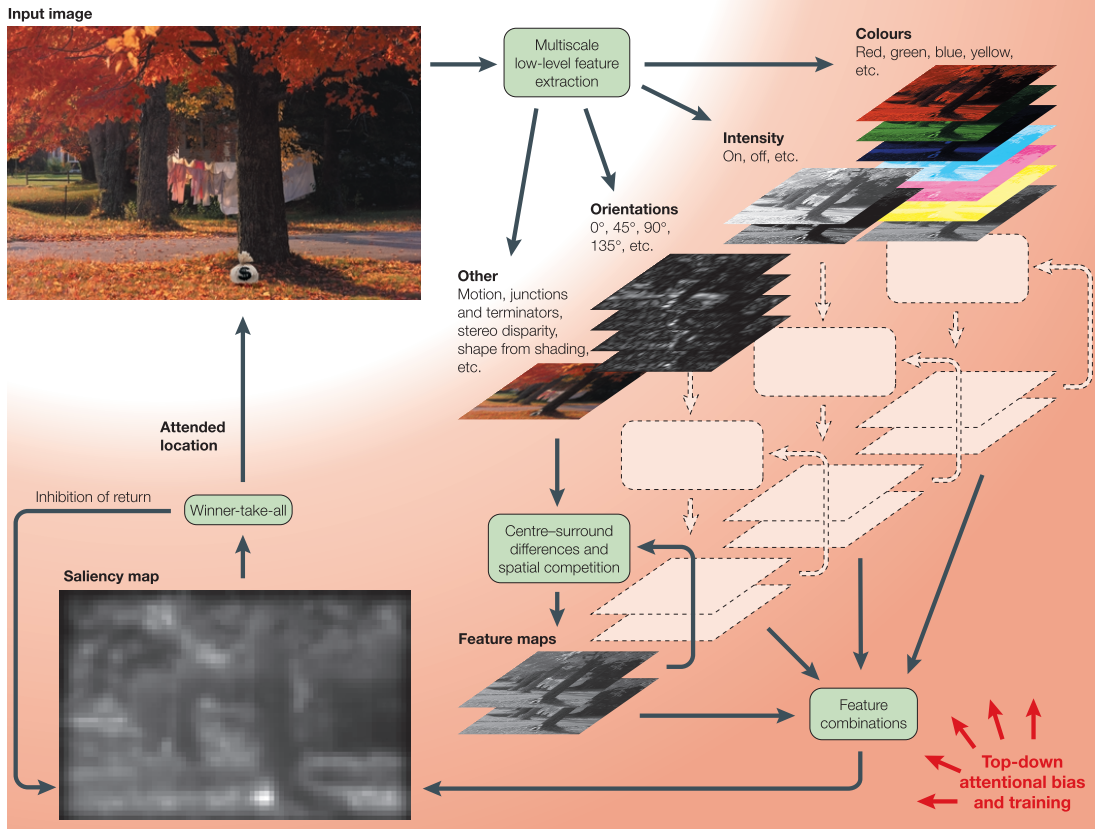


Figure 3.2: An overview of a visual attention method by Itti *et al.* (image source: [110]).

maps”:

$$C_I = f_I, \quad C_C = \mathcal{N}\left(\sum_{l \in L_C} f_l\right), \quad C_O = \mathcal{N}\left(\sum_{l \in L_O} f_l\right). \quad (3.2)$$

Finally, conspicuity maps are linearly combined once more to generate the saliency map S :

$$S = \frac{1}{3} \sum_{k \in \{I, C, O\}} C_k. \quad (3.3)$$

It was shown in [92] that the model, despite its simple architecture and low computational cost, performs well with complex natural scenes, and quickly detects salient traffic signs of varied shapes, colors and textures.

This model has been the basis of later models and the standard benchmark for comparison. As a very popular model among researchers in computer vision community, there are a few implementations of this model [103]. Results presented in this chapter are based on the implementation by

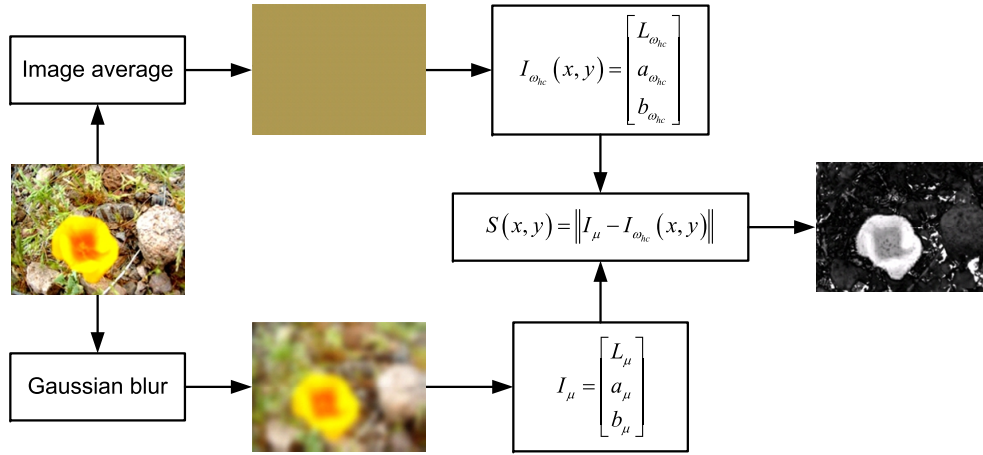


Figure 3.3: An overview of a visual attention method by Achanta *et al.* (image source: [104]).

Harel *et al.* [109], which is publicly available⁴⁵ and freely distributed.

3.3.2 A Frequency-Tuned Saliency Model

Achanta *et al.* [104] proposed a frequency-tuned approach to salient region detection using low-level features, such as color and luminance. The authors debated that the saliency map should have well-defined borders, uniformly highlighting the object if it is salient, and most of all, the saliency map should be in high resolution. Therefore, the authors proposed a method which generates the saliency map solely by center-surround contrast representation. Figure 3.3 shows an overview of this approach.

First, the input RGB image I is transformed to CIE*Lab* color space. Then, the scalar saliency map S for image I is computed as:

$$S(x,y) = \|I_{\mu} - I_{\omega_{hc}}(x,y)\|, \quad (3.4)$$

where x,y are the pixel coordinates inside the image, I_{μ} is the arithmetic mean image feature vector, $I_{\omega_{hc}}$ is a Gaussian blurred version of the original image (using a 5×5 separable binomial kernel) to eliminate fine texture details as well as noise and coding artifacts, and $\|\cdot\|$ is the L2 norm (Euclidean distance).

This approach is computationally efficient, fast and provides full resolution saliency map. When combined with adaptive thresholding and the mean-shift segmentation algorithm, the model achieves good results in salient object segmentation application [104].

We used software implementation of this approach available from the authors' web site⁴⁶.

⁴⁵ <http://www.klab.caltech.edu/~harel/share/gbvs.php>

⁴⁶ http://ivrg.epfl.ch/supplementary_material/RK_CVPR09

3.3.3 A Graph-Based Saliency Model

Harel *et al.* [109] introduced graph-based visual saliency (GBVS) model. They extract feature maps at multiple spatial scales, similar to Itti *et al.* [92], but perform maps normalization using a graph-based approach.

First, a scale-space pyramid is derived from image features: intensity, color, and orientation. Then, a fully-connected graph over all grid locations of each feature map M is built. Weights between two nodes are assigned proportional to the similarity of feature values and their spatial distance. The dissimilarity between two positions (i, j) and (p, q) in the feature map, with respective feature values $M(i, j)$ and $M(p, q)$, is defined as:

$$d((i, j) \parallel (p, q)) = \left| \log \frac{M(i, j)}{M(p, q)} \right|. \quad (3.5)$$

The directed edge from node (i, j) to node (p, q) is then assigned a weight proportional to their dissimilarity and their distance on lattice M :

$$\omega((i, j), (p, q)) = d((i, j) \parallel (p, q)) \cdot F(i - p, j - q), \quad (3.6)$$

where $F(a, b) = \exp\left(-\frac{a^2 + b^2}{2\sigma^2}\right)$ and σ is a free parameter of the method. The resulting graphs are treated as Markov chains by normalizing the weights of the outbound edges of each node to 1 and by defining an equivalence relation between nodes and states, as well as between edge weights and transition probabilities. Their equilibrium distribution is adopted as the activation and saliency maps. In the equilibrium distribution, nodes that are highly dissimilar to surrounding nodes will be assigned large values. The “activation maps” are finally normalized to emphasize conspicuous details, and then combined into a single overall saliency map.

Harel *et al.* performed evaluation of their model on real images of the natural world and discovered that GBVS model promotes higher saliency values in the center of the image plane, while “center-surround” (“c-s”) algorithms (e.g., Itti *et al.*) have troubles activating salient regions distant from object borders. The GBVS model demands very high computational cost [109].

The software implementation of this model by the author himself is available online⁴⁷ and we used it in evaluations presented in Section 3.6.

3.4 Application Scenarios

Based on visual attention modeling described in Section 3.3, we build the solution for automatic visual search. The system architecture is illustrated in Figure 3.4. First, visual attention modeling

⁴⁷ <http://www.klab.caltech.edu/~harel/share/gbvs.php>

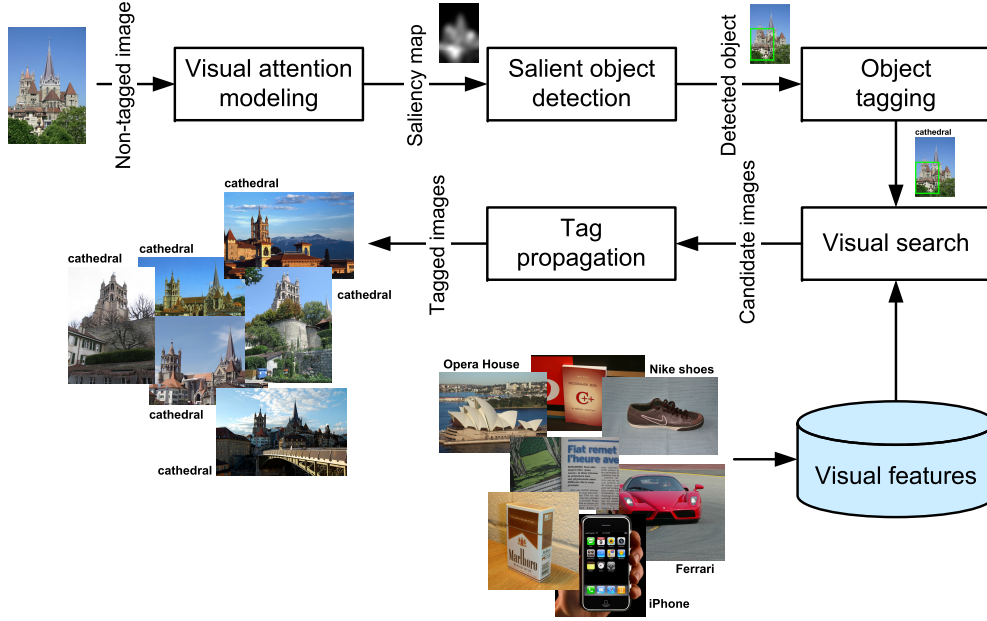


Figure 3.4: The architecture of the system which exploits visual focus of attention for automatic object detection and visual search.

is applied on an input color image and a saliency map is obtained using one of the three approaches described in Section 3.3. Then, the saliency map is thresholded in a suitable way to detect and extract a salient object, which is assumed to be the most informative part of the image. Once the salient object in the image is extracted, it is additionally tagged by a user and a search for visually similar objects is applied. Finally, the user-provided tags are propagated to retrieved images for automatic image annotation.

In this section, we describe approaches for automated salient object detection and visual search, and then their performance is measured and the results are reported and analyzed in Section 3.6.

3.4.1 Salient Object Detection

First, we consider the use of saliency maps in salient object detection. To detect a salient object or region, we need to binarize the saliency map such that ones (white pixels) correspond to salient object pixels while zeros (black pixels) correspond to the background. In this work, we assume that each input image has exactly one salient object. Therefore, if more than one salient object is detected in the input image, then the object with the largest number on pixels is considered as the most salient object in the image. Other objects are discarded.

The simplest way to obtain a binary mask for the salient object is to threshold the saliency map at a fixed threshold value.

Another method different from simple thresholding showed better performance, as shown by

Achanta *et al.* [104]. The authors performed an image-adaptive binarization of saliency maps. The threshold value T for binarization is equal to two times the mean saliency of a given image:

$$T = \frac{2}{W \times H} \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} S(x, y), \quad (3.7)$$

where W and H are the width and height of the saliency map in pixels, respectively, and $S(x, y)$ is the saliency value of the pixel at position (x, y) . Saliency map values bigger than threshold T are mapped to 255 (white pixels), while values lower than T are mapped to 0 (black pixels). White pixels correspond to the most salient object.

More sophisticated method for salient object detection is based on image segmentation. A mean-shift image segmentation is used, and the threshold for choosing salient regions is adaptive to the average saliency of the input image [104]. First, a saliency map of an input image is obtained. Then, an image segmentation is applied to decompose the original image into homogeneous regions (blobs or tiles), based on similarity in regions' pixel values [111, 112]. For each segmented region k the average saliency S_k is calculated and compared to $2 \times S_\mu$, where S_μ is the average saliency of the entire saliency map. If $S_k > 2 \times S_\mu$ then the region k is salient and all the pixels in this region are set to 255 (white pixels), otherwise pixels are set to 0 (black pixels). Again, white pixels correspond to the most salient object.

Finally, the coordinates of the salient object in the input image are determined by taking the far left, top, right, and bottom coordinates of areas with white pixels in the image. These four coordinates outline the most salient object with a bounding box (rectangle).

3.4.2 Visual Search

Once the most salient object in an input query image is extracted, a search for visually similar objects is applied, as shown in Figure 3.4. In performing visual search, content of images and objects in images are represented with numerical visual descriptors that facilitate comparison between images. Since salient objects in images are outlined by bonding-boxes in our research work, objects are treated as images of lower resolution (or cropped versions of the initial query images). At first, different low-level image characteristics, e.g., color, shape, or texture, are described using visual descriptors. Then, descriptor(s) of the query image (in our case, an image representing the salient object) is(are) compared with descriptors from a database of candidate images by measuring the similarity or distance of the two sets of descriptors. Images from the targeted database having the highest similarity or the smallest distance measure with the query image are retrieved. If the salient object is associated with tags, then these user-provided tags can be propagated to retrieved images for automatic image annotation. Since the focus of this chapter is on performance evaluation of different visual attention models in visual search applications, we do not consider here an issue of how to efficiently store descriptors extracted from images in a data structure which provides fast detection results. This issue has become important in recent

years due to the huge amount of images that people capture and share online. However, this issues is addressed in Chapter 2, where efficient system for large-scale object-based tag propagation is presented.

A number of low-level descriptors (also called features) are used for visual search among a collection of images or objects. Descriptors for visual search in images can be classified into global and local features. Global features describe an image or an object as a whole, while local features are extracted from particular parts of the object, such as key points, edges or corners. In the work presented in this chapter, we consider the following features: (1) global features: color histograms in HSV and CIE*Lab* color spaces, and texture descriptors (edge orientation histogram and homogeneous texture descriptor); (2) local features: scale-invariant feature transform, speeded up robust features and histogram of oriented gradients. Examples of three features are shown in Figure 3.5. Some other examples of features for visual search include: local binary patterns (LBP) [113], GIST descriptor [114], binary robust independent elementary features (BRIEF) [115], ferns [116], compressed histogram of gradients (CHOG) [117].

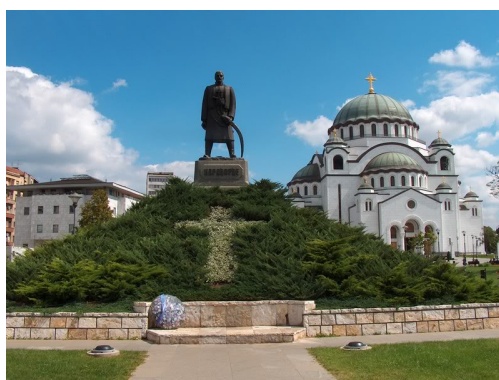
Similar images are then retrieved from the targeted database by measuring similarity between features of the query image and features extracted from candidate images by making use of similarity measurements. Similarity comparison between features in our experiments is performed using the Euclidean and the Bhattacharyya distance.

Color is a powerful clue that simplifies the recognition of objects in real-world scenes, from a human visual system (HVS) point of view. The same clue is used in the visual search. A standard descriptor for the color of an image is the color histogram [118], representing the distribution of colors in the image. Color histograms are invariant under translation and rotation, however spacial information of image content is not considered. Together with other audio-visual data descriptors at the global level, color histogram became part of the ISO standard MPEG-7. This standard defines a way of descriptor extraction, as well as description definition language, description schemes and audio-visual descriptors themselves [119, 120, 121].

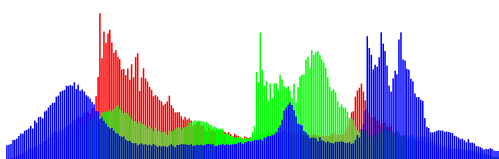
Feature extraction. Given a discrete color space defined by some color axes (e.g., RGB – red, green, blue), the color histogram is obtained by discretizing the image colors and counting the number of times each discrete color occurs in image pixels. In this work, 3-dimensional color histograms with 11 bins at each component (leading to feature size of $11 \times 11 \times 11$) are extracted for each image in HSV (hue, saturation, value) and CIE*Lab* (lightness, position between magenta and green, position between yellow and blue) color spaces. The HSV color space is developed to provide an intuitive representation of color and to approximate the way in which humans perceive and manipulate color [122]. CIE*Lab* is a perceptually uniform color space with an interesting capability for visual search that the Euclidian distance between colors in this space is strongly correlated with the human perception.

Similarity measurement. Once color histograms are extracted, they are compared to each other using the Euclidian distance.

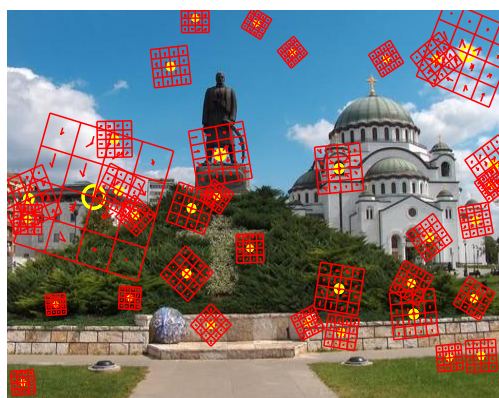
Implementation details. We used the implementation of 3-dimensional color histogram extrac-



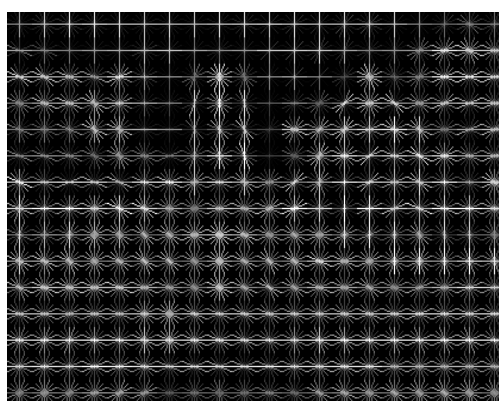
(a)



(b)



(c)



(d)

Figure 3.5: Three examples of features used for visual search: (a) the source image, (b) global color histogram, (c) local SIFT features, and (d) local HOG features. All features are extracted from the same source image shown in (a).

tion available online at File Exchange on MATLAB Central⁴⁸.

Edges are an important characteristic of an image content as they capture points in the image at which the image brightness has discontinuities. Discontinuities in image brightness are likely to correspond to discontinuities in depth or in surface orientation, and therefore it can help in segmenting objects in images [123].

Feature extraction. We extract the MPEG-7 edge orientation histogram (EOH) as it contains information about the spatial distribution of edges in an image. Each image is divided into 4×4 non-overlapping rectangular regions. In each region, a histogram based on orientation of local edges is computed, considering five types of edges: vertical, horizontal, 45° diagonal, 135° diagonal, and non-directional. Edges are detected by applying Canny edge detector [124]. The resulting edge histogram of size $4 \times 4 \times 5$ is then normalized.

Similarity measurement. In order to retrieve similar images to a query image, the extracted edge histograms are compared using the Bhattacharyya distance, which measures the approximate overlap of two populations [125].

Implementation details. We used the MATLAB implementation of MPEG-7 edge orientation histogram available online at Computer Vision Source Code⁴⁹.

Texture is another low-level image feature and it provides surface characteristics of an image. Unlike color, texture occurs over a region rather than a point and it gives us information about the spatial arrangement of color or intensities in an image, usually occurring as regular or repeated patterns in an image [126].

Feature extraction. Extraction of homogenous texture descriptor (HTD) is based on multi-channel Gabor filtering to mimic the operation of HVS for identifying different texture regions [127]. A bank of Gabor filters tuned to different spatial-frequencies and orientations is used to decompose an image into a number of filtered images. In this work, a bank of Gabor filters at 6 orientations (from 0 to $\pi - \frac{\pi}{6}$ with the step of $\frac{\pi}{6}$) and 10 different scales (from 2 to 20 with the step of 2) is used. Texture features are extracted from the filtered images as the mean and the standard deviation of the magnitude of filtered image for each scale and each orientation, and concatenated together. Therefore, final texture descriptor is of length 120 ($= 6 \cdot 10 \cdot 2$) elements.

Similarity measurement. Similar to the case of matching edge orientation histograms, texture features are compared to each other using the Bhattacharyya measure.

Implementation details. We used texture feature extraction tool available in a MATLAB toolbox implementing Computer Vision and Pattern Recognition related algorithms⁵⁰.

The *scale-invariant feature transform* (SIFT), proposed by Lowe [80], is used to detect and describe local features in images. The SIFT feature is among the most commonly used region descriptors, as it was shown to be invariant to scaling, rotation, and viewpoint change, and to be reliable and discriminable (descriptors for similar objects or image parts are also similar).

Feature detection and extraction. At first, interest key points are detected in images as local

⁴⁸ <http://www.mathworks.com/matlabcentral/fileexchange/22030-image-retrieval-query-by-example-demo>

⁴⁹ <http://clickdamage.com/sourcecode/index.php>

⁵⁰ <http://cvprtoolbox.sourceforge.net>

maxima and minima of the result of difference of Gaussians (DOG) function applied in scale-space to a series of smoothed and resampled images. Poorly localized key points along edges and low contrast key points are discarded. Then, gradient magnitude and orientation for each pixel are calculated from discrete approximation of the scale-space image, and dominant orientation from the neighboring pixels of the key point is assigned to each localized key point. An orientation histogram is formed from the gradient orientations of the pixels around the key point. Each sample added to the histogram is weighted by its gradient magnitude and by a Gaussian-weighted circular window within the neighborhood of the key point. This additional weight ensures more stable matching and recognition. Finally, the SIFT descriptor is formed from a 2D spatial location and orientation of a key point, and a vector containing the values of all the orientation histogram entries from 4×4 arrays of histogram with 8 bins of orientation around each key point, thus forming a descriptor of 128 ($= 4 \cdot 4 \cdot 8$) elements.

Similarity measurement. As the number of SIFT features extracted from a single image can be very high, one of the biggest challenges in visual search based on SIFT features is how to efficiently match large number of features. The best candidate match for each feature extracted from a query image is found by identifying its nearest neighbor in the database of features extracted from targeted images. We used the approach by Beis and Lowe [128], who proposed a modification of the k-d tree algorithm called the best-bin-first (BBF) search method, that can identify the nearest neighbors with high probability using only a limited amount of computation. The BBF algorithm uses a modified search ordering for the k-d tree algorithm so that bins in feature space are searched in the order of their closest distance from the query location. The nearest neighbors are defined as the features with minimum Euclidean distance from the given descriptor vector. The probability that a match is correct can be determined by taking the ratio of distance from the closest neighbor to the distance of the second closest. Lowe [80] rejected all matches in which the distance ratio is greater than 0.8. Once the best matches for each feature in a query image are determined, the similarity measure between images is computed as the ratio of the number of matched points and the minimum of the total key points in each image (query and each of the targeted images), as proposed by Mikolajczyk *et al.* [129]. The images with the highest similarity measure are retrieved from the targeted database of images.

Implementation details. We used the MATLAB/C implementation of the SIFT detector and descriptor by A. Vedaldi and B. Fulkerson, which is available online as VLFeat⁵¹ – An Open and Portable Library of Computer Vision Algorithms. Computation of a SIFT descriptor has many parameters that influence feature extraction process. We used the default values for all parameters from the original work of Lowe [80].

The *speeded up robust feature* (SURF) is a robust local feature detector, introduced by Bay *et al.* [79], which can be used for object recognition or 3D reconstruction. Similar to SIFT feature, it is a scale and rotation-invariant interest key point detector and descriptor, but computationally faster than SIFT.

Feature detection and extraction. The detection of interest key points is performed using an approximation of determinant of Hessian by 2D Haar-like features. Haar-like features are

⁵¹ <http://www.vlfeat.org>

approximated second order partial derivatives of Gaussian, which are used at several scales for the detection of key points (blobs) of various sizes. In addition to the approximation, the different scale-space images are produced by upsampling the filter (Haar-like features) and using integral images, instead of reducing the size of the original image as in SIFT detector. The approximations and use of integral images are the reasons why this algorithm is much faster than its predecessors, e.g., SIFT detector. Integral images drastically speed up the computation of the sum of elements inside a rectangle. The localization of the interest key points is performed by detection of maxima of determinant of the Hessian, across scale and space. Then, these interest key points are described based on Haar wavelet responses in horizontal and vertical directions in a circular neighborhood of the key point. After determining the dominant orientation of those responses in order to be invariant to image rotation, the key points is split up into smaller 4×4 square sub-regions. For each of those squares, Haar wavelet responses are computed again and summed up to generate sub-region descriptor. In order to bring in information about the polarity of the intensity changes, the sum of the absolute values of the responses is also extracted, therefore each sub-region has a four-dimensional descriptor. Again, this process is speeded up by the use of integral images. Finally, sub-region descriptors are concatenated in one region descriptor of size 64 ($= 16 \cdot 4$).

Similarity measurement. To match a query image descriptors to descriptors from the database of targeted images, the same approach is applied as for the SIFT similarity measure. Images with the highest similarity measure are retrieved.

Implementation details. We used SURFmex⁵², a MATLAB interface to C++ implementation of SURF detector in OpenCV⁵³. SURF descriptor extraction has also a lot of parameters which can influence the final detection performance. We used the default values for all parameters from the original work of Bay *et al.* [79].

The *histogram of oriented gradients* (HOG) is feature descriptor used in visual search for the purpose of object detection, which were shown to be especially efficient in human detection [130]. This method is based on evaluating normalized local histograms of image gradient orientations in a dense grid. The basic idea of the method is that local object appearance and shape can often be characterized well by the distribution of local intensity gradients, even without precise knowledge of the corresponding gradients and their positions.

Feature detection and extraction. The HOG descriptor counts occurrences of gradient orientation in localized regions (a dense grid of uniformly spaced cells) of an image. For each cell, both the horizontal and the vertical gradient magnitude values are calculated by applying the following filter kernels on the gray level version of the image: $[-1, 0, 1]$ and $[-1, 0, 1]^T$. In addition, gradient angle/orientation values are calculated for each pixel location using the horizontally and vertically filtered images. Then, the corresponding cell histograms based on gradient orientation are created. Every pixel within a cell contributes a weighted vote, according to the corresponding gradient magnitude, for the gradient orientation histogram of one cell. The orientation bins in histograms are evenly spread over 0° to 180° (“unsigned” gradient) or 0° to

⁵² <http://www.maths.lth.se/matematiklth/personal/petter/surfmex.php>

⁵³ <http://www.opencv.org>

360° (“signed” gradient). The cells can be either rectangular or radial. Cells are grouped into a block. Blocks overlap in a few cells in each direction. Gradient histograms are normalized locally at block level using L2 norm, in order to account for changes in illumination and contrast, and to improve detection accuracy. Concatenation of histograms over blocks creates the final HOG feature of an image.

Similarity measurement. Histograms of oriented gradients are compared to each other using the Bhattacharyya measure, similar to EOH and HTD descriptors.

Implementation details. We implemented extraction of this feature on our own in MATLAB. We assumed that gradient is “signed” and the cells are rectangular. Every image is divided into 25 cells, with 5 rectangular non-overlapping cells along each direction (horizontal and vertical). One block is comprised of 4 (2×2) cells. Blocks overlap one cell in each direction. For each cell, 20-bin orientation histogram of weighted gradients is created. As a result, the overall HOG descriptor for each of the images has length of 1280 ($= 20 \cdot 4 \cdot 4 \cdot 4$) elements.

3.5 Experiments

In this section, the evaluation methodology is presented. The experiments are divided into two application scenarios, where the first one considers salient object detection and the second one visual search. The considered database is briefly described in Section 3.5.1. In Section 3.5.2 the evaluation methodology and measures are presented.

3.5.1 Dataset

An evaluation database was created in order to evaluate the presented system for an automatic visual search. The database consists of 3200 images of various object classes, such as books, buildings, cars, gadgets, newspapers, shoes, text, and trademarks. Some sample images are shown in Figure 3.6. More details on this database are provided in Appendix A.2.

Each image contains targeted object which defines the class of that image. Objects in images are selected using bounding boxes which are used to evaluate visual attention models presented in Section 3.3. Ground truth bounding boxes are compared with predicted bounding boxes produced by visual attention models and results are analyzed in Section 3.6.

For the processing, all images are downsized to maximum dimension of 500×500 pixels and JPEG compressed. Reducing the image dimensions makes all visual attention approaches more computationally feasible, especially when these models are combined with mean-shift segmentation methods.



Figure 3.6: Sample images from the 160 objects within the database of images used for evaluation of visual attention models. More sample images of this database are provided in Appendix A.2.

3.5.2 Evaluation

This section provides instructions for the particular evaluation methodologies performed in our case to measure performance of visual attention and visual search algorithms.

First, the experiment on visual attention models is performed to evaluate how good these models are in salient object prediction. Experiment is repeated three times for each of the visual attention models described in Section 3.3. A saliency map is computed for each image in the evaluation database. The saliency map is then binarized by making use of adaptive thresholding and mean-shift segmentation with adaptive thresholding, as previously described in Section 3.4.1. The position of the most salient object in each image is predicted using a bounding box and compared with the bounding box of the ground truth object. Based on the overlap between the predicted bounding box B_p and ground truth bounding box B_{gt} , two evaluation measures are computed:

$$acceptance = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})}, \quad (3.8)$$

$$accuracy = \frac{area(B_p \cap B_{gt})}{area(B_{gt})}, \quad (3.9)$$

where $area(B_x)$, $x \in \{p, gt\}$, represents the number of pixels inside the bounding box B_x . To be considered as a valid detection, *acceptance* measure is required to exceed 50 %. This criterion is introduced in the evaluation of the detection task at the PASCAL Visual Object Classes Challenge [89]. The threshold of 50 % is set deliberately low to account for inaccuracies in bounding boxes in the ground truth data, for example, defining the bounding box for a highly non-convex object, e.g., a side view of a bridge or a car with an extended radio aerial, is somewhat subjective [89]. In the evaluation of the segmentation algorithms, these measurements, as well as similar measures, such as precision and recall, are usually based on the number of pixels in the predicted and ground truth object, and not on the number of pixels inside the bounding boxes around predicted and ground truth objects. The way we define measurements is suitable for our application scenario where a user associate a tag with an object in image, and usually annotation is done through outlining bounding box around the object. To distinguish between two object predictions which are accepted, but one of them is more accurate than the other one, we introduce *accuracy* measure. An example of visual comparison between *acceptance* and *accuracy* measure

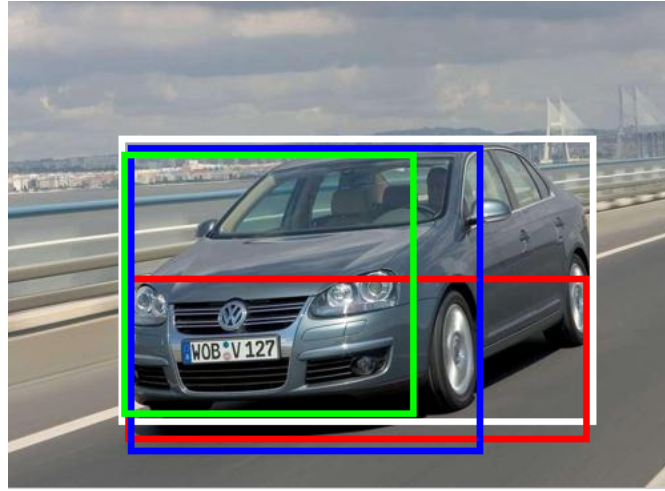


Figure 3.7: An example of visual comparison between *acceptance* and *accuracy* measures. White bounding box represents the ground truth, while red, green and blue are predicted bounding boxes by different visual attention models. Blue prediction has higher *acceptance* than red prediction, since blue bounding box has greater intersection with white bounding box while the union with white bounding box is the same for blue and red predictions. On the other hand, blue and green predictions have approximately the same *acceptance* measure, however blue prediction has higher *accuracy* as it has greater overlap with the white bounding box.

is shown in Figure 3.7. Blue prediction has higher *acceptance* than red prediction, since blue bounding box has greater intersection with white bounding box (ground truth) while the union with white bounding box is the same for blue and red predictions. On the other hand, blue and green predictions have approximately the same *acceptance* measure, however blue prediction has higher *accuracy* as it has greater overlap with the white bounding box.

Second, we continue the evaluation by adding visual search engine on top of the saliency detection. We selected the frequency-tuned saliency model by Achanta *et al.* [104], as a fast and reliable visual attention model, which will be discussed in Section 3.6.1. Furthermore, we considered seven types of visual features, which are explained in Section 3.4.2. When a query image is submitted, candidate images containing the same object are retrieved by making use of pairwise comparison between visual descriptors. To perform thorough analysis of visual search algorithms, we considered three types of query images: the entire image, the image of the ground truth object and the image of the object detected by a visual attention method. We performed leave-one-out cross-validation, meaning that one image from the database is used as the test data, and the remaining 3199 images as the training data. This is repeated such that each image from the database is used exactly once as the test data. For each query image, there are 19 positive sample images and the rest are negative samples. We used one of the standard evaluation measures, namely the recall at particular ranks (*recall@R*). It is a very good measure of the filtering capability of a visual search system and it measures the ratio of relevant images ranked in top *R*

positions. This measure is defined as:

$$recall@R = \frac{|retrieved@R \cap relevant|}{|relevant|}, \quad (3.10)$$

where *retrieved@R* denotes the set of *R* top-ranked images and *relevant* is the set of all images that are relevant to the query (in our case, 19 images). We calculated *recall@R* for $R \in \{5, 10, 20\}$.

3.6 Results and Analysis



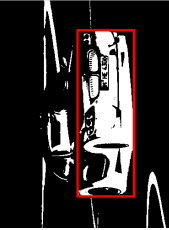

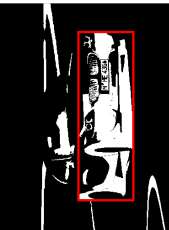

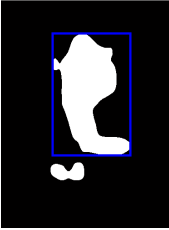
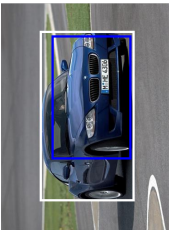


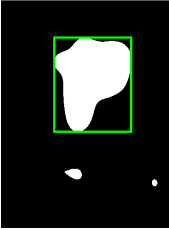
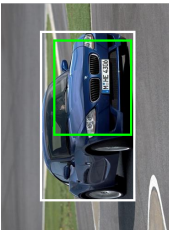
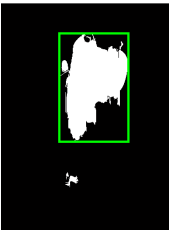
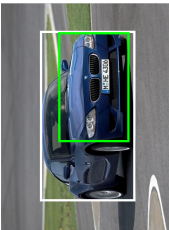

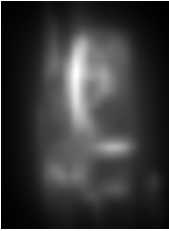
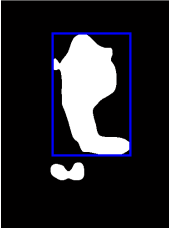
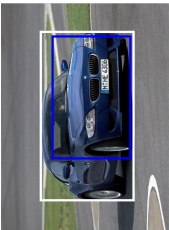


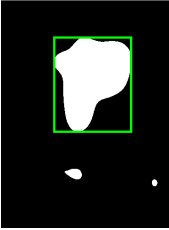
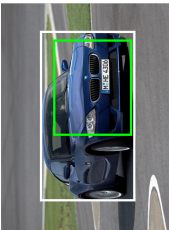
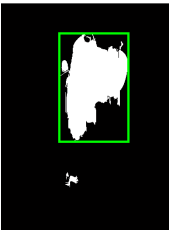
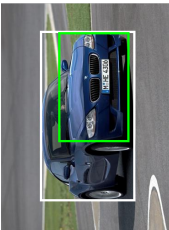






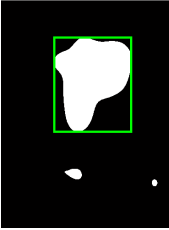
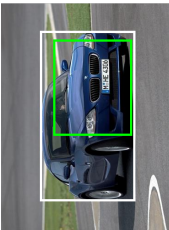
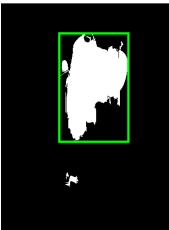
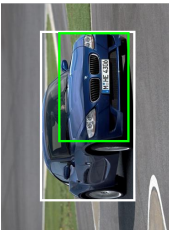








In this section, the performance of the both salient object detection and visual search application scenarios are presented and discussed.

3.6.1 Results of Salient Object Detection

The salient object detection module, shown in Figure 3.4, is powered by alternating the three visual attention models described in Section 3.3, and then the performance of the visual attention models is measured. Table 3.2 provides some examples of the obtained saliency maps and predicted salient objects for two methods: adaptive thresholding and mean-shift segmentation with adaptive thresholding, which are explained in Section 3.4.1. Positions of the predicted objects are then compared with the positions of the ground truth objects, and *acceptance* and *accuracy* measures are calculated.

Figure 3.8 shows the distribution of the images with respect to the *acceptance* value obtained by applying considered visual attention models on the evaluation database of images. Threshold of the *acceptance* value is set from 50 % to 95 % (with a step of 5 %), and the number of images that have *acceptance* values above this threshold is counted. Counted values are shown in Figure 3.8. General trend is that number of images decreases with the increase of the *acceptance* threshold. For salient object detection by adaptive thresholding (see Figure 3.8 (a)), the approach by Achanta *et al.* (red bars) performs better than other two approaches for *acceptance* thresholds greater than 60 %, while the approach by Harel *et al.* (blue bars) performs good around 50 % of *acceptance* threshold. With the increase of the *acceptance* threshold, performance of approaches by Harel *et al.* and Itti *et al.* (green bars) decrease faster than Achanta *et al.*. One of the reasons for this is that both approaches by Harel *et al.* and Itti *et al.* produce low resolution saliency maps, namely saliency map is downscaled by factor of 64 and 256, respectively, from the resolution of the source image. Therefore, the range of spatial frequencies in the source images is reduced and borders between salient objects and background regions are usually not accurately preserved. The approach by Achanta *et al.* produces the saliency map of the same resolution as the source image and reduces the spatial frequencies by lower factor [104]. For salient object detection by mean-shift segmentation with adaptive thresholding (see Figure 3.8 (b)), the approach by Harel *et al.* outperforms the two other approaches for *acceptance* thresholds below 70 %. However, above this threshold the approach by Harel *et al.* has fast decrease in performance and lacks

Table 3.2: Results of salient object detection scenario by making use of considered visual attention models. Obtained saliency maps are shown for each of the models in separated rows. In addition, saliency maps are binarized with two methods: adaptive thresholding and mean-shift segmentation with adaptive thresholding. Predicted salient objects are outlined in different colors (red, blue and green bounding-boxes) for each of the visual attention models, and are visually compared with ground truth (white bounding box).

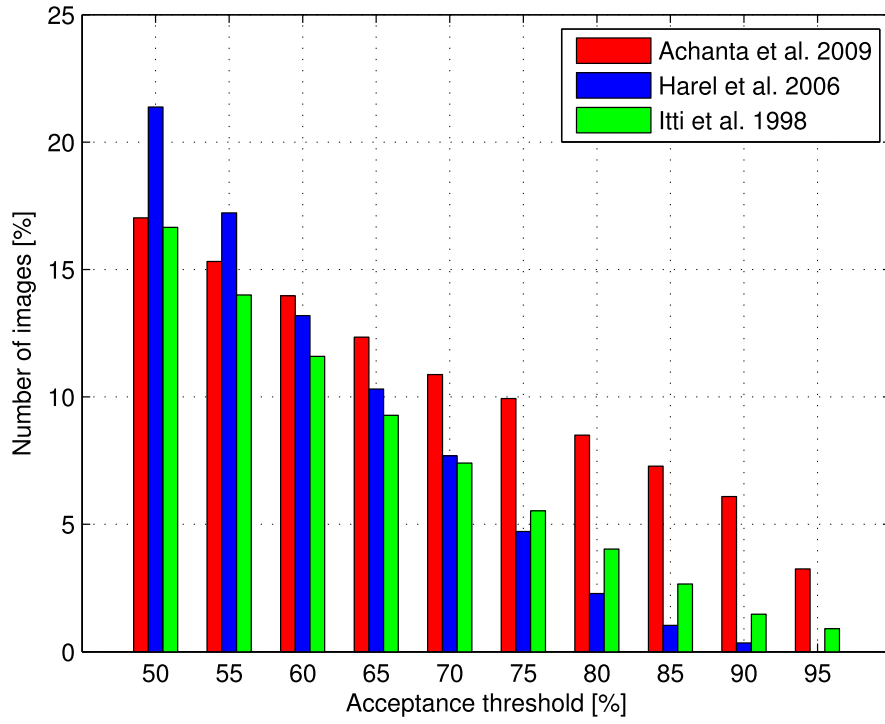
Visual atten- tion model	Original image	Saliency map	Adaptive thresholding		Mean-shift segmentation with adaptive thresholding	
			Binarized saliency map	Predicted object	Binarized saliency map	Predicted object
Achanta <i>et al.</i> [104]						
						
						
Harel <i>et al.</i> [109]						
						
						
Itti <i>et al.</i> [92]						
						
						

precise localization of salient objects (large portions of the background are deemed to be likely salient as well), which will be also confirmed later analyzing results in Figure 3.9. Throughout the observed range of *acceptance* thresholds, approaches by Achanta *et al.* and Itti *et al.* have similar performance (within 1 % of performance difference) and perform better than Harel *et al.*, for *acceptance* thresholds above 70 %. When the two salient object detection methods are compared among themselves, we can see that Harel *et al.* and Itti *et al.* perform better for mean-shift segmentation with adaptive thresholding, with a difference of 2–3 % from detection performed by adaptive thresholding. Therefore, there is a trade-off between the performance and the computational time, as a mean-shift segmentation is computationally very heavy method compared to simple adaptive thresholding. On the other hand, approach by Achanta *et al.* has similar performance regardless of the applied salient object detection method and keeps the same level of performance even when simple adaptive thresholding is applied.

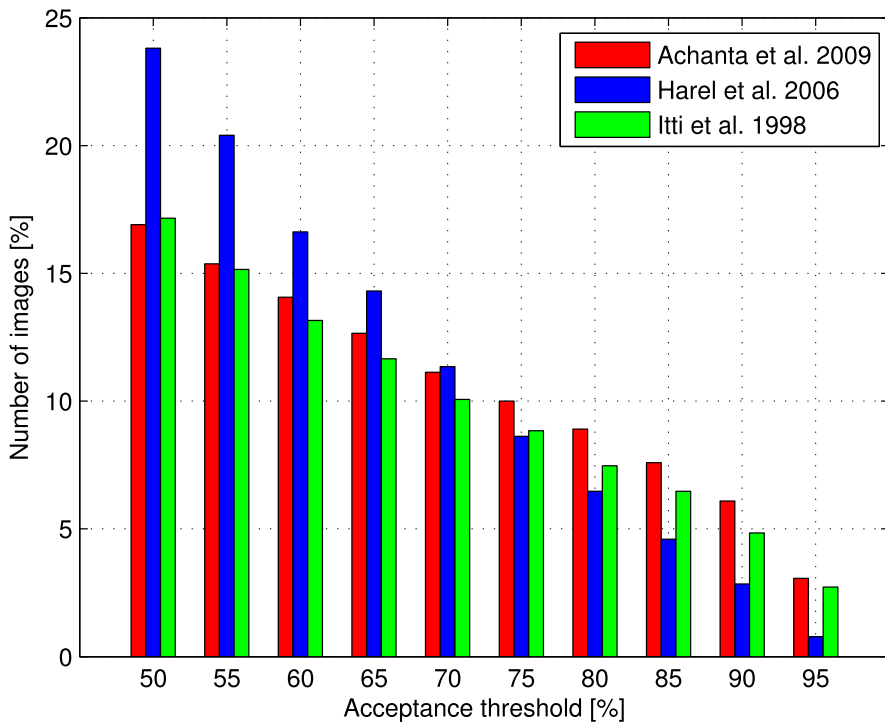
In order to evaluate how precise the considered visual attention models are, we fixed the *acceptance* threshold at 50 % for the reason we already explained in Section 3.5.2, and measured the *accuracy* value for all images that have the *acceptance* value above the threshold. Figure 3.9 shows the distribution of the images with respect to the *accuracy* values (ranging from 50 % to 100 %). This figure shows that if the objects in images are detected with the *acceptance* value above 50 %, then majority of the objects (45–50 %) are detected with high precision (above 90 % of *accuracy* value) when compared with the ground truth objects. This conclusion is valid for the visual attention approaches by Itti *et al.* and Achanta *et al.*, regardless of the object detection method. Approach by Achanta *et al.* has slightly better performance. However, approach by Harel *et al.* has lower precision than Itti *et al.* and Achanta *et al.* and around 30–35 % of images achieve above 90 % of *accuracy* value, which can be still considered as good result.

We then calculate the average *acceptance* values for each of the object classes from the evaluation database, namely books, buildings, cars, gadgets, newspapers, shoes, text, and trademarks. Results are shown in Figure 3.10. Both salient object detection methods, namely adaptive thresholding (see Figure 3.10 (a)) and mean-shift segmentation with adaptive thresholding (see Figure 3.10 (b)), have rather similar trends. The results are very much content dependent. The best performing class are shoes. One of the reasons for this is that background in this class of images is quite uniform, which leads to good segmentation and accurate prediction of salient objects for the three approaches by Achanta *et al.*, Itti *et al.* and Harel *et al.* On the other hand, classes text and newspapers perform bad, as these images have a lot of cluttered regions which are very hard to segment even by humans. The average *acceptance* values for these classes are below 15 %. The other object classes have the average *acceptance* values between 20–40 %. We can also confirm our previous conclusion that the more sophisticated salient object detection method based on mean-shift segmentation does not improve much the performance, as the maximum difference between this method and the simple adaptive thresholding is about 5–7 % of the average accuracy. The biggest improvement is achieved for the text class, where it reaches 10 % of the difference.

We also want to explore how the performance depends on the relative size of the object. We measured the relative size of the object in an image as the ratio between the pixels inside the

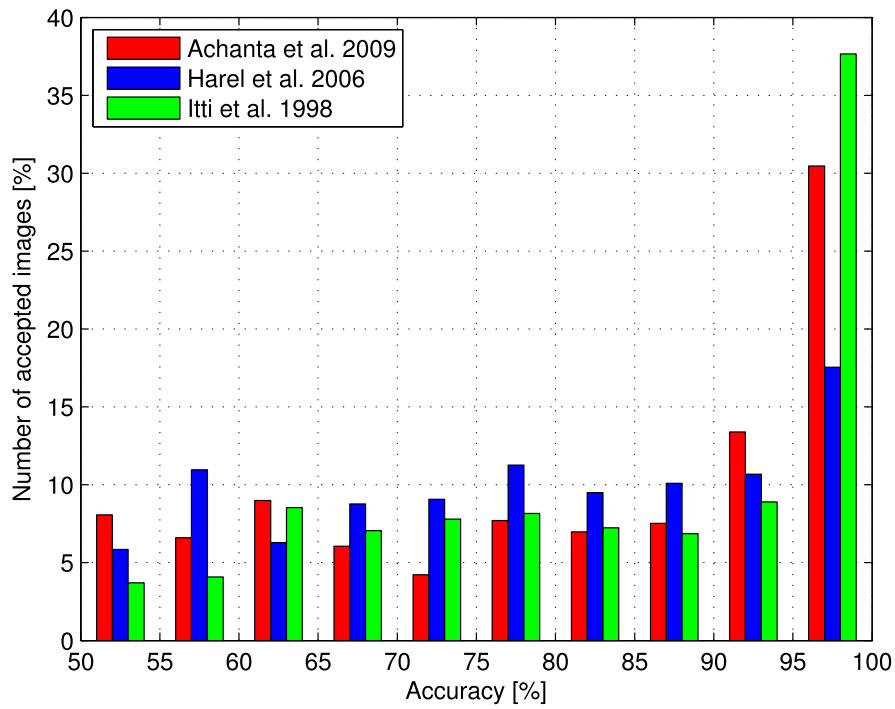


(a)

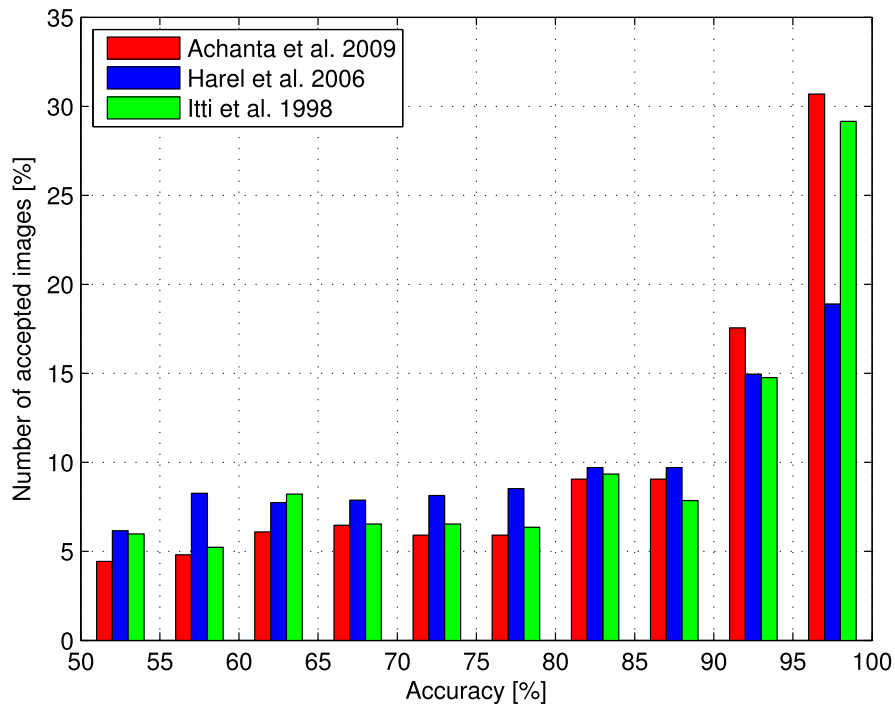


(b)

Figure 3.8: The distribution of the images with respect to the *acceptance* values obtained by applying considered visual attention models on the evaluation database of images. Two salient object detection methods are considered: (a) adaptive thresholding, and (b) mean-shift segmentation with adaptive thresholding.



(a)



(b)

Figure 3.9: The distribution of the images with respect to the *accuracy* values obtained by applying considered visual attention models on the evaluation database of images. Two salient object detection methods are considered: (a) adaptive thresholding, and (b) mean-shift segmentation with adaptive thresholding.

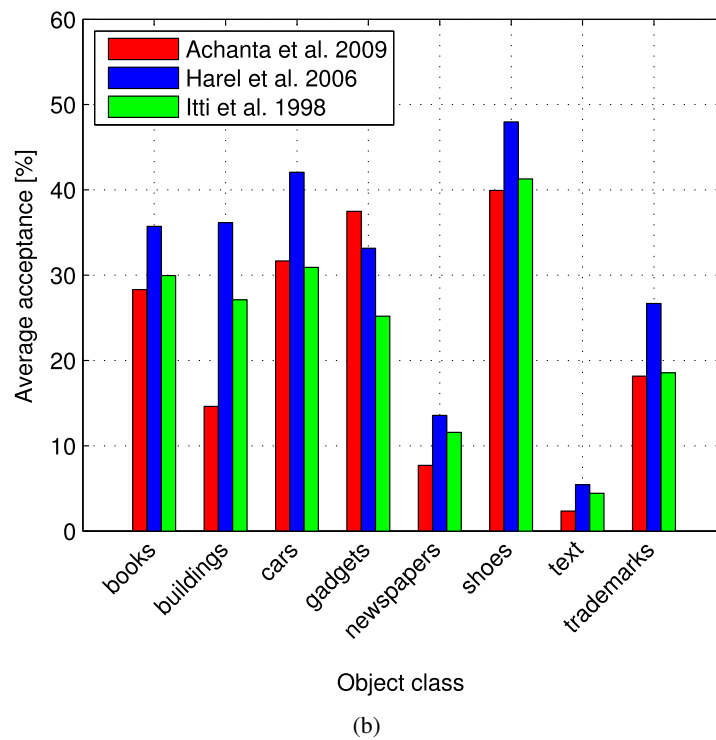
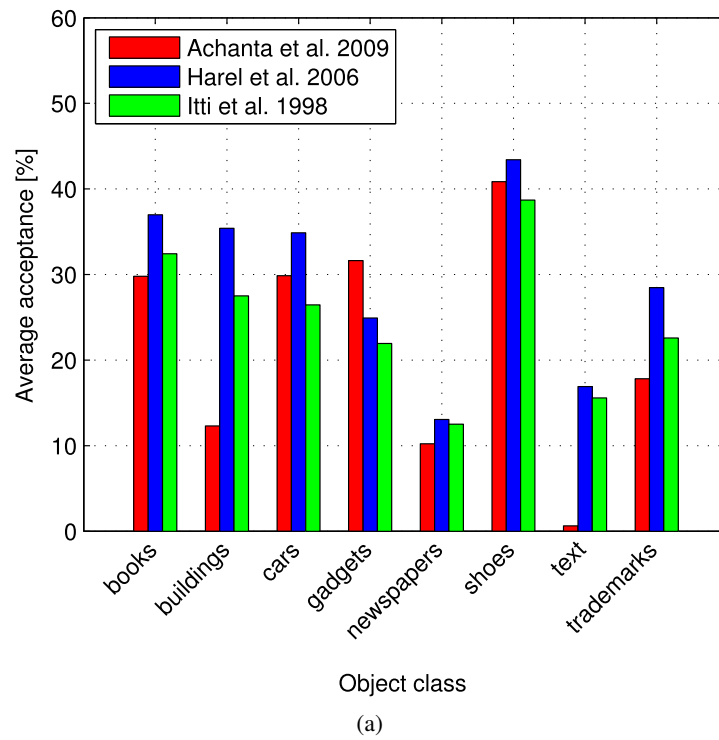


Figure 3.10: The average *acceptance* values for each of the object classes obtained by applying considered visual attention models on the evaluation database of images. Two salient object detection methods are considered: (a) adaptive thresholding, and (b) mean-shift segmentation with adaptive thresholding.

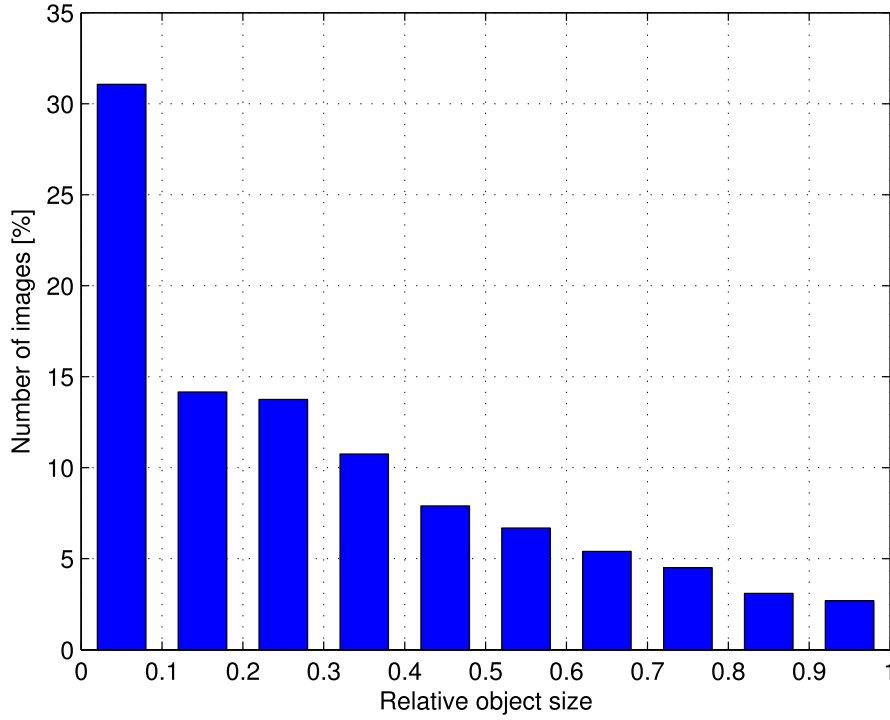
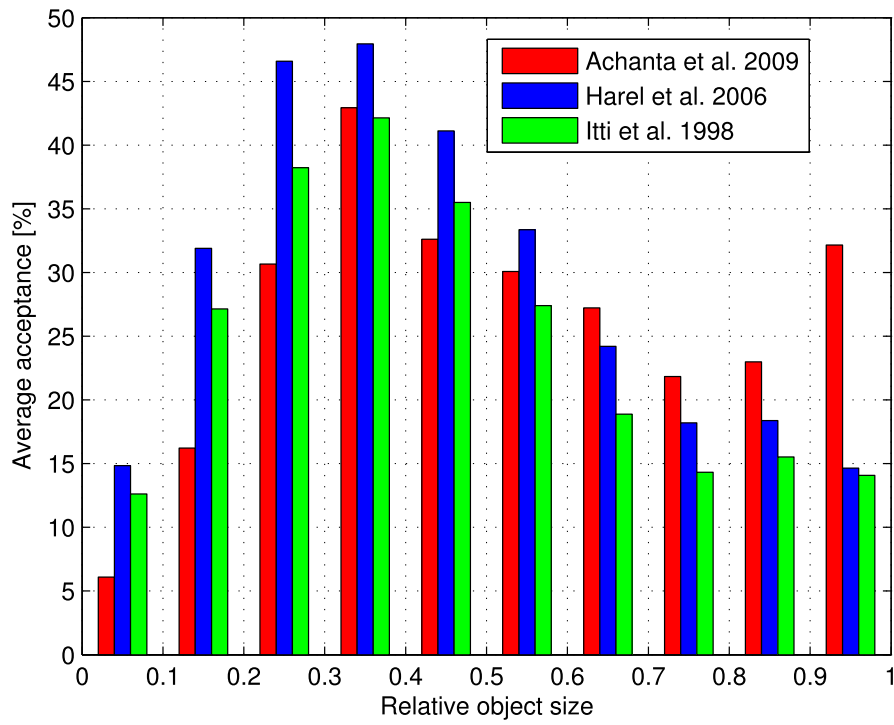


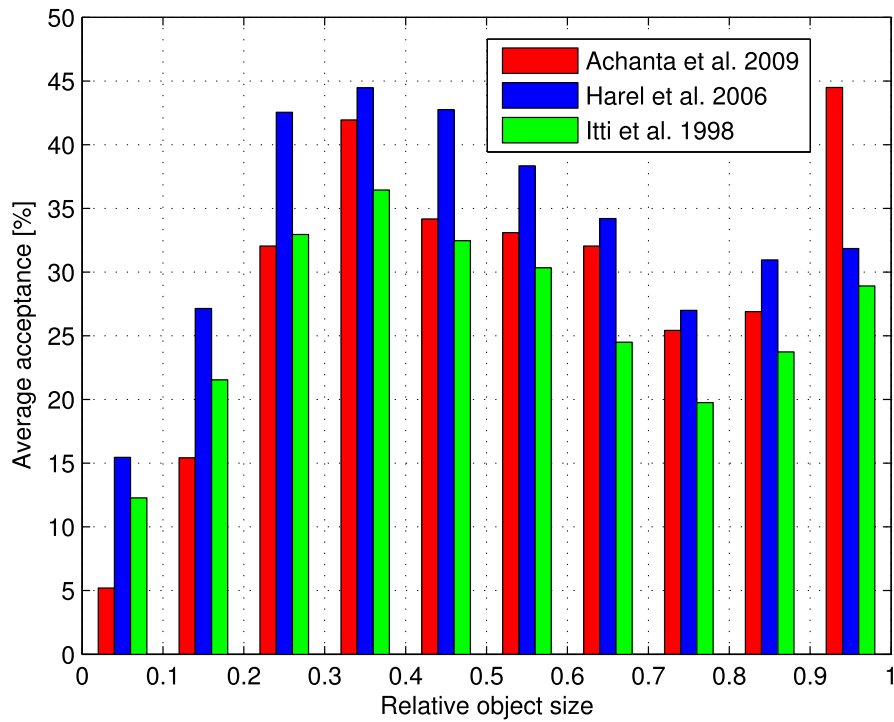
Figure 3.11: The distribution of the relative object sizes in the evaluation database of images.

bounding box outlining the object and the total number of pixels in the image. We calculated the histogram of the relative object sizes from the evaluation database of images. Relative sizes of the objects are split into bins of 0.1 from the image sizes and for each of the bins, number of images having object size in that particular range is calculated. The results are shown in Figure 3.11. Around 31 % of all images have objects smaller than 10 % of the entire image, while around 35 % of images have objects greater than 50 % of the image sizes. Furthermore, we calculate the average *acceptance* values for each of the bins of the relative object size. The results are shown in Figure 3.12. The best performance for all three visual attention models is achieved for the relative object size between 0.2–0.6, while for the lower and higher relative object sizes, the performance decreases. In this range from 0.2–0.6, the approach by Harel *et al.* outperforms the approaches by Itti *et al.* and Achanta *et al.* Interestingly, the approach by Achanta *et al.* shows good performance for the objects which cover almost the whole image (above 0.9 of relative object size) for both adaptive thresholding and mean-shift segmentation with adaptive thresholding. For the applications that require good precision of detected objects, the best would be to make use of the approach by Achanta *et al.*, as it shows good performance already beyond 0.6 of relative object size for the adaptive thresholding method. For the smaller-size objects (under 0.5 of relative object size), the approach by Harel *et al.* will be suitable.

At the end of the analysis of the considered visual attention models, we perform comparison in terms of computational time, which depends on the image resolution. Figure 3.13 shows the average time needed to detect salient objects for 50 randomly selected images of certain



(a)



(b)

Figure 3.12: The average *acceptance* values for each of the object classes obtained by applying considered visual attention models on the evaluation database of images. Two salient object detection methods are considered: (a) adaptive thresholding, and (b) mean-shift segmentation with adaptive thresholding.

resolutions from the evaluation database of images. The computational time includes the saliency map creation time and salient object detection time. Different resolutions were obtained by resizing (and cropping, if necessary) the source images. For salient object detection by adaptive thresholding (see Figure 3.13 (a)) we can see that the approach by Achanta *et al.* (red bars) outperforms other approaches for low resolution images. For higher-resolution images (above 1280×960 pixels), the approach by Itti *et al.* (green bars) is the fastest. The approach by Harel *et al.* (blue bars) is the slowest one. From Figure 3.13 (b) we can see that salient object detection by mean-shift segmentation with adaptive thresholding is very slow method, especially for the higher-resolution images where the calculation time differs by two orders of magnitude (factor of 100) compared to simple adaptive thresholding in Figure 3.13 (a). The majority of the calculation time in this case is taken by mean-shift segmentation algorithm, while the saliency map calculation time is negligible.

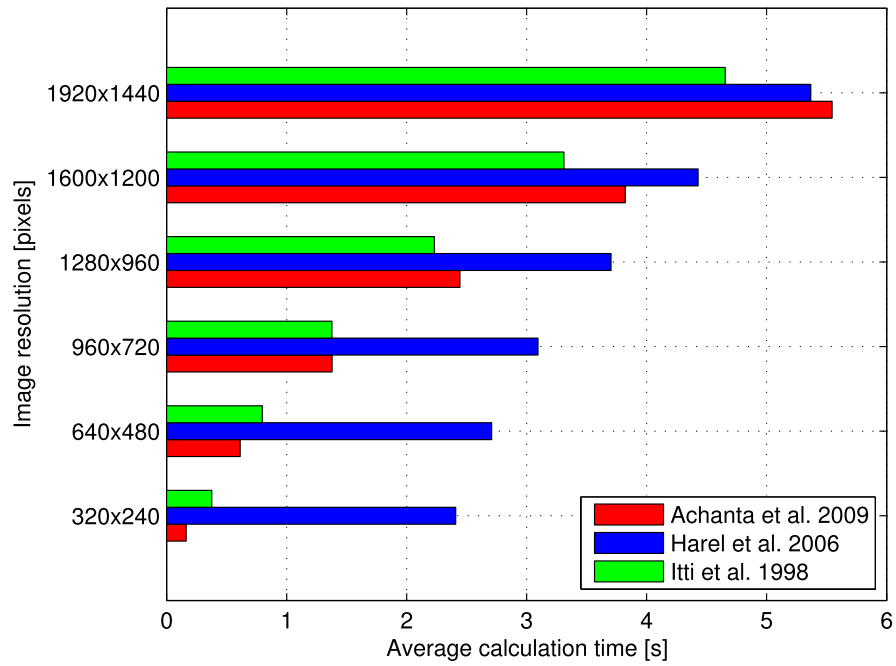
We showed here that the performance of the considered visual attention models vary across different image content and object sizes. There is also trade-off between the performance and complexity of the methods. The essential problems of current methods for saliency map calculation are that they produce low resolution maps with not accurately preserved borders between salient objects and background regions. Another problem of the current methods is that they emphasize saliency of the edges of the objects, but are not uniform over the entire object. However, the analysis showed that presented approaches are good enough for the visual search and tag propagation scenario. Especially the approach by Achanta *et al.* is precise enough to estimate the position of the most interesting objects in images, as its performance is good for higher accuracy values and it is fast when combined with adaptive thresholding for salient object detection to be used in a real-time tag propagation scenario.

3.6.2 Results of Visual Search

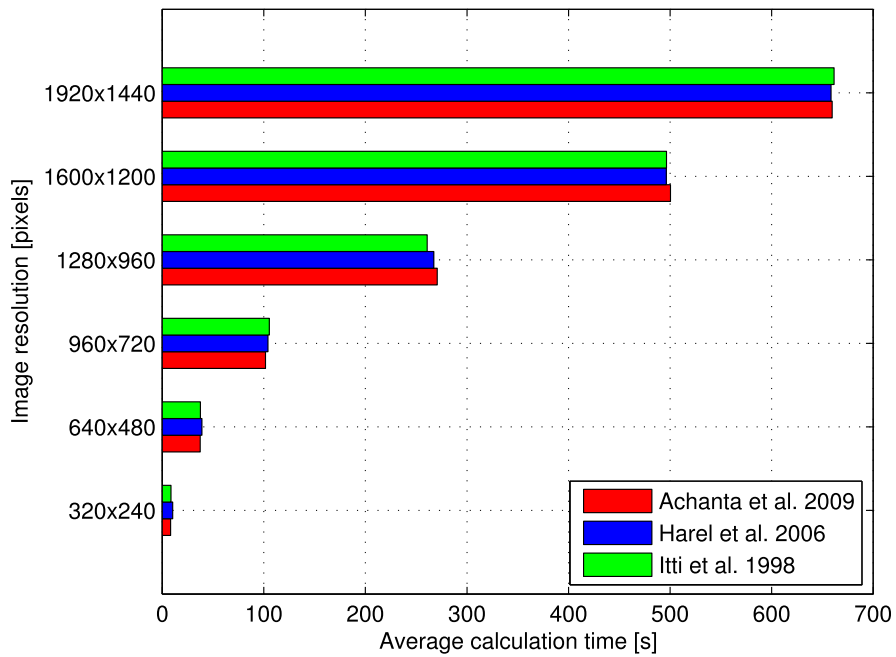
Once salient objects in images are predicted, they are fed as query images into the visual search module shown in Figure 3.4. For the thorough performance evaluation, the ground truth objects selected by the bounding boxes and the entire source images are also fed into the visual search module and their performance are measured.

The performance of different features used for the visual search are shown in Figures 3.14 and 3.15. Different subfigures show the results for different object classes. Within one subfigure, the results are presented for different types of a query image (namely, an entire image, a ground truth object, and an automatically selected object) within one vertical bar and for different evaluation measures (namely, *recall@5*, *recall@10*, and *recall@20*) as three connected vertical bars.

First, we compare global (colHSV, colLab, EOH, HTD) and local (SIFT, SURF, HOG) features. We can clearly see that local features perform significantly better for all classes, except for shoes class. One of the reasons is that shoes do not have enough discriminative local features, as it is shown in Table 3.3, and therefore there is only a small improvement when using local



(a)



(b)

Figure 3.13: The average calculation time for each of the considered visual attention models applied on the evaluation database of images. Two salient object detection methods are considered: (a) adaptive thresholding, and (b) mean-shift segmentation with adaptive thresholding. All timings were obtained using Intel Core2 Duo P9400 CPU running at 2.4 GHz with 4 GB of RAM memory.

Chapter 3. Saliency-Driven Automatic Extraction of Informative Image Content

Table 3.3: Statistical overview of the SIFT and SURF local descriptors extracted from the evaluation database. The average number of features per image is given for different classes, as well as, different types of query images.

Object class	Average number of features per image					
	SIFT			SURF		
	e. i. ^a	g. t. o. ^b	a. s. o. ^c	e. i. ^a	g. t. o. ^b	a. s. o. ^c
books	352.8	158.4	123.5	240.2	87.2	64.8
buildings	1242.6	525.8	220.5	707.1	282.3	110.4
cars	878.6	556.5	182.9	614.1	415.8	129.3
gadgets	785.3	622.3	346.5	555.2	406.1	218.4
newspapers	1119.6	156.6	400.9	495.9	60.6	169.6
shoes	400.5	254.6	159.7	223.4	134.7	71.5
text	2586.3	390.2	236.7	899.4	180.6	68.7
trademarks	919.8	244.2	216.9	651.1	116.5	145.3

^a entire image

^b ground truth object

^c automatically selected object

SIFT features. For the class of buildings, EOH features perform as good as local features, since buildings have a lot of corners and edges by its nature and this helps distinguishing between them. The largest performance improvement of local features over global features is for the class of trademarks, as trademark logos usually cover only a small area of an image, therefore global features perform very bad. The SURF feature is the most discriminative feature for the text and the newspapers classes.

Then, we compare values for *recall@5*, *recall@10*, and *recall@20*, and the trend is that these values always increase with respect to this order. However, this increase varies for different visual features, as well as, for different object classes. For example, the largest difference between *recall@20* and *recall@10*, and between *recall@20* and *recall@5* is achieved for the local features, such as SIFT and SURF, for the classes of text and newspapers. On the other hand, the smallest improvement between *recall@5* on one side and *recall@10* and *recall@20* on the other side is made for the global features of the class trademarks, for the already explained reasons. Similarly, the global features of the class cars make very small improvement from *recall@5* towards *recall@20*, probably due to very challenging images of this class in the evaluation database including variety of colors, view points, and object sizes, which do not help improving the performance.

At the end, we compare performance across different types of a query image: an entire image, a ground truth object, and an automatically selected object. In most cases, entire images perform better than ground truth objects. However, for the class of cars and gadgets, entire images show close performance to ground truth objects. Due to imperfections of visual attention models which are discussed in Section 3.6.1, automatically selected objects perform worse than ground truth

objects when global features are applied. Therefore, for the further analysis we consider only local features, in particular SIFT and SURF features. For these types of features, localization of an object (with a ground truth data) in most cases either improves or retains the performance. For example, for the trademarks class, improvement from entire image to ground truth object is around 17 % in terms of *recall@20* and SURF features. However, as we already showed that visual attention models have problems with precise localization of objects, this problem is reflected here and automatically selected objects in most cases perform worse than ground truth objects and entire images, or rarely reach the same level of performance as entire images. Therefore, to achieve the best performance in the object-based tag propagation scenario, additional user input is necessary. The user will have to manually adjust the borders of the bounding box around the object that is detected by one of the visual attention models. In this way, the performance of the system are expected to be close to the performance of the local SIFT and SURF features applied on ground truth objects, as depicted in Figures 3.14 and 3.15.

3.7 Conclusion

The ability of human visual system to detect visual saliency is extraordinarily fast and reliable. However, computational modeling of this basic intelligent behavior still remains a challenge. Efforts have been made in computational modelling the mechanism of human visual attention, especially the bottom-up attention mechanism.

In this chapter, we study three bottom-up attention approaches for extracting saliency in images, namely Itti *et al.* [92], Achanta *et al.* [104] and Harel *et al.* [109]. These methods are applied for automatic detection of interesting objects in images and visual search for object-based tag propagation. We perform an objective comparison of the accuracy of the saliency maps for three state-of-the-art methods using a ground truth of 3200 images depicting different object classes. We also evaluate features for visual search extracted from different object classes.

We showed here that the performance of the considered visual attention models vary across different objects in images and object sizes. There is also trade-off between the performance and complexity of the methods. For the applications that require good precision of detected objects, the best would be to make use of the approach by Achanta *et al.*, as it is precise enough to estimate the position of the most interesting objects in images and fast when combined with adaptive thresholding for salient object detection to be used in a real-time tag propagation scenario.

We also showed that local features perform significantly better than global features for majority of the object classes in visual search scenario. We compared performance across different types of a query image: an entire image, a ground truth object, and an automatically selected object by making use of visual attention model. Automatically selected objects in most cases perform worse than ground truth objects and entire images, or rarely reach the same level of performance as entire images. Therefore, to achieve the best performance in the object-based tag propagation scenario, additional user input, e.g., adjusting the borders of the bounding box

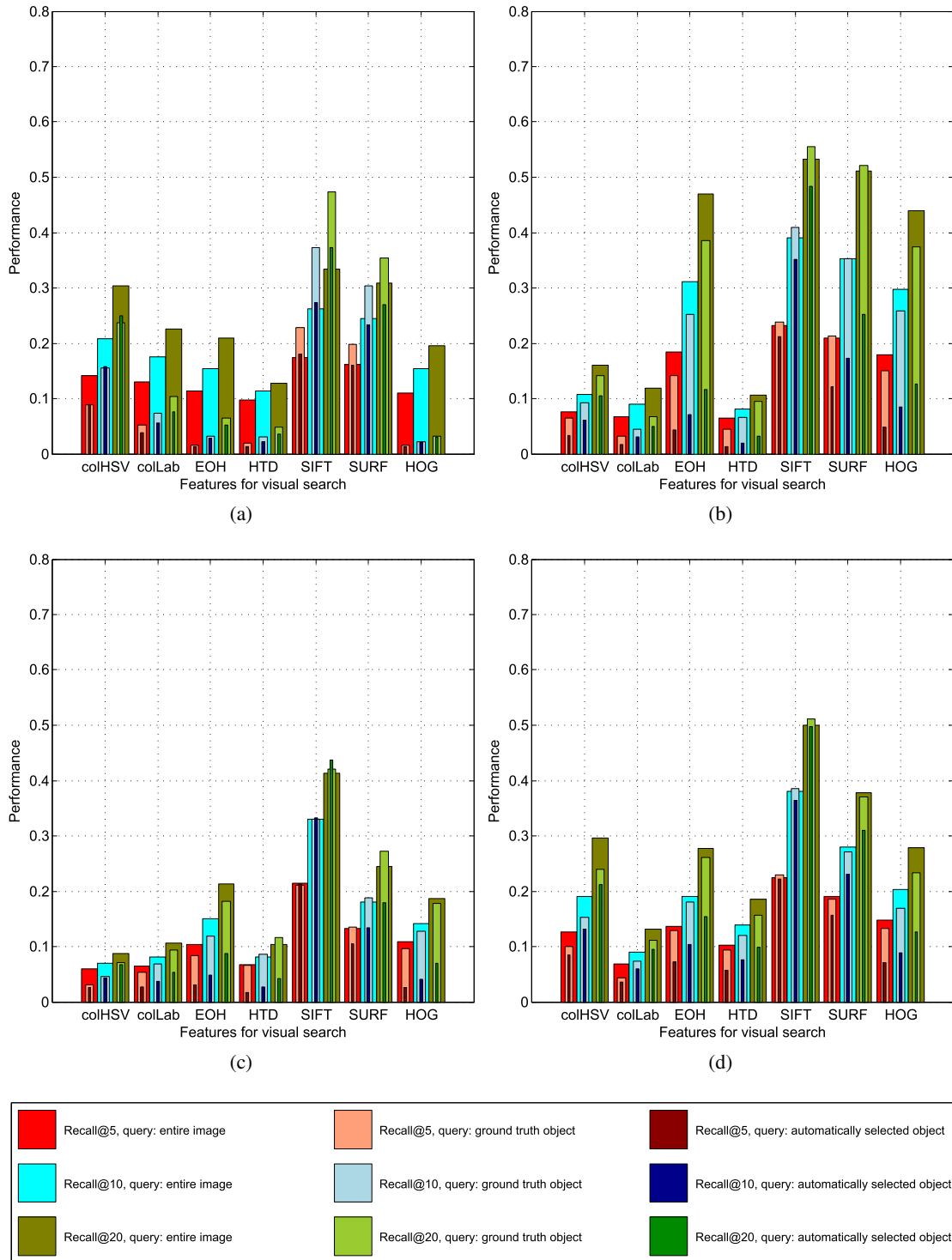


Figure 3.14: Performance evaluation of different features for visual search across object classes: (a) books, (b) buildings, (c) cars and (d) gadgets. Performance is measured as a rate of relevant images found in the top 5, 10 or 20 images. Visual search is performed using different query images: entire image, image of the ground truth object and image of the object selected by visual attention model.

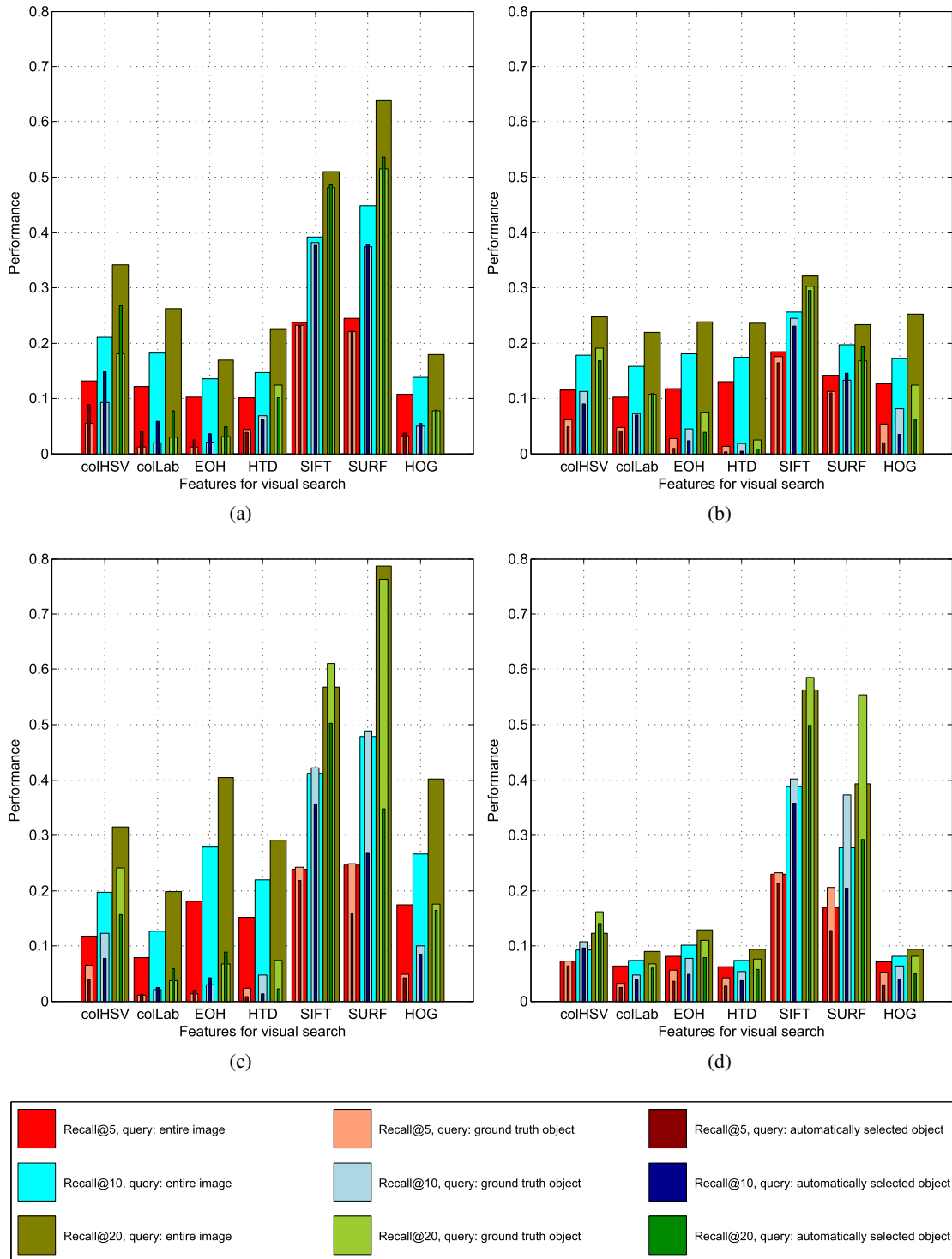


Figure 3.15: Performance evaluation of different features for visual search across object classes: (a) newspapers, (b) shoes, (c) text and (d) trademarks. Performance is measured as a rate of relevant images found in the top 5, 10 or 20 images. Visual search is performed using different query images: entire image, image of the ground truth object and image of the object selected by visual attention model.

Chapter 3. Saliency-Driven Automatic Extraction of Informative Image Content

around the predicted object, is necessary.

In this study we have experimented only with a single interesting object in an image. However, the study could be extended to include detection of multiple salient objects in images. The future work can also focus on more sophisticated visual search algorithms or other object classes. It would be interesting to see if visual attention models could predict the most representative photos of one photo album. We address this challenge through social gaming, as presented in Chapter 6.

User Enrichment Part II

4 User Trust Modeling for Automatic Landmark Tagging

Many images uploaded to social networks are related to travel, since people consider travelling to be an important event in their life. People often share their travel experiences by uploading photos with annotations, descriptions, and comments. However, a significant amount of travel images on the Internet lack proper geographical annotations or tags. In many cases, the images are tagged manually. One way to make this time-consuming manual tagging process more efficient is to propagate tags from a small set of tagged images to the larger set of untagged images automatically. In this chapter, we propose a system for automatic geotag propagation in images based on the similarity between image content (famous landmarks) and its context (associated geotags). For each tagged image, we find similar untagged images using the robust graph-based object duplicate detection and propagate the known tags accordingly. In such scenario, however, a wrong or a spam tag can damage the integrity and reliability of the automated propagation system. Users may make mistakes in tagging, or irrelevant tags and content may be added maliciously for advertisement or self-promotion. Therefore, for reliable geotags propagation, we suggest adopting user trust models based on a social feedback from the users of the photo-sharing system. By conducting experiments on an image database containing various landmarks, we compare different user trust models and demonstrate their effectiveness in a social tagging system.

Portions of this chapter are published in:

I. Ivanov, P. Vajda, J.-S. Lee, L. Goldmann, and T. Ebrahimi, “Geotag propagation in social networks based on user trust model,” *Multimedia Tools and Applications*, vol. 56, no. 1, pp. 155–177, 2012

I. Ivanov, P. Vajda, J. S. Lee, and T. Ebrahimi, “In tags we trust: Trust modeling in social tagging of multimedia content,” *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 98–107, 2012

I. Ivanov, P. Vajda, J.-S. Lee, P. Korshunov, and T. Ebrahimi, “Geotag propagation with user trust modeling,” in *Social Media Retrieval* (N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, and X.-S. Hua, eds.), Computer Communications and Networks, pp. 283–304, Springer-Verlag London, Jan. 2013

I. Ivanov, P. Vajda, P. Korshunov, and T. Ebrahimi, “Comparative study of trust modeling for automatic landmark tagging,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, 2013

4.1 Introduction

Social networks and photo sharing websites have become increasingly popular in recent years, since people use them to interact with each other by sharing their own interests or activities and exploring shared content (e.g., photo, video, text, and audio) of others. This sharing trend has resulted in a continuously growing volume of publicly available photos on Flickr, Picasa, and Facebook. For instance, 219 billion photos have been uploaded on Facebook since 2005 [36]. Tagging is one of the popular mechanisms that helps managing large volume of photos. Tags, when combined with search technologies, are essential in resolving user queries targeting shared photos. However, tagging a lot of photos by hand is a time-consuming task. Users typically tag a small number of the shared photos only, leaving most of the other photos with incomplete metadata. This lack of metadata decreases the precision of search, because photos without proper annotations are typically much harder to retrieve than correctly annotated photos.

A significant subset of shared photos in social networks or photo sharing websites is travel related. Travel is an important type of event for which people like to share, annotate and search pictures. In a large-scale analysis of users' tagging behavior and the information provided by tags, Sigurbjörnsson and van Zwol [62] found that 28 % of the tags in a random set of 52 million photos from Flickr corresponded to the location type of WordNet [131] categories. Travel images are mostly annotated with names of locations where images were captured. For the majority of travel images on the Internet, however, proper geographical annotations are not available. Usually, the most salient region in the image corresponds to a specific landmark or object. When users annotate such images, they link a geotag to the object depicted in the image. In order to speed up the time-consuming manual tagging process, geotags can be propagated based on the similarity between image content (usually famous landmarks) and context (associated geotags). Therefore, we propose to use object duplicate detection for the propagation of geotags since it is robust in detecting the same object and discarding similar objects. Untagged images are automatically annotated based on the detection of the same object from a small set of training images with associated tags.

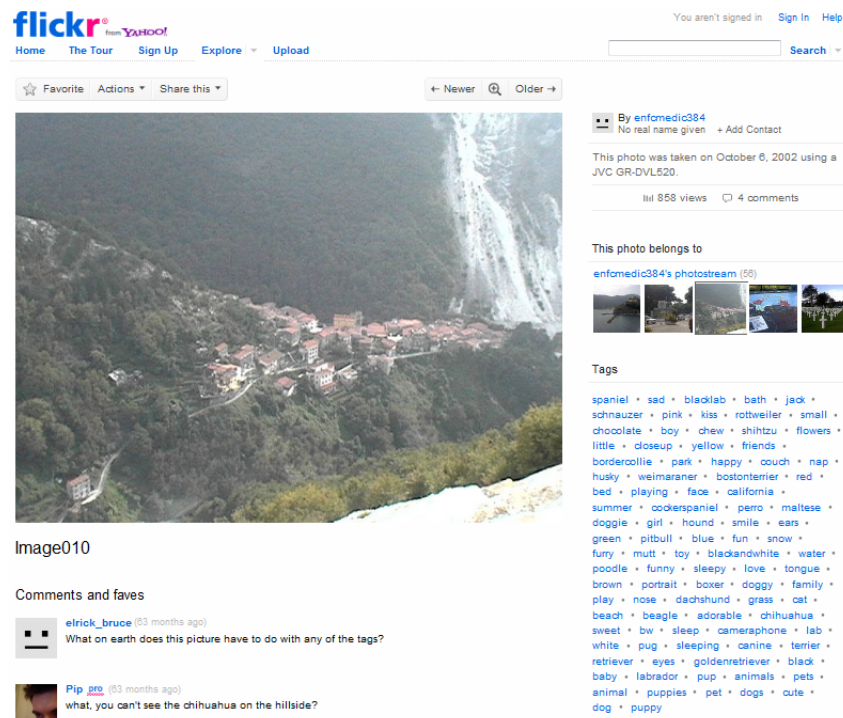
One important challenge in tagging is to identify most appropriate tags for given content, and at the same time, to eliminate noisy or spam tags. Shared photos can be assigned with inappropriate tags for several reasons. First of all, users are human beings and make mistakes. Moreover, it is possible to provide wrong tags on purpose for advertisement, self-promotion, or to increase the rank of a particular tag in automatic search engines. Consequently, free-form keywords (tags) assigned to photos carry a significant risk that wrong or irrelevant tags eventually prevent users from the benefits of annotated photos. Finally, wrong machine tags, such as longitude and latitude, can be automatically assigned to images captured with cameras equipped with GPS devices due to bad or noisy communication channels with GPS satellites or wireless access points. Kennedy *et al.* [132] analyzed the Flickr website and revealed that the tags provided by users are often imprecise and only around 50 % of tags are truly related to an image. Figure 4.1 shows examples of imprecise or spam tags and incorrect geotags in a popular photo sharing website. Beside the tag-photo association, spam objects can take other forms, i.e. possibly manifesting as

a spam photo or a spam user (spammer). Therefore, for the practical tag propagation system, it is important to consider user trust information derived from users' tagging behavior.

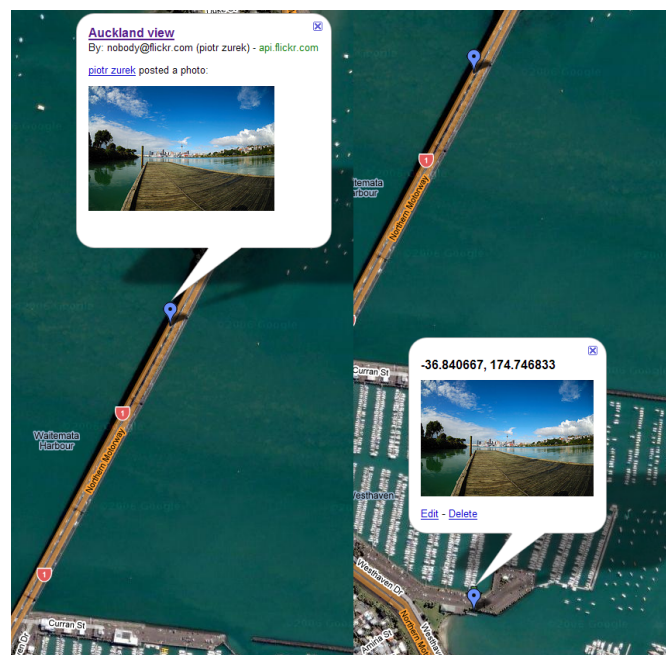
Trust provides a natural security policy stipulating that users or photos with low trust values should be investigated or eliminated. Trust can predict the future behavior of users in order to avoid undesirable influences of untrustworthy users. Trust-based schemes can be used to motivate users to positively contribute to social networks and/or penalize adversaries who try to disrupt the network. The distribution of the trust values of the users or photos in a social network can be used to represent the health of that network.

In this chapter, we consider traveling as one of the human activities highly influenced by social media in recent years [134]. As majority of travelers read blogs and reviews, and watch photos to select destinations, it is important to increase photos "findability" through geotagging of famous landmarks in images. For accurate and reliable automatic geotagging, we propose to combine the object (e.g., famous landmark) duplicate detection with the user trust modeling, which we published in [51, 54]. The user trust modeling reduces the risk of propagating wrong tags caused by spamming or faulty annotation. In a real-life scenario, an image with unknown landmark will be automatically tagged with either one geotag or none, depending on the level of similarity with the known (trained) landmarks. This scenario is denoted as open set problem. The less complex scenario is closed set problem, where each test image depicting unknown landmark is assumed to correspond to exactly one of the known landmarks. Therefore, the test image gets assigned to the most probable trained landmark and the corresponding geotag is propagated to the test image. We first evaluate the accuracy of our automatic geotag propagation system without including users and their mistakes in the annotation process, for both open and closed set problems. Although we use the proposed automatic geotagging system, in this chapter we do not focus on its evaluation of computational efficiency, nor on comparison of different geotagging methods. Furthermore, we present our socially-driven user trust model, previously published in Ivanov *et al.* [51], and compare it with four other techniques for trust modeling in social tagging systems, namely, Koutrika *et al.* [90], Liu *et al.* [14], Xu *et al.* [135], and Krestel and Chen [136]. The effectiveness of these models is demonstrated through a set of experiments on an image database with different landmarks. To create an environment for trust modeling where object duplicate detection performs well, we consider only closed set problem. This constrain allows us to thoroughly analyze and compare trust models, while having less errors introduced in the system by the object duplicate detection module. Using robust trust model and robust object duplicate detection, we can create more accurate system. Spam in social systems cannot be controlled, however, the performance of object duplicate detection can be improved, which will lead to more accurate system for automatic geotagging with trust modeling.

The remaining sections of this chapter are organized as follows. We introduce related work of geotagging in Section 4.2. Our and four other selected trust models are presented in Section 4.3. Section 4.4 describes our approach for geotag propagation between images and discusses two application scenarios. Methodology and results for the performance evaluation of the proposed geotagging system and comparison of trust models are presented in Sections 4.5 and 4.6, respec-



(a)



(b)

Figure 4.1: Examples of imprecise or spam tags and incorrect geotags in Flickr: (a) wrong tags – only a few tags in the list are related to the image, while the rest is irrelevant (e.g., yellow, love, doggy) (screenshot retrieved in March 2011), (b) incorrect geotags – the left side of the picture shows the point placed at incorrect location acquired from GPS-enabled camera, while the right side shows the manually created Google Maps point with correct data and the incorrect point slightly above on the bridge (image source: [133]).

Chapter 4. User Trust Modeling for Automatic Landmark Tagging

Table 4.1: Summary of representative recent techniques that combine geographical context and visual content for automatic geotagging of images.

Reference	Descriptor	Method	Application
Hays and Efros [138]	visual features	the probability distribution for the location of an unknown image is found on the globe using a purely data-driven scene matching	non-landmark (scene) location recognition
Kennedy and Naaman [139]	visual and textual features	for a given location diverse and representative images are generated based on geotagged community images	visual summary of landmarks
Zheng <i>et al.</i> [140]	visual features & GPS coordinates	travel blogs and geotagged images are analyzed and a list of tourist landmarks is established based on the information from nearest neighbors	landmark recognition
Quack <i>et al.</i> [73]	visual and textual features & GPS coordinates	objects and events are retrieved from a large-scale collection of geotagged images using pair-wise similarity	event/scene understanding

tively. Finally, Section 4.7 concludes the chapter with a summary and some perspectives for the future work.

4.2 Related Work

The proposed system is related to different research fields including visual analysis, geographic information systems, and social networking and tagging systems. Therefore, the goal of this section is to review the most relevant work in the fields of joint analysis of visual content and geographical context, and manual tagging, while the next section provides insight into trust modeling in social tagging systems.

In the last several years, an important trend in multimedia understanding is modeling and extracting value from geographical context, such as GPS coordinates, and visual content, such as photo description. Different research problems and significant approaches in this field are summarized by Luo *et al.* [137]. In this section, we focus on some of the representative image retrieval approaches that rely on a variety of image or landmark descriptors combined with geographic information. These approaches are summarized in Table 4.1.

A pioneering paper in this area by Hays and Efros [138] proposed an algorithm called IM2GPS to estimate the locations of a single image using a purely data-driven scene matching approach.

Given a test image, the algorithm finds the visual nearest neighbors in the database and estimates a geolocation of the image from the GPS coordinates of the tagged nearest neighbors. The estimated image location is represented as a probability distribution over the Earth's surface. However, the IM2GPS approach showed low recognition accuracy due to low-level features. While IM2GPS uses a set of more than 6 million training images, its general applicability is inconclusive, because the performance was verified only on 237 hand-selected test images.

Kennedy and Naaman [139] presented a method to search representative landmark images from a large collection of geotagged images. This method uses tags and the geographical location representing a landmark. The visual features (global color and texture features, and scale-invariant feature transform (SIFT)) are analyzed to cluster landmark images into visually similar groups. The method has been proven to be effective for extraction of the representative image sets for a given landmark. But since it cannot be applied to untagged images, its applicability is limited.

The recent work of Zheng *et al.* [140] automatically finds frequently photographed landmarks from a large collection of geotagged photos. The authors perform clustering on GPS coordinates and visual texture features from the image pool and extract landmark names as the most frequent tags associated with the particular visual cluster. Additionally, they extract landmark names from the travel guide articles, such as Wikitravel, and visually cluster photos gathered by querying Google Images. However, the test set they use is quite limited – 728 images in total for a 124-category problem, or less than 6 test images per landmark.

Another application that combines textual and visual techniques has been proposed by Quack *et al.* [73]. The authors developed a system that crawls photos on the internet and identifies clusters of images referring to a common object (physical items at fixed locations), and events (special social occasions taking place at certain times). The clusters are created based on the pair-wise visual similarities between the images, and the metadata of the clustered photos are used to derive labels for the clusters. Finally, Wikipedia articles are attached to the images and the validity of these associations is checked. Gammeter *et al.* [74] extends this idea towards object-based auto-annotation of holiday photos in a large database that includes landmark buildings, statues, scenes, pieces of art, with help of external resources such as Wikipedia. In both [73] and [74], GPS coordinates are used to pre-cluster objects which may not be always available.

A commercial application by Google, called Google Goggles, is created for landmark search on mobile phones. It also detects logos, book and DVD covers, artworks, and products.

Most of the photo-sharing websites (e.g., Flickr, Picasa, Panoramio), provide information about where images were taken in form of maps or groups. This information is either provided by an external GPS sensor and stored as image metadata (Exchangeable Image File Format (EXIF) [141], International Press Telecommunications Council (IPTC) [142]), or manually annotated via geocoding. Our goal is to obtain this information by comparing the content of the image with a small set of already tagged images.

The main disadvantages of the above systems is that they rely on GPS coordinates to derive

geographical annotation, which is not available for the majority of web images and photos, since only a few camera models are equipped with GPS devices. Furthermore, a GPS sensor in a camera provides only the location of the photographer instead of that of the captured landmark, which may be up to several kilometers away. Therefore, the GPS coordinates alone may not be enough to distinguish between two landmarks within a city. Describing landmarks through location names rather than GPS coordinates is not only more reliable but also more expressive. A recent study by Hollenstein and Purves [143] indicated that geotagging should follow the way people actually describe locations, i.e. it is more convenient to use: Church of Saint Sava in Belgrade, rather than: latitude 44.798083 and longitude 20.46855. Therefore, there is a growing interest in the research community to derive geographic locations of the scenes in photos based on visual and text features.

4.3 Trust Modeling in Geotagging Applications

When information is exchanged on the Internet, malicious individuals are everywhere trying to take advantage of the information exchange structure for their own benefit, while bothering and spamming others. In this section, we present and discuss several techniques for combatting noise and spam through trust modeling in social tagging systems. First, we remind the reader of the model of a social tagging system, which is previously introduced in Section 1.2.2. Then, we present in details our (see Section 4.3.1) and four other recent techniques for user trust modeling that are suitable for a specific application of geotagging and can be used in geotagging the shared content and efficient propagation of such tags throughout the untagged content. Other techniques for trust modeling in social tagging systems are discussed in the next chapter in Section 5.2.

As previously described in Section 1.2.2, the entities that make up the model of a social tagging system [13] are: *users* who interact with the system, *content* (resources or documents), which is a piece of information such as photo, video, textual document, or web pages; and *tags*, the descriptions assigned to the piece of the content by users. The action of associating a tag to a content by a user is usually referred to as *tag assignment* [14].

Trust modeling methods can be categorized into two classes according to the target of trust, i.e. content and user trust modeling, as we previously reported in [52]. *Content trust modeling* is to classify content as spam or legitimate. In this case, the target of trust is a content, and thus a trust value is given to each content. In *user trust modeling*, trust is given to each user based on the information extracted from a user's account, his/her interaction with other participants within the social network, and/or the relationship between the content and tags that the user contributed to the social network. Given a user trust value, the user might be flagged as a legitimate user or spammer. Table 4.2 summarizes five user trust models, which we then describe in more details (in the same order as they are presented in the table). And Table 4.3 summarizes the notations used for their detailed description. These methods are different in the targeted media content, for which the geotagging is used, the applications they are intended for, and the required level of participation from the users of the geotagging system.

4.3. Trust Modeling in Geotagging Applications

Table 4.2: Summary of five trust modeling techniques used for combatting noise and spam in social tagging systems.

Reference	Content	Method	Database
Ivanov <i>et al.</i> [51]	images	an approach based on the feedback from other users who agree or disagree with a tag associated with an image	Panoramio, real
Koutrika <i>et al.</i> [90]	bookmarks	a coincidence-based model for query-by-tag search, which estimates the level of agreement among different users in the system for a given tag	Delicious, real & simulated
Liu <i>et al.</i> [14]	bookmarks	an iterative approach to identify spam content by its information value extracted from the collaborative knowledge	Delicious, real
Xu <i>et al.</i> [135]	bookmarks	an iterative approach to compute the goodness of each tag with respect to a content and the authority scores of the users	MyWeb 2.0, real
Krestel and Chen [136]	bookmarks	a TrustRank-based approach using features which model tag co-occurrence, content co-occurrence and co-occurrence of tag-content	BibSonomy, real

Table 4.3: Notation used in this chapter.

Notation	Explanation
U	Set of users
D	Set of documents (content)
T	Set of tags
$P \in U \times D \times T$	Set of tag assignments
$u \in U$	A user
$d \in D$	A document (content)
$t \in T$	A tag
$p \in P$	A tag assignment
$trust^{model}(u)$	Trust value of a user u calculated for particular model
L	Total number of users
M	Number of training documents (photos)
N	Number of test photos
$ x : condition $	Number of $x \in \{u, d, t, p\}$ which satisfy <i>condition</i>

4.3.1 User Reliability Based Model

In this section, we describe our own approach for user trust modeling in image tagging, which was published in Ivanov *et al.* [51].

First, we evaluate the trust or reliability of users by making use of their past behavior in tagging. We want to distinguish between users who provide reliable geotags, and those who do not. After user evaluation and trust model creation, tags will be propagated to other photos in the database only if the user is trusted. Assuming that there are L users who tag M training images, a matrix $R(i, u)$, $i \in [1 \dots M]$ and $u \in [1 \dots L]$, is defined as:

$$R(i, u) = \begin{cases} 1, & \text{if user } u \text{ tags image } i \text{ correctly;} \\ 0, & \text{otherwise.} \end{cases} \quad (4.1)$$

The process of comparing the propagated tags to ground truth tags can be done automatically using tag similarity measures, for example WordNet [144] or Google distance [145] measures. As another example, automatic tag checking can be done by making use of WordNet [131] and external resources with images and text (e.g., Wikipedia). Given the predicted geotag, WordNet returns a set of the closest words (tags) to that geotag, and this set of tags is used to acquire ground truth images from Wikipedia. Then, the object duplicate detection is performed on retrieved images to see if the tags from WordNet correspond to the given image. Nevertheless, we considered only manually defined ground truth for our experiments.

A trust value for user u , $trust^{Ivanov}(u)$, is computed as the percentage of the correctly tagged images among all images tagged by user u :

$$trust^{Ivanov}(u) = \frac{\sum_{i=1}^M R(i, u)}{M}. \quad (4.2)$$

In this approach, ground truth data are used for the estimation of the user trust value. However, for a practical photo sharing system, such as Panoramio, it is not necessary to collect ground truth data since user feedback can replace them. The main idea is that users evaluate tagged images by assigning a true or a false flag to the tag associated with an image. If the user assigns a false flag, then he/she needs to suggest a correct tag for the image. The more misplacements a user has, the more untrusted he/she is. By applying this method, spammers and unreliable users can be efficiently detected and eliminated. Therefore, the user trust value is calculated as the ratio between the number of true flags and all associated flags over all images tagged by that user. The number of misplacements in Panoramio is analogous to the number of wrongly tagged images in our approach.

In case that a spammer attacks the system, other users can collaboratively eliminate the spammer. First, the spammer wants to make other users untrusted, so he/she assigns many false flags to

the tags given by the trusted users and sets new wrong tags to these images. In this way, the spammer becomes trusted. Then, other users correct the tags given by the spammer, so that the spammer becomes untrusted and all of his/her feedbacks in the form of flags are not considered in the whole system. Finally, previously trusted users, who were untrusted due to spammer attack, recover their status. Following this scenario, the user trust value can be constructed by making use of the feedbacks from other users who agree or disagree with the tagged location. However, due to the lack of a suitable database which provides user feedback, the evaluation of the user trust scenario is based on the simulation of the social network environment.

4.3.2 A Coincidence-based Model

Koutrika *et al.* [90] were the first to explicitly discuss methods of tackling spamming activities in social tagging systems. The authors studied the impact of spamming through a framework for modeling social tagging systems and user tagging behavior. They proposed a method for ranking content matching a tag based on taggers' reliability in social bookmarking service Delicious. Their coincidence-based model for query-by-tag search estimates the level of agreement among different users in the system for a given tag. A bookmark is ranked high if it is tagged correctly by many reliable users. A user is more reliable if his/her tags more often coincide with other users' tags.

In more formal way, the following calculations are performed:

$$c(u) = \sum_{d,t:\exists P(u,d,t)} \sum_{u_i \in U: u_i \neq u} |p : \exists P(u_i, d, t)|, \quad (4.3)$$

$$score(d, t) = \frac{\sum_{u:\exists P(u,d,t)} c(u)}{\sum_{u \in U} c(u)}, \quad (4.4)$$

$$trust^{Koutrika}(u) = \sum_{d,t:\exists P(u,d,t)} score(d, t), \quad (4.5)$$

where $c(u)$, coincidence factor of the user u , is the number of other users u_i who assigned the same tag t to the same document d as the user u did. Score of the document d with respect to the tag t , denoted as $score(d, t)$, is calculated as a normalized value of c over all users who assigned t to d . Finally, a trust value of the user u , $trust^{Koutrika}(u)$, is the sum of $score(d, t)$ over all tag assignments by u .

Koutrika *et al.* performed a variety of evaluations of their trust model on controlled (simulated) database by populating a tagging system with different user tagging behavior models, including a good user, bad user, targeted attack model and several other models. Using controlled data,

interesting scenarios that are not covered by real-world data could be explored. It was shown that spam in tag search results using the coincidence-based model is ranked lower than in results generated by, e.g. a traditional occurrence-based model, where content is ranked based on the number of posts that associate the content to the query tag.

4.3.3 A Wisdom of Crowds Model

Liu *et al.* [14] proposed a simple but effective approach for detecting spam content in Delicious, by harvesting the wisdom of crowds. An information value of a bookmark is defined as the average number of times that each tag of the content is assigned by different users. A low information value of a bookmark indicates a divergence from crowds, which can be considered as a spam content. Furthermore, this method was extended to user trust modeling by aggregating the information values for each user.

All measures are defined as follows:

$$it(d, t) = \frac{|u : \exists P(u, d, t)|}{\sum_{t' \in T} |u : \exists P(u, d, t')|}, \quad (4.6)$$

$$ic(u, d) = \frac{\sum_{t: \exists P(u, d, t)} it(d, t)}{|t : \exists P(u, d, t)|}, \quad (4.7)$$

$$I(d) = \frac{|u : \exists P(u, d, \cdot)|}{\sum_{d' \in D} |u : \exists P(u, d', \cdot)|}, \quad (4.8)$$

$$trust^{Liu}(u) = \sum_{d: \exists P(u, d, \cdot)} I(d) \cdot ic(u, d), \quad (4.9)$$

where $it(d, t)$ represents the tag's t tagging information value with respect to document d , and $ic(u, d)$ is the information value of the content (document) d with respect to user u . The importance of the document d is defined by $I(d)$. Finally, a trust value of the user u , $trust^{Liu}(u)$, is calculated as the weighted average of the information value of the content tagged by user u , with the importance of the document as weight.

An interesting point is that, for the time being, Liu *et al.* collected the largest database for trust modeling by crawling Delicious [51]. This database had around 82 thousand users, 1.1 million tags, 9.3 million bookmarks and 17.4 million tag-bookmark associations.

4.3.4 An “Authority” Model Based on Goodness of Tags

Xu *et al.* [135] introduced the concept of “authority” in social bookmarking systems, where they measured the goodness of each tag with respect to a content by the sum of the authority scores of the users who have assigned the tag to the content. Authority scores and goodness are iteratively updated by using Hyperlink-Induced Topic Search (HITS) algorithm, which was initially used to rank web pages based on their linkage on the web [146].

Following measures are defined and iteratively calculated:

$$s_{i+1}(d, t) = \sum_{u: \exists P(u, d, t)} trust_i^{Xu}(u), \quad (4.10)$$

$$trust_i^{Xu}(u) = \frac{\sum_{d, t: \exists P(u, d, t)} s_i(d, t)}{|t : \exists P(u, ., t)|}, \quad (4.11)$$

where $i \in [1 \dots Q]$, $s_i(d, t)$ is the goodness of each tag t with respect to a content d , and $trust_i^{Xu}(u)$ represents a trust value (authority score) of the user u . Initial settings in this iterative approach are: $s_0(d, t) = 0, \forall t, d$ and $trust_0(u) = 1, \forall u$. The number of iterations is set to $Q = 100$.

4.3.5 A Co-occurrence Model

In contrast to the approach of Xu *et al.* [135], Krestel and Chen [136] iteratively updated values for users only. The authors proposed to use a spam value propagation technique to propagate trust values through a social graph in BibSonomy, where edges between nodes (in this case, users) indicate the number of common tags supplied by users, common content annotated by users and/or common tag-content pairs used by users. Starting from a manually assessed set of nodes labeled as spammers or legitimate users with the initial spam values, a TrustRank metric is used to calculate and iteratively update spam values for all users. TrustRank metric is previously introduced in [49] to semi-automatically separate reputable from spam web pages. This metric relies on an important empirical observation called approximate isolation of the good set: good pages seldom point to bad ones.

All measures are calculated as follows:

$$W(u_1, u_2) = |t : \exists P(u_1, ., t), P(u_2, ., t)| + |d : \exists P(u_1, d, .), P(u_2, d, .)| + |d, t : \exists P(u_1, d, t), P(u_2, d, t)|, \quad (4.12)$$

$$Tr(u_1, u_2) = \frac{W(u_1, u_2)}{\sum_{v \in U} W(u_1, v)}, \quad (4.13)$$

$$trust_i^{Krestel}(u) = \alpha \cdot \sum_{v \in U} Tr(u, v) \cdot trust_{i-1}^{Krestel}(v) - (1 - \alpha)d(u), \quad (4.14)$$

where $i \in [1 \dots Q]$, $W(u_1, u_2)$ is the weight of the edge between users u_1 and u_2 in the social graph and $Tr(u_1, u_2)$ is the corresponding transition matrix. A trust value of the user u , $trust_i^{Krestel}(u)$, is iteratively calculated. Initial setting in this iterative approach is: $trust_0(u) = d(u), \forall u$, where $d(u)$ represents the trust values of the seed users. The number of iterations is set to $Q = 100$.

The approach of Krestel and Chen is more sophisticated than the approach of Xu *et al.* [135] in that multiple relationships, such as tag co-occurrence, content co-occurrence and tag-content co-occurrence, can be taken into account, rather than considering only the tag-content pairs shared by users.

4.4 System for Automated Landmark Tagging

Based on the user reliability trust modeling described in Section 4.3.1, we build the solution for geotag propagation between images. The main innovation of such system is the combination of object duplicate detection and user trust modeling for accurate and reliable geotag propagation. The system architecture is illustrated in Figure 4.2. It contains three functional modules, each of which has a specific task: object duplicate detection, tag propagation, and user trust modeling. The user trust modeling module is powered by alternating the five trust modeling approaches described in Section 4.3, and then the performance of the trust module and the whole system is measured, and results are reported and analyzed in Section 4.6. As the focus of this chapter is on trust modeling and its combination with object duplicate detection, the object duplicate detection [85] module is only summarized briefly below, while tag propagation and user trust modeling are described in details in the next sections.

The system takes a small set of training images with associated geotags to create the corresponding object (landmark) models. These object models are used to detect objects duplicated in a set of untagged images. As a result, matching scores between the models and the images are obtained. According to the scores, the tag propagation module makes decisions about which geotags should be propagated to the individual images. Given a user trust model which describes the tagging reliability of each user, only the tags from the users who are trusted are propagated to the photos in the database.

4.4.1 Object Duplicate Detection

The goal of the object duplicate detection module is to detect the presence of a target object in an image based on an object model created from training images. Duplicate objects may vary from their perspective, have different size, or be modified versions of the original objects after minor manipulations, as long as such manipulations do not change their identity. This is especially true for images related to travel, where tourists tend to take a lot of photos from different distances

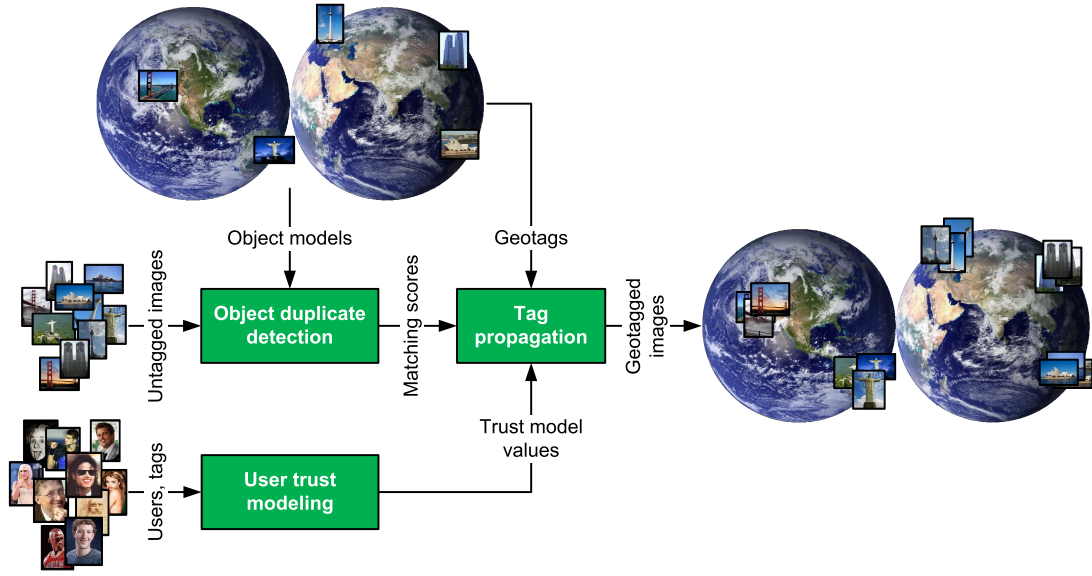


Figure 4.2: Overview of the system for geotag propagation in images. The object duplicate detection is trained with a small set of images with associated geotags. The created object (landmark) models are matched against untagged images. The resulting matching scores serve as an input to the tag propagation module, which propagates the corresponding tags to the untagged images. Given a user trust model, only the tags from reliable users are propagated.

and viewpoints around a famous landmark. The basic idea of applying object duplicate detection for geotag propagation is that travel images typically depict distinctive landmarks (e.g., buildings, mountains, bridges), which can be considered as object duplicates.

Training is performed as follows: given a set of images, features are extracted and a spatial graph model describing the object, i.e. landmark, is created for each of the landmarks. In our case, one training image per landmark is used to create a graph model. First, regions of interest (ROIs) in an image are extracted using the Hessian affine detector [81] and each of these regions is described using SIFT features [80]. These features are robust to arbitrary changes in viewpoints. Then, hierarchical k-means clustering [82] is applied to the features, to group them based on their similarity, as described in more details in Section 2.3.1.2. The result of the hierarchical clustering is used for the fast approximation of the nearest neighbor search, to efficiently resolve feature matching in the test phase. Finally, a spatial graph model is constructed to improve the accuracy of the feature matching with a test image. The graph model considers the scale, orientation, position, and neighborhood of features. The nodes of the graph are the features of the training images. The edges of the graph connect features with their spatial nearest neighbors. The attributes of edges are the distance and orientation of the neighbors. These attributes are important for the matching step in the test phase.

To detect the presence of the landmark within a test image, the features are extracted from the image in the same way described above. These features are matched to those in the graph model

derived from the training images. Feature matching is performed using a one-to-one nearest neighbor matching, where the hierarchical clustering is used to efficiently resolve the nearest neighbor search. Considering only matched features and their positions, a spatial graph model of the query image is constructed in the same way described in the training phase. Then, graph matching is applied between two graph models to identify the local correspondences between regions in the training and the test image. Finally, for the global object matching and matching score computation, the generalised Hough transform [87] is applied on the nodes of the matched graph. The matching scores represent the pair-wise comparison of training and test images.

More details about the object duplicate detection approach are presented in [85, 64].

4.4.2 Tag Propagation

The goal of the tag propagation module is to propagate the geotags from the tagged to the untagged images according to the matching scores, provided by the object duplicate detection module. As a result, labels from the training set are propagated to the same object found in the test set.

The geographical metadata (geotags) embedded in the image file usually consist of location names and/or GPS coordinates, but may also include altitude, viewpoint, etc. Two of the most commonly used metadata formats for image files are EXIF and IPTC. In this chapter, we consider the existing IPTC schema and introduce a hierarchical order for a subset of the available geotags, namely: city (name of the city where image was taken) and sublocation (area or name of the landmark), for example, Paris (Eiffel Tower) and Budapest (Parliament).

The tag propagation module supports two application scenarios: closed and open set problem, as shown in Figure 4.3. In the closed set problem, each test image is assumed to correspond to exactly one of the known (trained) landmarks. Therefore, the test image gets assigned to the most probable trained landmark, based on the matching scores provided by the object duplicate detection module, and the corresponding tag is propagated to the test image. This is comparable to an identification task in biometrics. However, in the open set problem the test image may correspond to an unknown landmark. This problem is comparable to a watchlist task in biometrics where the goal is to distinguish between known and unknown persons (landmarks) and to propagate the tags only for the known ones. For example, in Figure 4.3 we assume that the system is trained with three known landmarks: Budapest (Parliament), Belgrade (Church St. Sava) and Tokyo (Tower). Given the input test image of Paris (Eiffel Tower), our system gives different results for the closed and open set problems. In case of the closed set problem, our system finds that Tokyo (Tower) is the most suitable model for the test image. If we consider the open set problem, the system does not retrieve any of the trained models since the matching scores between the object models and the test image do not exceed a predefined threshold. The open and closed set problems are separately evaluated in Section 4.5 as detection and recognition tasks, respectively. However, as the main focus of this chapter is on trust modeling, in the evaluation of

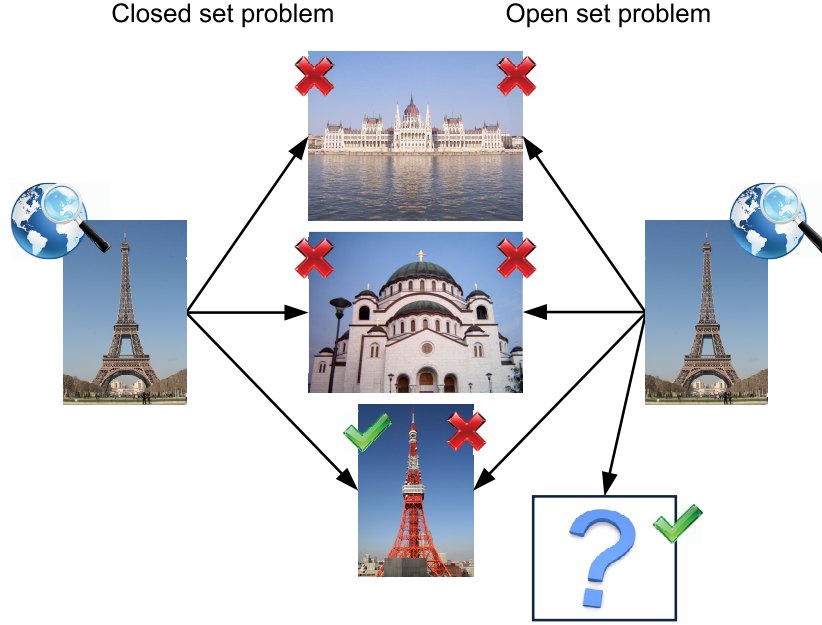


Figure 4.3: The closed and the open set problems. In the closed set problem, each test image is assumed to correspond to one of the known (trained) landmarks. However, in the open set problem the test picture may also correspond to an unknown landmark. In both the closed and the open set problems, the most suitable matches between the test image and trained landmarks are marked with green check mark, otherwise they are marked with red “x” mark.

trust models we will consider only closed set problem to create an environment for trust modeling where object duplicate detection performs well, as already explained in Section 4.1.

In a more detailed way, the tag propagation can be explained as follows. The object duplicate detection module provides a matching score matrix $S_{i,j}$. It represents the pair-wise comparison of the trained images (landmarks) $i, i \in [1 \dots M]$, and the test images $j, j \in [1 \dots N]$, where M and N are number of training and test images, respectively.

In the closed set problem, we find the maximum score for each test image j and propagate the geotag of the corresponding training image i . The assignment matrix $C_{i,j}, i \in [1 \dots M]$ and $j \in [1 \dots N]$, is formed in the following way:

$$C_{i,j} = \begin{cases} 1, & \text{if } S_{i,j} = \max_{i \in [1 \dots M]} \{S_{i,j}\}; \\ 0, & \text{otherwise.} \end{cases} \quad (4.15)$$

In this case, each test image gets assigned with exactly one tag from the training photo database.

In the open set problem, the tag propagation is only done if the corresponding score exceeds a

predefined threshold \hat{S} . The assignment matrix $O_{i,j}$, $i \in [1 \dots M]$ and $j \in [1 \dots N]$, is defined as:

$$O_{i,j} = \begin{cases} 1, & \text{if } S_{i,j} = \max_{i \in [1 \dots M]} \{S_{i,j}\} \wedge S_{i,j} \geq \hat{S}; \\ 0, & \text{otherwise.} \end{cases} \quad (4.16)$$

In this case, each test image can get assigned zero or one tag from the training set depending on the value of the threshold \hat{S} .

Based on the assignment matrix $C_{i,j}$ or $O_{i,j}$ the tags are propagated. If the corresponding value is 1, the tag associated with training image i is propagated to the test image j . If the corresponding value is 0, no tag is propagated.

4.4.3 User Trust Modeling

The goal of the user trust modeling module is to measure trust or reliability of the users in geotagging.

Different users are introduced into the system. They perform annotation of a small set of images through geotagging. The trust module collects all tags from the users or their feedback on tags, and aggregates them to produce trust values. For a given trust model described in Section 4.3, the user trust values $trust(u)$, $u \in [1 \dots L]$, are calculated. Only tags from users who are trusted are propagated to other photos in the dataset. In other words, if the user trust value $trust(u)$, exceeds a predefined threshold \hat{T} , then all his/her tags are propagated. Otherwise, none of his/her tags are propagated. Also, based on this user trust value, the user might be flagged as a legitimate user, if $trust(u) \geq \hat{T}$ or spammer, if $trust(u) < \hat{T}$.

As a final outcome of the whole system, only geotags coming from trusted users are propagated to the same landmark found in the non-tagged set of images.

4.5 Experiments

The effectiveness of the proposed system for automatic geotagging based on user trust modeling is demonstrated through a set of experiments on an image database containing various landmarks. The performance of the proposed geotag propagation method is evaluated and analyzed in two application scenarios, namely, with and without including users and their mistakes in the annotation process. Furthermore, several techniques for trust modeling, described in Section 4.3, are compared and contrasted to each other in detail.



Figure 4.4: Some example images from the database of famous landmarks used for the performance evaluation of the system for automated landmark tagging. More sample images of this database are provided in Appendix A.5.

4.5.1 Database

We created a database of images depicting geographically unique landmarks in order to evaluate the proposed geotag propagation method. This database is reported in [51] and described in details in Appendix A.5. It consists of 1320 images: 22 cities (such as Amsterdam, Barcelona, London, Moscow, Paris) and 3 landmarks for each of them (objects or areas in those cities, such as Bird's Nest Stadium, Sagrada Familia, Reichstag, Golden Gate Bridge, Eiffel Tower). Figure 4.5 summarizes all cities and landmarks contained in the database. Each landmark has 20 image samples, taken from variety of view points and distances. Some example images are shown in Figure 4.4.

The database is split into a training and a test set. Training images are chosen carefully so that they provide a wide angle view of those landmarks without other dominating objects. One image from each landmark is chosen as a training image, leading to 66 training images in total. All other images from the database are test images.

We manually formed a ground truth data by assigning several tags describing landmarks depicted in images.

In order to make our approach more computationally feasible, all images are downsized to a maximum size of 500×500 pixels and JPEG compressed before further processing.

4.5.2 Scenarios

The performance of the described geotag propagation method is evaluated and analyzed in two application scenarios: tag propagation and user trust scenario.

In the *tag propagation scenario*, we evaluate our automatic geotag propagation algorithm without including users and their mistakes in the annotation process. First, training images are selected for each landmark. Moreover, for each training image, negative and positive test pictures are selected. For each landmark, there are 19 positive images in the test set. Negative images are all images which do not contain the ground truth landmark, namely all images which depict one of the other 65 landmarks. This leads to $19 \times 65 = 1235$ negative images in the test set. In the evaluation of this scenario in Section 4.6.1, we consider both the open and closed set problems.

In the *user trust scenario*, we simulate a social network environment. As explained in Section 4.3.1, due to the lack of a suitable database, which provides user geotags and feedback from photo sharing website Panoramio, the evaluation of the user trust scenario in this chapter is based on the simulation of the social network environment. We recruited $L = 47$ participants, among whom 66 % were males and 34 % were females, aged 16–63 (average age was 29), with different backgrounds (architects, researchers, engineers, doctors, high school students) and cultural differences (from 8 different countries located mostly in Europe). Participants were asked to tag $M = 66$ photos from the database, putting the name of the landmark depicted in the image. We collected 3295 tags (658 of them were unique tags) and they were used to create different user trust models as per Section 4.3. For the model of Krestel and Chen [136], 12 users were manually selected as reliable in geotagging. They were researchers and architects who are assumed to have reliable knowledge of landmarks due to frequent travels and educational background. After having created the user trust values $trust(u)$, $u \in [1 \dots L]$, for all trust models described in Section 4.3, we perform tag propagation based on those annotated images. Selected trust models are then compared.

4.5.3 Evaluation

In this section, the evaluation methods for user driven tag propagation system are described. While the tag propagation scenario is evaluated as a closed and an open set problem, the user trust scenario is only evaluated as a closed set problem, as explained in Section 4.1.

An open set problem can be evaluated as a typical detection task, where an image has to be classified as known or unknown. Given the ground truth and the predicted labels, the numbers of true positives (TP), false positives (FP) and false negatives (FN) are computed. Precision-recall (PR) curves can be also derived, which plot the recall (R) versus the precision (P) with:

$$P = \frac{TP}{TP + FP} , \quad (4.17)$$

$$R = \frac{TP}{TP + FN} . \quad (4.18)$$

The F-measure is calculated to determine the optimum threshold (\hat{S} in Equation 4.16) for the object duplicate detection. It can be computed as the harmonic mean of the P and R values:

$$F = \frac{2 \cdot P \cdot R}{P + R} . \quad (4.19)$$

Thus it considers precision and recall equally weighted.

First, ground truth matrix (GT) is created for each test image j and of the corresponding training

image i :

$$GT_{i,j} = \begin{cases} 1, & \text{if } Landmark(i) = Landmark(j); \\ 0, & \text{otherwise.} \end{cases} \quad (4.20)$$

where $i \in [1 \dots M]$, $j \in [1 \dots N]$, M is the number of training images and N is the number of test images.

Then, given the assignment matrix $O_{i,j}$ defined in Section 4.4.2, TP , FP and FN can be calculated as:

$$TP = \sum_{i,j} GT_{i,j} \cdot O_{i,j}, \quad (4.21)$$

$$FP = \sum_{i,j} (1 - GT_{i,j}) \cdot O_{i,j}, \quad (4.22)$$

$$FN = \sum_{i,j} GT_{i,j} \cdot (1 - O_{i,j}). \quad (4.23)$$

A closed set problem can be evaluated using the recognition rate (accuracy, RR). It is defined as the ratio between the numbers of correctly suggested tags Tc and overall samples A :

$$RR = \frac{Tc}{A}. \quad (4.24)$$

First, ground truth matrix GT is calculated with Equation 4.20.

Second, for tag propagation scenario using object duplicate detection (ODD) method, Tc and A are calculated as:

$$Tc^{ODD} = \sum_{i,j} GT_{i,j} \cdot C_{i,j}, \quad (4.25)$$

$$A = \sum_{i,j} GT_{i,j}, \quad (4.26)$$

where $C_{i,j}$ assignment matrix was defined in Section 4.4.2.

Finally, for user trust scenario, which combines the tag propagation method and user trust modeling, Tc and A are defined as:

$$Tc^{Trust} = \sum_{i,j,u: trust(u) \geq \hat{T}} GT_{i,j} \cdot C_{i,j} \cdot U_{i,u}, \quad (4.27)$$

$$A = \left(\sum_{i,j} GT_{i,j} \right) \cdot \left(\sum_{u: trust^{model}(u) \geq \hat{T}} 1 \right), \quad (4.28)$$

where \hat{T} is the threshold for the user trust value, index $u \in [1 \dots L]$ and L is the number of users. In other words, A is the number of tags (one for each image) for each trusted user and Tc^{Trust} is the correctly propagated tags among those. A propagated tag is considered correct only if the annotated tag was the same as the ground truth tag.

4.6 Results

4.6.1 Results of the Tag Propagation Scenario

In this section, the results of the tag propagation scenario based on object duplicate detection [85] are discussed.

First, the closed set problem is evaluated as a recognition task. Considering only object duplicate detection in the tag propagation scenario without trust modeling, we compute the recognition rate for all landmarks in the database as shown in Figure 4.5. Each field in the figure is highlighted with the color ranging in the spectrum from red to blue, with red corresponding to the recognition rate closer to 1 and blue to the rate closer to 0. First column of the figure shows the average recognition rate for each city, sorted from highest to lowest values. In the database, we have three landmarks for each city as reflected in the three right columns of the figure.

The performance of object duplicate detection varies considerably for different cities, but also across the individual landmarks within a city. To investigate these variations, all landmarks, according to common visual properties, were divided into different groups such as castles, churches, bridges, towers/statues, stadiums, and ground structure. The further evaluation of the tag propagation scenario is based on these categories. Figure 4.6 shows that average recognition rates vary considerably for each of these categories. With the average rate of 71 % across all the landmarks, the object duplicate detection demonstrates the highest recognition rate for the castles' category and the lowest for the stadiums'. These results demonstrate that if we rely only on the object duplicate detection for tags propagation, on average, we can expect from the accuracy of propagation to be no more than 71 %. This expectation, however, is based on the assumption that all propagated tags are correct and reliable, with no spam tags or tags mistakenly assigned to images. Therefore, in the extended more practical scenario, where we assume some images to be tagged wrongly, the accuracy of the propagation may decrease dramatically. To minimize the impact from the spam or mistaken tags on tag propagation, in the next section, we consider different user trust models and compare their efficiency in filtering out the wrong tags.

Secondly, the open set problem is evaluated as a detection task through the PR curves shown in Figure 4.7. The PR curves show significant difference between the different groups of landmarks. The proposed tag propagation scenario performs well with castles or other buildings which have

Sydney	Harbour Bridge	Luna Park	Opera House
Oxford	Radcliffe	All Souls College	Ashmolean Museum
Budapest	Parliament	Buda Castle	Hero Square
Paris	Eiffel Tower	Louvre	Arc De Triomphe
Moscow	Christ Savior Cathedral	St. Basil	Kremlin
Delhi	Lotus Temple	Akshardham Temple	Humayun Tomb
Venice	Lion Statue	Campanile Di San Marco	St. Mark Bell Tower
Rome	Pantheon	St. Peter Basilica	Colosseum
London	Big Ben	Buckingham Palace	Tower Bridge
Berlin	TV Tower	Brandenburg Gate	Reichstag
Beijing	Temple Of Heaven	Birds Nest Stadium	Tiananmen
Barcelona	Sagrada Familia	Casa Mila	Olympic Communication Tower
Mexico City	Angel De La Independencia	Torre Latinoamericana	Palace Of Fine Arts
San Francisco	Coit Tower	Golden Gate Bridge	Twin Peaks
Amsterdam	Church Of St. Nicholas	Rijksmuseum	Royal Palace
Rio De Janeiro	Cristo Redentor	Paco Imperial	Maracana
Belgrade	Parliament	Winner Statue	St. Sava Church
Zurich	St. Peter	Fraumunster	Grossmunster
Tokyo	Tower	Metropolitan Government Center	National Museum
Istanbul	Blue Mosque	Hagia Sofia	Galata Tower
Lausanne	EPFL	Riponne	Cathedral
New York	Brooklyn Bridge	Statue Of Liberty	Twin Towers

Figure 4.5: The recognition rate for all landmarks. Each row represents one city from our database and the right three columns represent three landmarks in each city. The values in the first column are the average recognition rates for each city, sorted from highest to lowest.

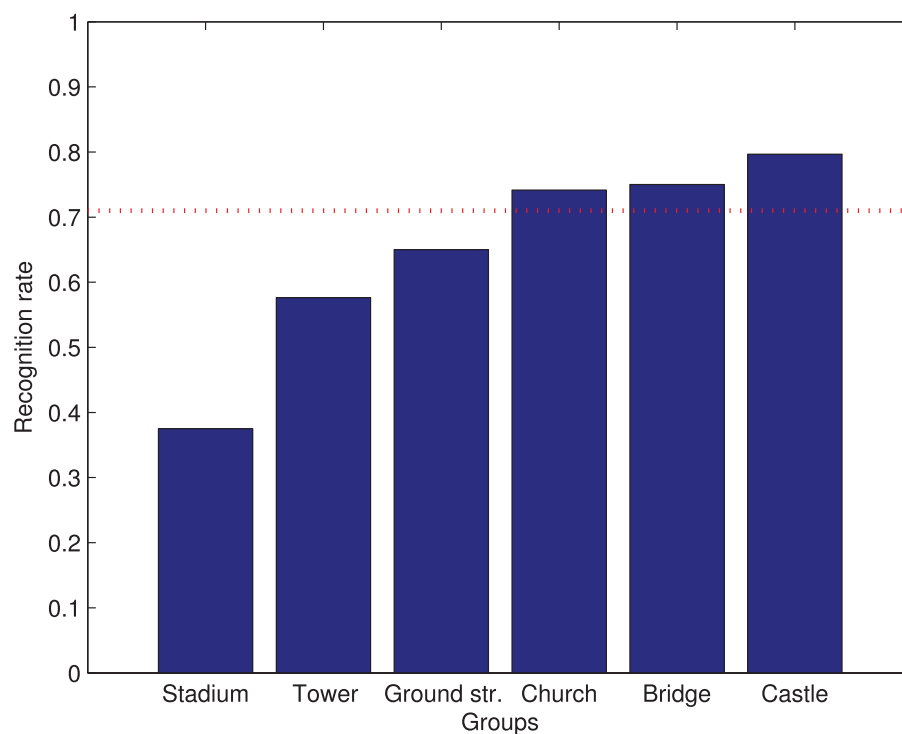


Figure 4.6: The recognition rate across the different landmark categories in the closed set problem (bars) and the recognition rate of all landmarks (dashed line). Landmarks have been grouped according to their visual characteristics. The average recognition rate across all the landmarks is 71 %.

more salient regions. In case of towers, it performs worse because the landmark does not have enough discriminative features. However, in case of stadiums, the performance is low due to the large variety of viewpoints.

The F-measures for the different detection thresholds \hat{S} are calculated to determine the optimal threshold value. Figure 4.8 shows the F-measures across the different groups. The F-measures for the different thresholds are calculated and the optimal threshold is chosen for the maximum F-measure and shown by green markers. The optimal threshold value does not vary much depending on landmarks (standard deviation of 13 %). The final F-measure for the open set problem averaged over the whole database is 73 %.

4.6.2 Results of the User Trust Scenario

In this section, comparison of different approaches in user trust scenario is discussed.

To compare different user trust models, we first analyze the distribution of their trust values given the manually assigned tags by the human participants (see Section 4.5.2 for more details). The values for each trust model were computed as described in Section 4.3. Obtained user trust values were normalized to 1 for each trust model. Then, the trust values were split into five equally distributed histogram bins with the following ranges: 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, and 0.8–1. Figure 4.9 shows the distribution of the total number of users with trust values in different bins for each of the trust model. From the results, it can be noted that the distributions for most of the user trust models are not normal (or Gaussian) distributions with mean value around 0.5. However, the tags to our dataset assigned by the human participants can be regarded following a normal distribution, assuming, participants unbiasedly tagged the depicted generally well-known landmarks. Therefore, useful, adequate, and practical user trust model should also reflect this distribution in the gathered tags from participants. From Figure 4.9, we can notice that only one out of five compared user trust models, Ivanov *et al.* [51], demonstrates the normal distribution around 0.5 in its assignment of the trust values to the participated users, while the rest of the models mark majority of the users as untrusted.

To understand the reasons for such bias in some of the user trust models, we plotted user trust values given by each approach to all users in Figure 4.10. The figure shows that trust values vary considerably between different users, but also across different models. For example, trust values of the users enumerated with 2, 16 and 47 span almost the entire range of the normalized trust value, namely, from 0 to 1, for all selected models. Since it is difficult to compare selected trust models for each user separately, we also grouped the participants according to their background. In our experiments, users are split into 6 different categories: researchers (13 users), architects (7 users), engineers (12 users), doctors (4 users), high school students (2 users), and others (9 users who did not indicate their background). By looking at the average trust levels for each group of users, as shown in Figure 4.11, we can observe that the trust values from Ivanov *et al.*, Koutrika *et al.* are higher for researchers and architects than for engineers, doctors, and high school students.

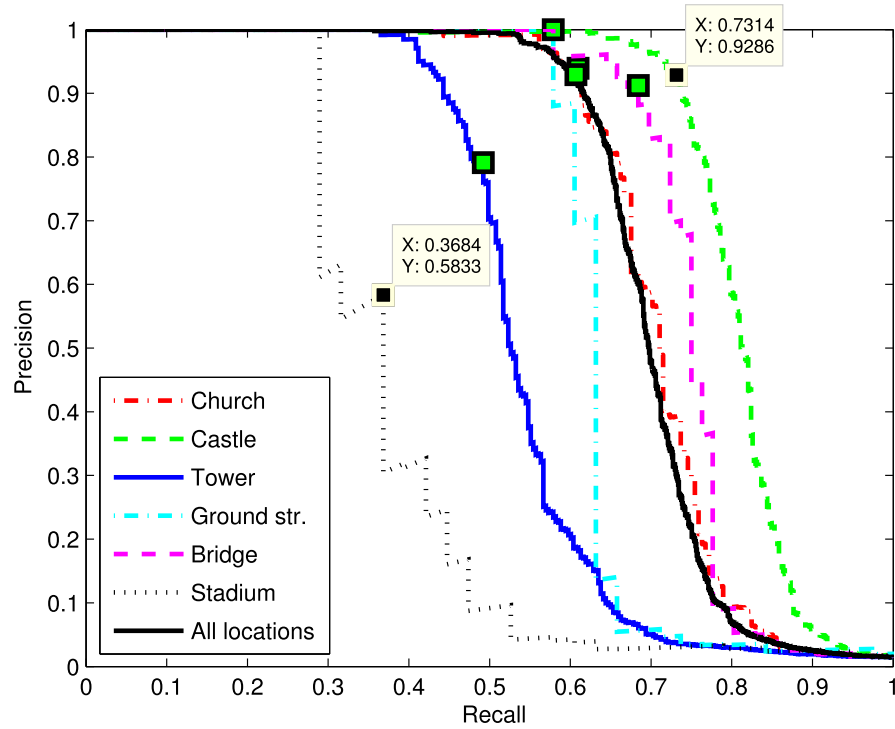


Figure 4.7: Precision versus recall curves for the open set problem across the different landmark groups. Markers show the optimal precision and recall, considering the F-measure.

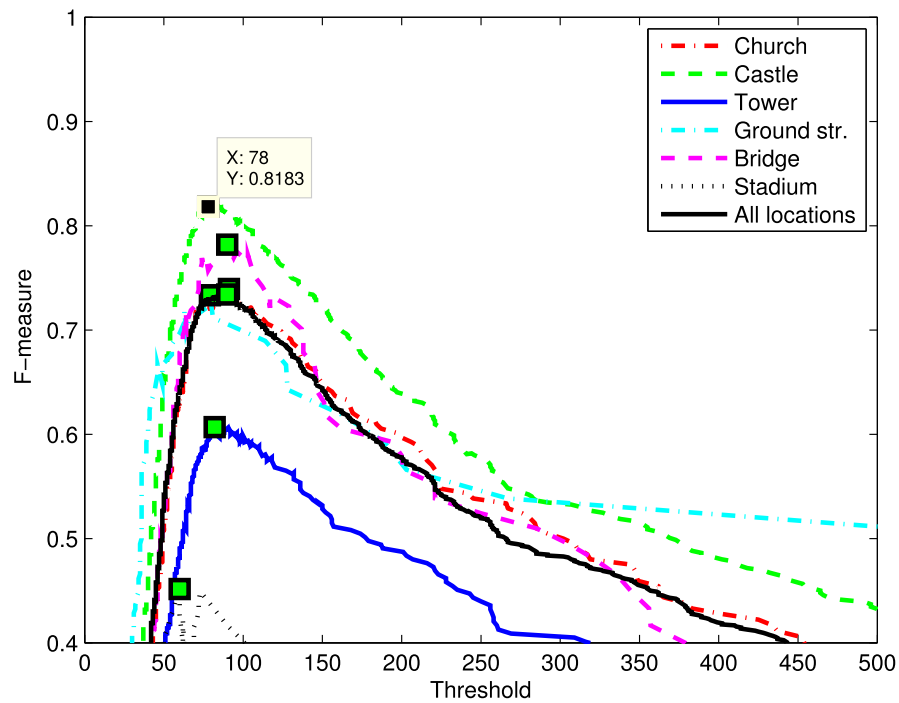


Figure 4.8: F-measure versus detection threshold \hat{S} across different landmark groups. Green markers show the optimal thresholds.

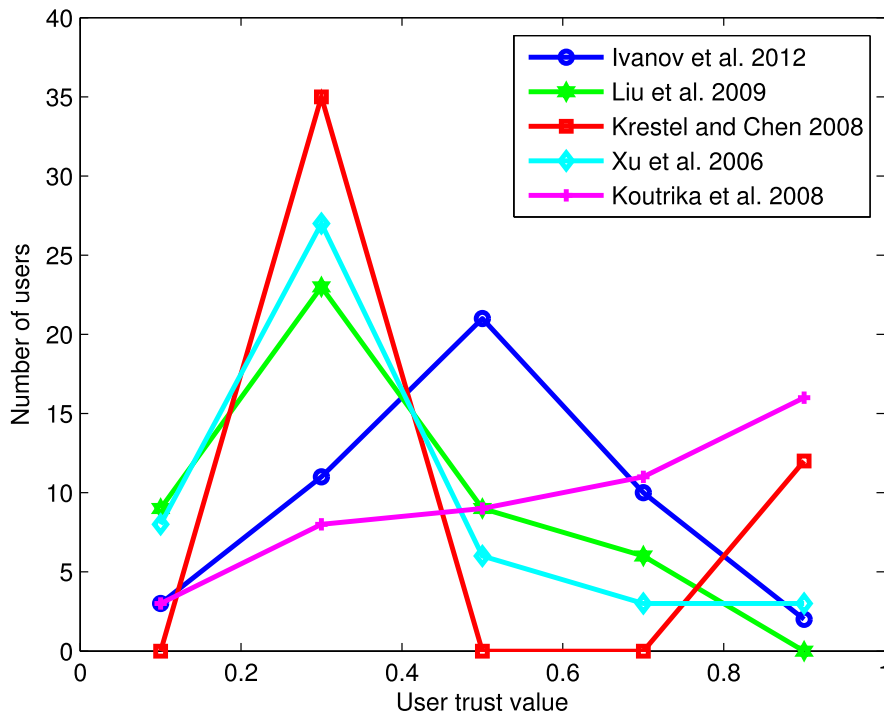


Figure 4.9: The distribution of the normalized trust values for different user trust models. Different user trust models are depicted with different line colors and different markers. The results show wide variety of distributions, mainly not normal (or Gaussian) distributions, which leads to a conclusion that users possess different knowledge in landmarks recognition and thus people are more or less reliable in geotagging.

Researchers and architects had larger number of geotags, since they travel often and, therefore, have a good knowledge of landmarks and their locations. High school students had considerably fewer geotags. Therefore, the trust values of researchers and architects produced by Ivanov *et al.* and Koutrika *et al.* are high, since these models give higher trust values to users who have more common tags with other users. Doctors and high school students might travel less, and thus their trust values are lower. On the other hand, models by Liu *et al.* [14] and Xu *et al.* [135] give larger trust values to the high school students. The reason is that these models assign higher trust weight to users who provide less tags in total but more of which are reliable, i.e., common with tags of the other users. One model that stands out of the rest is by Krestel and Chen [136]. Figure 4.10 demonstrates that one group of users have the highest trust value 1, while the rest of users are given value about 0.2. The users with trust value 1 are the 12 users that were manually chosen to be the trusted users to 'seed' the algorithm by Krestel and Chen. The reason for such significant disparity in trust values between the trusted users and others is that Krestel and Chen's model was designed for scenarios when most of the data is spam [136]. Their user trust model is very sensitive to the tags, which deviate from the tags by the original set of the trusted users. Therefore, this model demonstrates such significant bias when used with our set of geotags from the typical non-malicious participants.

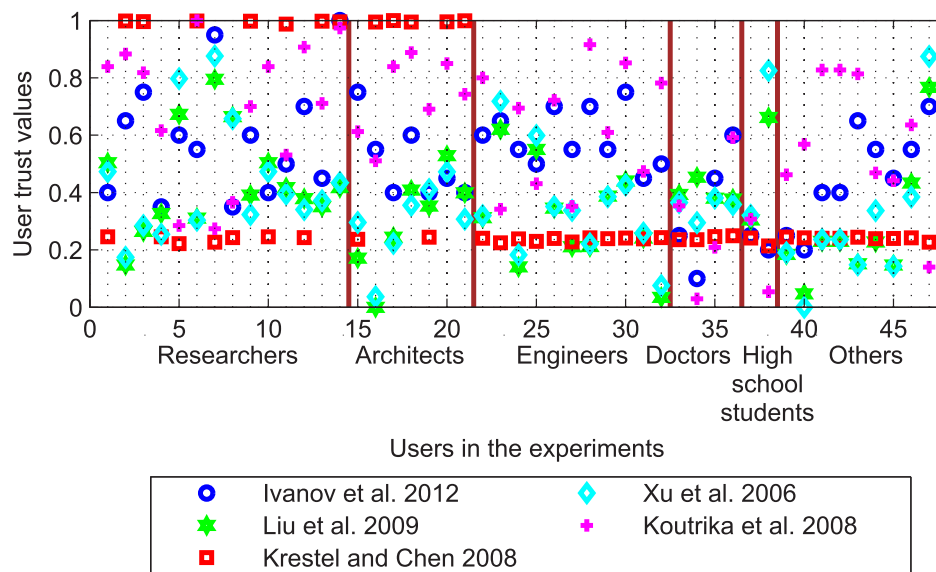


Figure 4.10: The distribution of the normalized trust values for all users by different trust models. Each user has five trust values given by each of the trust models. Different trust models are depicted with different marker colors. The results show that the trust values vary considerably between different users, but also across different approaches.

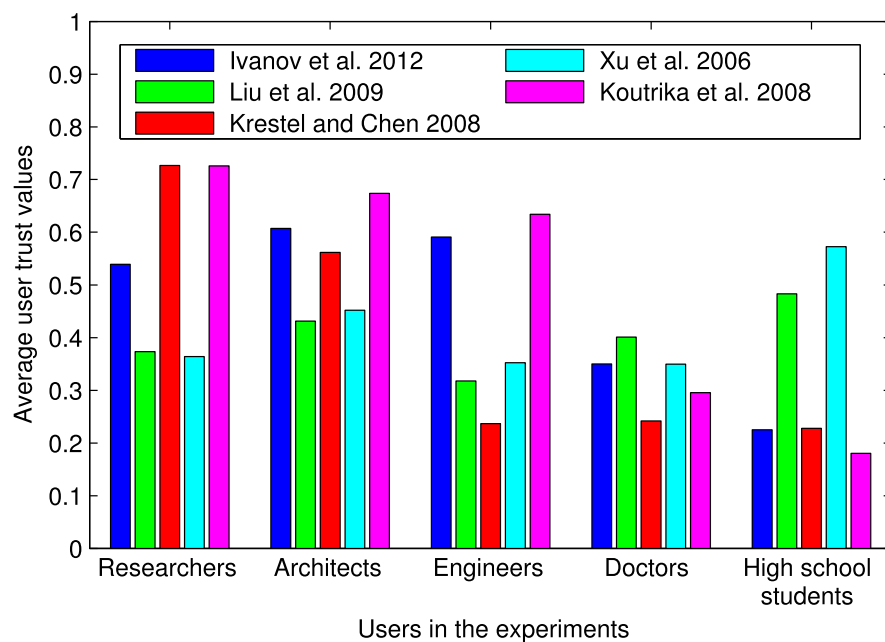


Figure 4.11: The average normalized trust values for different groups of users. The user category “others” is not shown, since the background of these users was not known. Different user trust models are depicted with different bar colors. Trust values mainly decrease from group of users who provided large number of correct geotags (researchers and architects) towards users who provided incorrect and low number of geotags (high school students).

To analyze the performance of the studied user trust models in tags propagation scenario, we plot the accuracy of propagation vs. the number of tags in Figure 4.13. We vary the number of propagated tags by adjusting the threshold on what is the acceptable user trust value of each model. The accuracy of the tags propagation is computed as the ratio of the correct (based on the ground truth) tags to the total number of tags assigned (propagated) to images. The maximum number of propagated tags can be much higher than the number of images, since several tags can be assigned to an image by different users. Each tag is propagated to different images. Therefore the curves in the Figure 4.13 show a trade-off between propagating tags to more images but less accurately and propagating tags to less images but more accurately. The black marker indicates the average tagging accuracy of the system when neither the user trust model nor automated tag propagation is used. In our experiments, it corresponds to users assigning $47 \times 66 = 3102$ tags to images (47 users in our experiments with each of them tagging 66 images at least once). The resulted average tagging accuracy is 52 %. This accuracy is equivalent to what currently Flickr or Panoramio have, where users simply tag photos independently with no propagation used.

Interestingly, by taking a closer look at the results in Figure 4.10, one can perceive that there exists a good correlation between the trust values for some of the approaches. The most straightforward way to compare the results obtained from the independent approaches is to analyze the scatter plots of trust values and to calculate the Pearson and Spearman correlation coefficients. The Pearson coefficient measures the distribution of the points around the linear trend, while the Spearman coefficient measures the monotonicity of the mapping, that is, how well an arbitrary monotonic function describes the relationship between two sets of data. Scatter plots of trust values for each pair of the considered trust modeling approaches are shown in Figure 4.12. The scatter plots and the correlation coefficients give an indication of the strong, positive correlation between the results of Liu *et al.* [14] and Xu *et al.* [135], with the Pearson (r) and Spearman (r_s) correlation coefficients of $r = 0.9517$ and $r_s = 0.9071$, which are statistically significant (the significance level is greater than 0.95). The other pairs of the trust values are less correlated. For example, there is a statistically significant positive correlation of $r = 0.3737$ and $r_s = 0.3368$ between approaches of Ivanov *et al.* [51] and Koutrika *et al.* [90]. The scatter plots show that the trust values of Ivanov *et al.* are usually shifted towards higher estimated trust levels, when compared to the values obtained by the approach of Koutrika *et al.*. On the other hand, the trust values of Xu *et al.* [135] and Koutrika *et al.* [90] show a negative correlation of $r = -0.4705$ and $r_s = -0.2818$, and are statistically significant. This negative correlation trend leads to the conclusion that higher trust values of Xu *et al.* correspond to the lower trust values of Koutrika *et al.*, and vice versa.

However, by using automated tag propagation that relies on the trust model based on user feedback, we can improve the accuracy of the tagging system and propagate more tags to the untagged images in the dataset. This improvement is illustrated by the left part of the blue curve (our method), which is above the average user trust value of 52 %. It means that more than 6600 tags (see Figure 4.13) can be propagated, twice more than without a trust model, from the trusted users, while keeping accuracy higher above 52 %. Other trust-based methods, such as by Koutrika *et al.*, also perform well, though, they show less impressive results than the tag

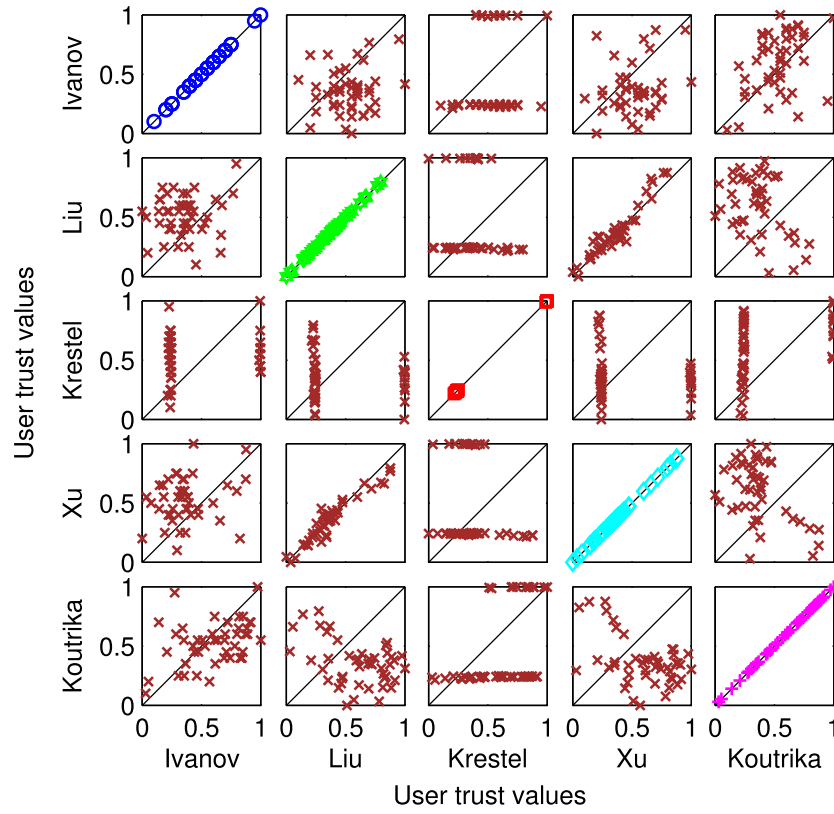


Figure 4.12: The scatter plots between the normalized trust values from different trust models. It can be noted that approaches of Liu *et al.* and Xu *et al.* have positive correlation between one another, while Xu *et al.* and Koutrika *et al.* show a negative correlation.

propagation based on our user trust model. However, the advantage of the algorithm by Koutrika *et al.* is that it is simple and does not need any ground truth or seed data. Methods of Liu *et al.*, Xu *et al.*, and Krestel and Chen are not able to perform well in the tag propagation scenario. Our method showed good performance in this simulated social network environment, since the algorithm includes users' tagging behavior through feedback from other people as an important factor in calculating trust value, rather than simply relying on the user contributed tags.

To further justify the usage of trust modeling in the automatic landmark tagging system, we measure the accuracy of propagation and the percentage of the number of propagated tags versus the threshold set for the user trust modeling. The results are shown in Figure 4.14 for the socially-driven approach by Ivanov *et al.* The optimal accuracy using object duplicate detection for geotag propagation is 71%. However, in this scenario the error of the user tagging step leads to a decrease of the performance. This error is caused by wrong tags given by the users. The optimal results can be reached if we set the threshold \hat{T} to a high value, but then the number of propagated tags becomes very low. On the other hand, when the threshold is low, more tags are propagated. These curves could be used to determine an appropriate threshold for the user trust model. The higher the threshold for the user trust model is, the more reliable the geotag

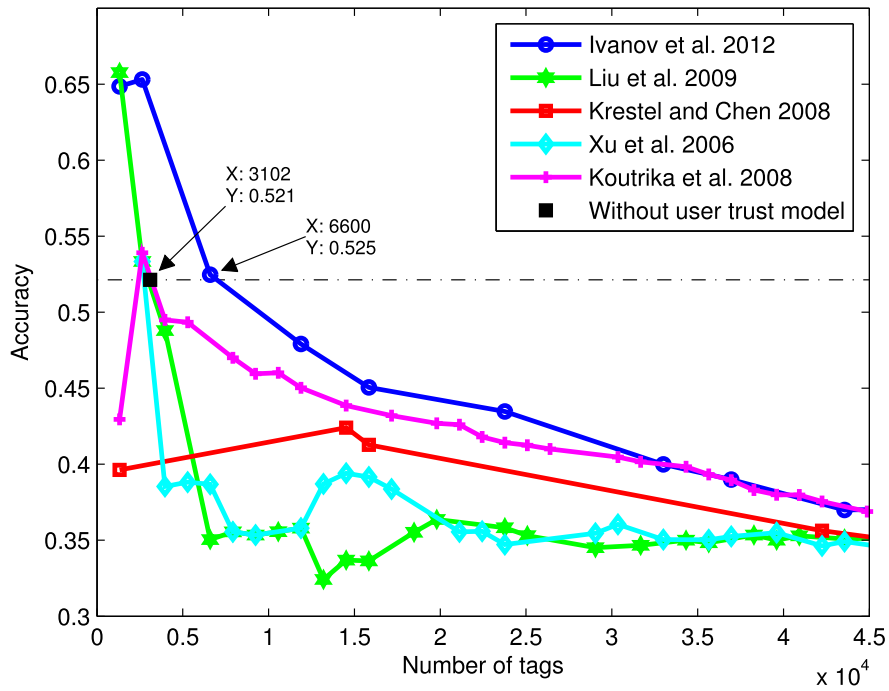


Figure 4.13: The recognition rate of the geotag propagation system versus the number of the propagated tags. Different user trust models are depicted with different line colors and markers.

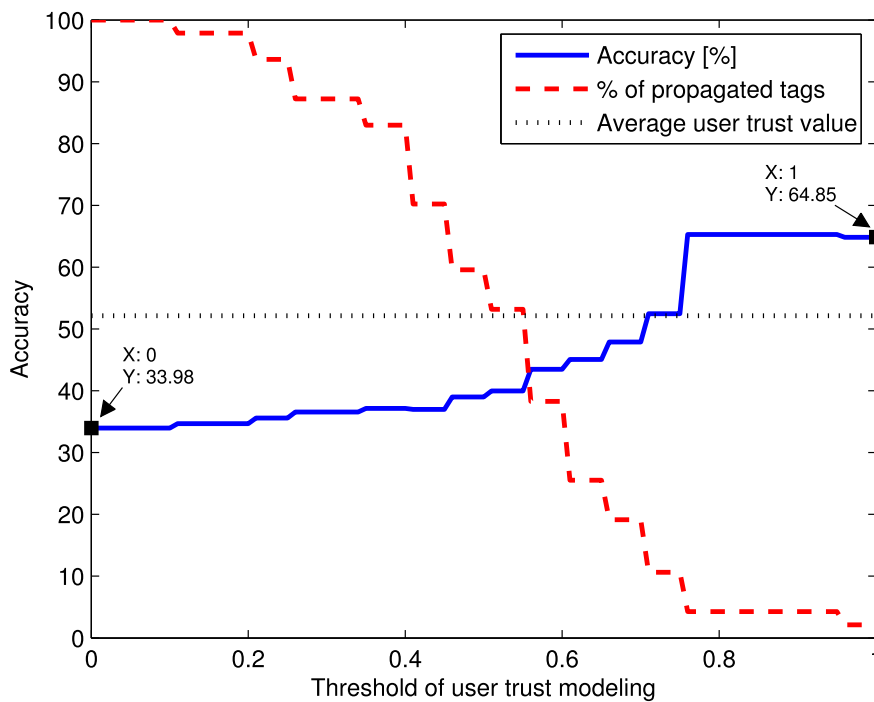


Figure 4.14: The recognition rate of the geotag propagation system and the percentage of the propagated tags versus the threshold \hat{T} for the user trust modeling in the socially-driven approach by Ivanov *et al.* [51].

propagation system is. At a threshold of 0, the accuracy of the system is equal to that without a user trust model, since all the user tags are propagated. In this case the accuracy of the system is 34 %. This additionally confirms that the accuracy of propagation can be significantly improved by including trust modeling in the automatic landmark tagging system.

4.7 Conclusion

In this chapter, we have presented different techniques for user trust modeling that are suitable for geotagging and can be used in geotag propagation systems. The problem of having trustworthy geotags of the content is important in social networks, because of their increasing popularity as means of sharing interests and information. Especially photo sharing and tagging is becoming more and more popular. Among other tags, geotags in form of geographical locations provide efficient information for grouping or retrieving images. Since manual annotation of these tags is time-consuming, automatic tag propagation based on visual similarity offers a very interestingly good solution.

The particular focus of this chapter is on the system for automatic geotag propagation by associating locations with distinctive landmarks and using object duplicate detection for tag propagation. The adopted graph-based approach reliably establishes the correspondence between a small set of tagged images and a large set of untagged images. Based on these correspondences and a trust value of the model derived for each user, only reliable geotags are propagated, which leads to a decrease of tagging efforts. We have analyzed the performance of the tag propagation alone which leads to a promising average accuracy of 71 % over all the landmarks. We have also shown that the performance varies considerably among different landmark types depending on their visual characteristics. We have analyzed the influence of wrongly annotated tags, which causes even more wrongly propagated tags in the database. Several trust models were evaluated and compared. The results show that by propagating tags based on the trust modeling relying on users' tagging behavior, the larger number of tags (more than twice) can be propagated with the same accuracy compared to using other trust models that simply rely on the user contributed tags or if using no trust modeling at all.

Since this type of comparative study is a pioneering work, a future study may consider a more careful selection of participants, for example, equal distribution of participated users in terms of group sizes and background. In this chapter, we compared trust modelling for automatic tagging considering closed set problem, as we could precisely measure number of tags in the system. However, we expect that the open set problem would also work fairly good, granted that we have a "good" thresholding method for object duplicate detection step. Most of the current techniques for noise and spam reduction focus only on textual tag processing and user profile analysis, while visual features of multimedia content can also provide useful information about the relevance of the content and content-tag relationship. In the future, a promising research direction would be to combine multimedia content analysis with conventional tag processing and user profile analysis. The challenge addressed in this chapter has been extended towards applying machine learning

Chapter 4. User Trust Modeling for Automatic Landmark Tagging

approach to facilitate the process of identifying legitimate users and spammers in a social tagging system, as presented in Chapter 5.

5 Fighting Spammers in Social Tagging Systems

Tagging in online social networks is very popular these days, as it facilitates search and retrieval of diverse resources available online. However, noisy and spam annotations often make it difficult to perform an efficient search. Users may make mistakes in tagging and irrelevant tags and resources may be maliciously added for advertisement or self-promotion. Since filtering spam annotations and spammers is time-consuming if it is done manually, machine learning approaches can be employed to facilitate this process. In this chapter, we propose and analyze a set of distinct features based on user behavior in tagging and tags popularity to distinguish between legitimate users and spammers. The effectiveness of the proposed features is demonstrated through a set of experiments on a database of social bookmarks.

Portions of this chapter are published in:

I. Ivanov, P. Vajda, J. S. Lee, and T. Ebrahimi, “In tags we trust: Trust modeling in social tagging of multimedia content,” *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 98–107, 2012

S. Yazdani, I. Ivanov, M. Analoui, R. Berangi, and T. Ebrahimi, “Spam fighting in social tagging systems,” in *Proceedings of the International Conference on Social Informatics*, pp. 448–461, Dec. 2012

5.1 Introduction

Social systems (networks) allow users to store, share, search and consume content (resources) online. Tagging in social systems has become increasingly popular since the transition to Web 2.0 [11], as it simplifies and eases search and retrieval of information, and allows users to access these information globally while interact and collaborate with each other. Tags can be assigned to different types of resources, such as images, videos, publications and bookmarks, making it a valuable asset to search engines on the Internet and in social tagging systems, as we already discussed in Chapter 1.

A few challenges have been identified in research community as important in social tagging systems, namely tag recommendation, tag propagation and tag relevance. For example, tag recommendation approaches suggest appropriate tags to resources (e.g., videos) in order to make it easy for users to search and access information in social systems [46]. In order to speed up the time-consuming manual tagging process, tags can be automatically assigned to images by making use of tag propagation techniques based on the similarity between image content (e.g., famous landmarks) and its context (e.g., associated geotags), as we already discussed in Chapter 4 and reported in [50]. Since user-contributed tags are known to be uncontrolled, ambiguous and personalized, one of the fundamental issues in tagging is how to reliably determine the relevance of a tag with respect to the content it is describing [47]. The fact that tags are user-contributed enables spammers to pollute social systems with irrelevant or wrong information (spam) to mislead other users, and to damage the integrity and reliability of social systems. In general, spam on the Internet is created to trick search engines by giving the spam content higher rank in the search results for advertisement or self-promotion purposes. Various techniques have been proposed in the literature for combatting spam, for example, Google's PageRank [48] and TrustRank [49].

Tags play a vital role in social systems, since it is important that resources in these systems are assigned with relevant tags. Injection of irrelevant tags and inappropriate content in social systems can be performed mainly in two ways. First, spammers can use legitimate resources and assign irrelevant tags to them for the purpose of advertisement or self-promotion [147]. Second, spammers can use popular and high ranking tags to describe a spam resource and boost its rank [90]. Therefore, one of the most important issues in social tagging systems is to identify appropriate tags and at the same time filter or eliminate spam content or spammers. Figure 5.1 shows an example of spam content on a popular social tagging systems.

In this chapter, we propose a set of distinct features that can efficiently identify spam users in social tagging systems. The introduced features address various properties of social spam and users activities in the system, and provide a helpful signal to discriminate legitimate users from spammers. The effectiveness of the proposed features is demonstrated through a set of experiments on a database of social bookmarks.

The rest of the chapter is organized as follows. Section 5.2 reviews the most recent related

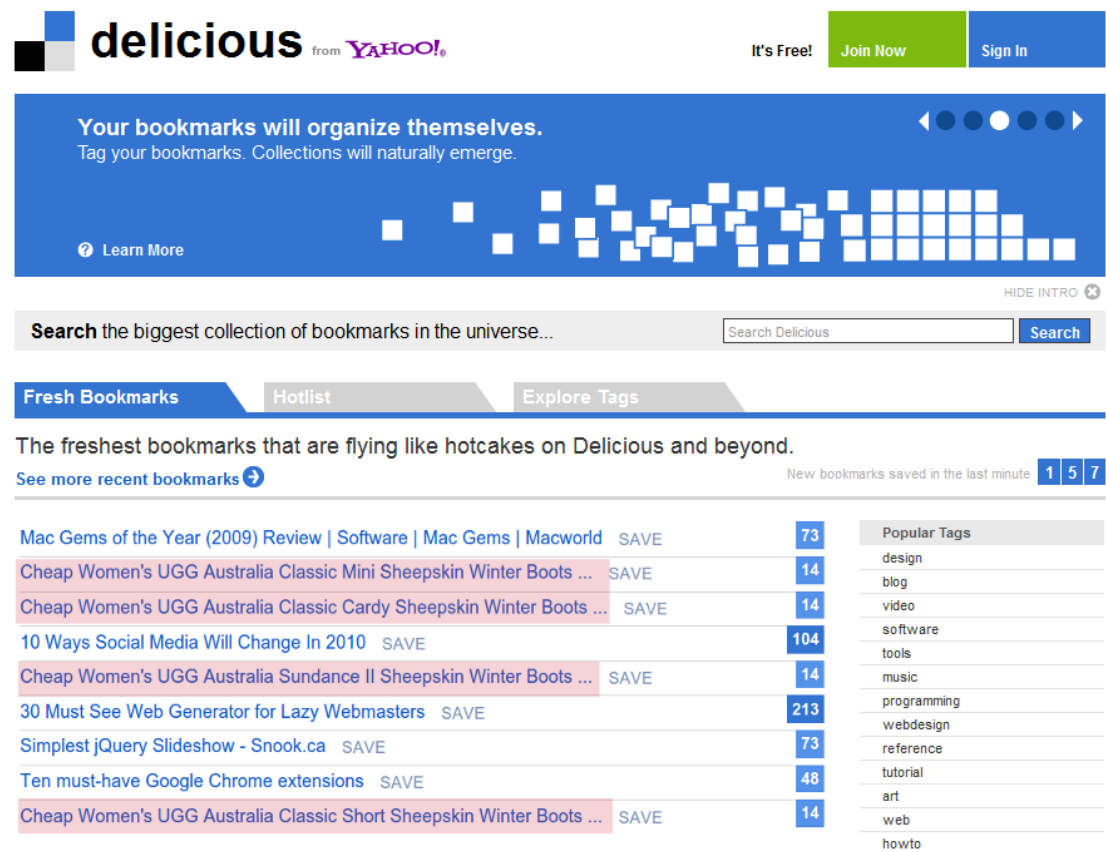


Figure 5.1: An example of spam content on a popular social tagging system: spam bookmarks in Delicious – all bookmarks are seeded by the same account and tagged by the same users (screenshot retrieved in December 2009).

work. In Section 5.3, we propose a set of distinct features for spammer detection based on user behavior in tagging and tags popularity. Evaluation methodology and database are presented in Section 5.4. In Section 5.5, we compare several supervised learning approaches applied to the proposed features and analyze their performance. Finally, Section 5.6 concludes the chapter with a summary and some perspectives for future work.

5.2 Related Work

Before social tagging became popular, spam content was observed in various domains: first in e-mail (e.g., [148]), and then in web search (e.g., [149]). Peer-to-peer (P2P) networks have been also influenced by malicious peers, and thus various solutions based on trust and reputation have been proposed, which dealt with collecting information on peer behavior, scoring and ranking peers, and responding based on the scores [150]. Nowadays, even blogs are spammed [151].

Ratings in online reputation systems, such as eBay⁵⁴, Amazon.com⁵⁵ and Epinions.com, are very similar to tagging systems and they may face the problem of unfair ratings by artificially inflating or deflating reputations [152]. Several filtering techniques for excluding unfair ratings are proposed in the literature (e.g., [153], [154]). Unfortunately, the countermeasures developed for the e-mail and web spam do not directly apply to social networks and photo sharing websites [155].

In order to reduce or eliminate spams in social networks, various anti-spam methods have been proposed in the state-of-the-art research. Heymann *et al.* [155] classified anti-spam strategies into three categories: prevention, detection, and demotion. *Prevention-based approaches* aim at making it difficult for spam content to contribute to social networks by restricting certain access types through interfaces (such as CAPTCHA [156] or reCAPTCHA [157]) or through usage limits (such as tagging quota, e.g., Flickr introduced a limit of 75 tags per photo [158]). *Detection approaches* identify likely spams either manually or automatically by making use of, for example, machine learning (such as text classification) or statistical analysis (such as link analysis), and then deleting the spam content or visibly marking it as hidden to users. Finally, *demotion-based approaches* reduce the prominence of content likely to be spam. For instance, rank-based methods produce ordering of a network's content, tags or users based on their trust scores. The prevention-based approaches can be considered as a type of precaution to prevent spammers. However, they cannot completely secure a social network. Some studies, e.g., [159], showed that CAPTCHA systems can be defeated by computers with around 90 % of accuracy, using, for example, optical character recognition or shape context matching. Even if prevention methods were perfect, there would be still possibility that the social networks get polluted with spam (malicious) or irrelevant tags. Therefore, detection and demotion techniques are required to keep a network free of noise and spam.

In a social network with tagging capability, spam or noise can be injected at three different levels: spam content (might be any piece of information – photos, videos, textual documents or web pages), spam tag-content association and spammer [160]. Spam fighting can be performed at each level separately (e.g., [160]) or different levels can be considered jointly to fight against noise and spam, for example, to assess a user's reliability, one can consider not only the user profile, but also the content that the user uploaded to a social network (e.g., [161]). Although BibSonomy is the most popularly explored domain for spam fighting, there are researchers who developed techniques for other social systems, such as Delicious, YouTube, MySpace or Twitter. We surveyed recent advances in techniques for combatting noise and spam in social tagging systems, classified the state-of-the-art approaches into a few categories and qualitatively compared and contrasted them. This work was published in [52].

Most of the techniques for combatting spam in social tagging systems use machine learning approaches applied to features specific to considered social network domains. Here we discuss some representative techniques.

⁵⁴ <http://www.ebay.com>

⁵⁵ <http://www.amazon.com>

Bogers *et al.* [147] proposed an approach to identify spammers in social bookmarking systems such as BibSonomy and CiteULike. The approach is based on user language models assuming that spammers and legitimate users use different language jargons when posting. To detect spam users, they learned a language model for each post, and then measured its similarity to the incoming posts by making use of Kullback-Leiber (KL) divergence. The spam status of a new post takes the status of the most similar language model. Status of a user is determined by grouping all users' posts. This approach was evaluated on BibSonomy database for spam detection, proposed at the "ECML PKDD Discovery Challenge 2008".

Krause *et al.* [161] employed a machine learning approach to detect spammers in BibSonomy. They investigated a framework for detecting spammers. The authors assumed that spammers usually use different strategies for polluting social bookmarking systems such as creating several accounts, publishing a particular post several times, and using semantically diverse tags to describe a bookmark and teaming up with other spammers to give good votes to each other. The authors investigated features considering information about a user's profile, location, bookmarking activity and semantics of tags. By making use of these features, and naïve Bayes, support vector machine (SVM) classifiers, logistic regression and J48 decision trees, they were able to distinguish legitimate users from malicious ones. This study represents a good foundation for future machine learning spam detecting approaches.

Markines *et al.* [160] proposed six different tag-, content- and user-based features for automatic detection of spammers in BibSonomy. The authors used features representing the probability of a tag being spam, number of advertises per post and number of valid resources per user posts. It was shown that "TagSpam" feature (tag diversity in posts) is the best predictor of spammers among all other features, because spammers tend to use certain "suspect" tags more than legitimate users. Although their work showed promising results, most of the proposed features rely on an infrastructure to enable access to the content, and must be recalculated periodically to remain reliable. Therefore, the feasibility of the proposed features depends on the circumstances of a particular social tagging system.

5.3 Distinct Features

In this section, we first recall a model of a social tagging system and then introduce a set of distinct features to distinguish between legitimate users and spammers in social systems.

Social tagging systems allow users to assign tags to resources shared online in order to enrich a resource with metadata and facilitate search for a particular resource, as previously explained in this chapter. The model of a social tagging system is introduced in Chapter 1 and is represented as a hyper-graph structure where the set of nodes consists of three kinds of objects: users, resources and tags, and hyper-edges connect these objects based on their relations [17]. The hyper-graph G can be defined with a quaternary structure $G = (U, T, R, P)$, where U represents the set of users u in the system, T is the set of tags t posted by users, R shows the set of resources r and

P defines the relation existing between tags, users, and resources. A relation linking a user, a tag and a resource represents a post. A post p in a social system can be represented with a triple $p = (u, r, T_u)$ which relates a user u who associated a resource r with a set of n tags $T_u = \{t_1, t_2, \dots, t_n\}$. We already presented an example of a social tagging system with 3 users, 4 tags and 3 resources in Figure 1.2.

Distinguishing between legitimate users and spammers in social tagging systems can be regarded as a classification problem. The most important part in any classification problem is the extraction of a good set of features from data. Features should represent data well to achieve good classification rate. Features are used to reduce the dimensionality of data while keeping important and relevant information. After studying the BibSonomy user behavior, we introduce 16 distinct features for each user from the evaluation database. Each user is represented with a feature vector consisting of 16 features which can be used by any known classifier to fight spam. In the following, we describe the proposed features in details, discuss the observation behind them and explain how to extract them out of a social tagging system.

5.3.1 LegitTags/SpamTags

We studied users' behavior in BibSonomy and found out that spammers and legitimate users tend to use different languages for their posts. Spammers often use a fraction of legitimate user vocabulary, mostly popular tags, to gain higher ranks. Apart from this fact, they have a very distinctive jargon which is barely used by legitimate users. Based on these observations, we propose two features: *LegitTags* and *SpamTags*.

LegitTags calculates the number of tags a user has posted which are mostly used by legitimate users. However, spammers also have habit to use popular tags that are previously posted by legitimate users. Therefore, we introduce a feature *LegitTags* which defines the probability that a particular tag is used only by legitimate users. Let U_t be the set of all users in a social tagging system who associated at least one resource with a tag t , T_u be the set of all tags posted by a user u , S_t be a subset of spammers in U_t and L_t be a subset of legitimate users in U_t . Then, the feature *LegitTags* for user u can be calculated as follows:

$$LegitTags_u = \frac{1}{|T_u|} \sum_{t \in T_u} \delta(u, t), \quad (5.1)$$

where $\delta(u, t)$ returns 1 if $|S_t|/|U_t|$ is less than a predefined threshold Th_{Legit} , otherwise it returns 0.

Analogously, a feature *SpamTags* is defined as:

$$SpamTags_u = \frac{1}{|T_u|} \sum_{t \in T_u} \sigma(u, t), \quad (5.2)$$

Chapter 5. Fighting Spammers in Social Tagging Systems

Table 5.1: Summary of tags popularity based features. All features are accumulated and averaged for each user separately.

Distinct feature	Description
<i>LegitPopularity</i>	Number of times tag t is assigned to posts by users in L_t
<i>SpamPopularity</i>	Number of times tag t is assigned to posts by users in S_t
<i>TagPopularity</i>	Number of times tag t is assigned to posts by users in U_t
<i>DistinctLegitPopularity</i>	Number of users in L_t who assign tag t to at least one resource
<i>DistinctSpamPopularity</i>	Number of users in S_t who assign tag t to at least one resource
<i>DistinctTagPopularity</i>	Number of users in U_t who assign tag t to at least one resource

where $\sigma(u, t)$ returns 1 if $|L_t|/|U_t|$ is less than a predefined threshold Th_{Spam} , otherwise it returns 0.

Optimal threshold values for Th_{Legit} and Th_{Spam} are experimentally found, and for our evaluation database they are set to 0.21 and 0.13, respectively.

5.3.2 Tags Popularity Based Features

One characteristic of spammers is that they tend to use popular tags when annotating online resources to gain higher rank in a search by keyword scenario [90], as already discussed in Sections 5.1 and 5.2. Based on this finding, we propose six features which address the popularity of tags shared in a social tagging system, namely, *LegitPopularity*, *SpamPopularity*, *TagPopularity*, *DistinctLegitPopularity*, *DistinctSpamPopularity* and *DistinctTagPopularity*.

For a particular tag t , we define a feature *LegitPopularity* as the number of times users in L_t used tag t in their posts. In an analogous way, features *SpamPopularity* and *TagPopularity* represent the number of times tag t was assigned to resources by users in S_t and U_t , respectively.

We propose three additional features representing tags popularity, namely *DistinctLegitPopularity*, *DistinctSpamPopularity* and *DistinctTagPopularity*. They represent the number of users in L_t , S_t and U_t who assigned tag t to at least one resource, respectively.

Each tags popularity based feature is accumulated for a particular user and averaged across all tags posted by the user to create a single feature value for every user.

5.3.3 User Activity Based Features

User activity based features take advantage of user's posting behavior in a social system to better discriminate between legitimate users and spammers. These features are explained in the following and summarized in Table 5.2. All features are computed for each user separately.

Feature *AverageTagsPerPost* shows the average number of tags a user assigned to different

Table 5.2: Summary of user activity based features. All features are computed for each user separately.

Distinct feature	Description
<i>AverageTagsPerPost</i>	Average number of tags a user assigned to different resources
<i>AverageDistinctTagsPerPost</i>	Average number of unique tags a user assigned to different resources
<i>NewTags</i>	Number of unprecedented tags a user added to the global dictionary of tags
<i>Legit2Spam</i>	Ratio between the number of legitimate and spam tags assigned by a user
<i>TagsPerUser</i>	Total number of tags a user assigned to different resources
<i>DistinctTagsPerUser</i>	Total number of unique tags a user assigned to different resources
<i>Posts</i>	Number of posts shared by a user
<i>DistinctTagRatio</i>	Ratio between number of unique tags and total number of tags assigned by a user

resources. The rationale behind this feature is that posts from legitimate users usually have more tags describing resources compared to posts shared by spam users. With the same rational, we introduce a feature *TagsPerUser*, defined as the total number of tags a user assigned to different resources.

Based on our observation that spammers tend to use different popular tags for different posts and, at the same time, the intersection between sets of tags in two arbitrary posts from one spammer is none or very small, we introduce a feature called *AverageDistinctTagsPerPost*. This feature measures the average number of unique tags a user assigned to different resources. With the same rational, we present two other features: *DistinctTagsPerUser*, defined as the total number of unique tags a user assigned to different resources, and *DistinctTagRatio*, which represents the ratio between number of unique tags and total number of tags assigned by a user.

Furthermore, number of new tags introduced by spammers to the global dictionary of tags is relatively higher than number of tags introduced by legitimate users. Based on this fact, we introduce a feature *NewTags*. This feature is defined as the number of unprecedented tags a user added to the global dictionary of tags.

We present here two other user activity based features. A feature *Legit2Spam* represents the ratio between the number of legitimate and spam tags assigned by a user, while a feature *Posts* is defined as the number of posts shared by a user.

Discussion on the performance of all proposed features on discriminating legitimate users from spammers is presented in Section 5.5.

Chapter 5. Fighting Spammers in Social Tagging Systems

Table 5.3: Statistics of the original database (“ECML PKDD Discovery Challenge 2008” bookmarks database) and a reduced database used for evaluation.

Statistics of databases	Original database			Evaluation database		
	Legitimate	Spam	Total	Legitimate	Spam	Total
Number of users	2467	29248	31715	500	500	1000
Number of resources	401250	2060707	2461957	172452	65378	237830
Number of tags	816197	13258759	14074956	477794	473544	951338
Average number of posts per user	162	70	77	344	131	238
Average number of tags per user	330	453	506	955	947	951
Average number of tags per post	2	7	6	3	8	4

5.4 Evaluation

In this section, we present a database and classification metrics used to evaluate the set of proposed features.

5.4.1 Database

We used database collected from BibSonomy. As already described in Chapter 1, BibSonomy is a social tagging system that allows users to share bookmarks and publication references. The system is aimed for researches and academic institutions which require a system without irrelevant information and commercial content. Therefore, this system has a rigorous policy against spammers. Moderators in this system manually find and remove spammers from the system [147]. If a user is labeled as a spammer, his/her posts will be no longer visible to other users. However, spammer posts will not be removed from the system and this fact gives an illusion to spammers that they are still able to pollute the system.

We used a public database released by BibSonomy as a part of the “ECML PKDD Discovery Challenge 2008” on spam detection in social bookmarking systems. More details about this database are provided in Appendix A.7. Table 5.3 summarizes statistics of the database. This database consists of around 32000 users who are manually labeled either as spammers or legitimate users, user-contributed tags and resources (bookmarks) which can be either web pages or BibTeX files. However, as shown in the second column of Table 5.3, an important skewness is present in this database since a majority of the users are spammers. This means that if a classifier

labels all users as spammers, we would achieve a classification accuracy of over 0.92. Therefore, we selected randomly a subset of users (500 legitimate users and 500 spammers) to achieve a balance with respect to the number of users. Statistics of the database used for evaluation in this chapter is shown in the third column of Table 5.3, denoted as Evaluation database.

5.4.2 Classification Metrics

After having extracted proposed features from the evaluation database, several supervised classification methods, such as SVM, AdaBoost and decision trees, were applied on the extracted features to classify users as legitimate or spammers. Given the ground truth and the predicted labels, a confusion matrix is created and the numbers of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are computed.

Different metrics are used to evaluate the proposed features. The accuracy of the classification when shown solely is not a good indicator of a classifier behavior, and therefore, we calculated some complementary measures to thoroughly evaluate the proposed features. In addition to the classification accuracy defined as:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}, \quad (5.3)$$

we also calculated:

- (1) false positive rate (FPR) as:

$$FPR = \frac{FP}{FP + TN}, \quad (5.4)$$

- (2) precision (P) as:

$$P = \frac{TP}{TP + FP}, \quad (5.5)$$

- (3) recall (R) as:

$$R = \frac{TP}{TP + FN}, \quad (5.6)$$

- (4) F-measure as:

$$F = \frac{2 \cdot P \cdot R}{P + R}, \quad (5.7)$$

- (5) area under receiver operating characteristics (AUC ROC) which represents the probability that an arbitrary legitimate user is ranked higher than an arbitrary spammer.

Finally, we determined Matthews correlation coefficient (MCC) [162] to validate our result. As a less known performance metric, we explain it here in more details. MCC is a performance quality measure used in two-class classification problems. It is often used as a performance metric in bioinformatics. MCC is defined as:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \quad (5.8)$$

MCC has values between -1 and $+1$, where $+1$ indicates perfect classification (prediction), -1 shows total disagreement between prediction and observation, and 0 represents a random classification.

5.5 Results and Analysis

In this section, we discuss the prominence of the proposed features for detection of spammers. First, performance of each feature separately is estimated and then some of them are aggregated to improve the classification performance. Finally, performance of different classifiers are compared and analyzed. All performance criteria were evaluated by making use of classifiers in Weka⁵⁶, a software library of most distinguished machine learning algorithms [163]. Evaluation is performed using 10-fold cross-validation and default values for all parameters in Weka.

Figure 5.2 shows how well each of the proposed 16 features discriminates spammers. A decision stump classifier in Weka is applied on extracted features and the performance of each proposed feature is measured as accuracy, AUC ROC and F-measure. As we can see from the accuracy metric, each feature is able to correctly classify at least 60 % of users. Feature *LegitTags* has the best performance with more than 0.91 of accuracy in classification, and it is followed by *SpamTags*, *DistinctLegitPopularity* and *Legit2Spam* with 0.87, 0.76, 0.73 of accuracy, respectively. For classification of randomly selected users, as it can be seen from AUC ROC, again *LegitTags* and *SpamTags* have the best performance with 0.96 and 0.93 of AUC ROC. F-measure follows the trend of accuracy and AUC ROC, showing that *LegitTags* and *SpamTags* are the adequate features. Having considered all these measures, we can conclude that after *LegitTags* and *SpamTags*, tags popularity based features are the best performing set of features.

Feature *LegitTags* has the ability to very well separate spammers from legitimate users when fed solely into the classifier, as discussed previously in this section. Therefore, we explore the performance of this feature in more details. 10-bins histogram of *LegitTags* values calculated from the evaluation database is shown in Figure 5.3 (a). When this feature is combined with the second best performing feature *SpamTags* and feature values are plotted in the feature space, we obtain the distribution shown in Figure 5.3 (b). These distributions give a visual intuition for how well feature *LegitTags* alone or combined with other feature separates two types of users. We can clearly see that the distributions of legitimate users and spammers can be easily separated by a

⁵⁶ <http://www.cs.waikato.ac.nz/ml/weka>

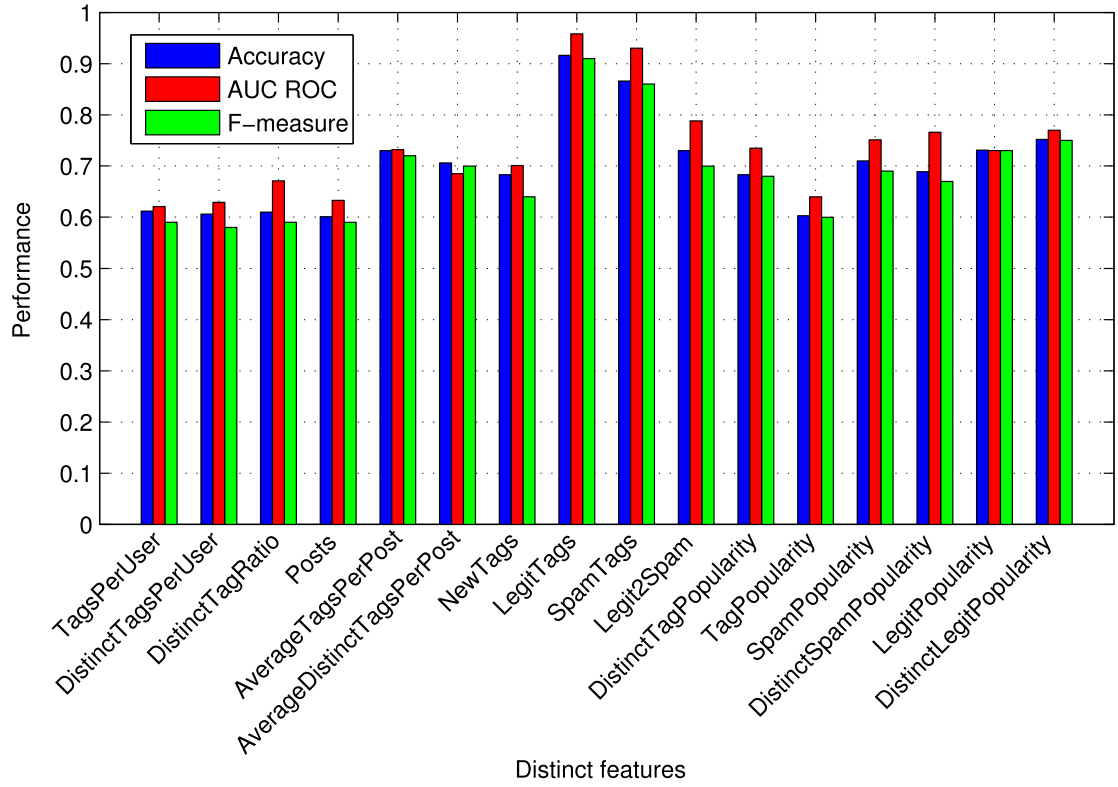


Figure 5.2: The performance of each proposed feature plotted as accuracy, AUC ROC and F-measure.

simple threshold, for case (a), or line, for case (b). Therefore, linear discrimination classifiers are enough for spammers detection when using *LegitTags* and *SpamTags* features.

After *LegitTags* and *SpamTags*, tags popularity based features are the most powerful set of features, as shown in Figure 5.2. To further evaluate these features, we applied a standard discrimination function, the χ^2 statistics. The χ^2 (chi-square) statistics measures the goodness and powerfulness of features used for classification [164]. Again, we used Weka to apply this discrimination function. Figure 5.4 shows the consistent ranking of our six tags popularity based features to discriminate spammers from legitimate users.

It is well known that classification accuracy can be significantly improved by aggregating weak features rather than feeding different features separately into a classifier [165]. We can see from Figure 5.2 that each tag popularity and user activity based features have less than 0.75 and 0.73 of accuracy, respectively. Nevertheless, combination of these features results in a performance improvement. Figures 5.5 (a) and (b) show how classification performance can be improved by separately aggregating tag popularity based features and user activity based features. Results are shown for two classifiers, namely AdaBoost and LibSVM. By combining all tag popularity based features we can improve classification accuracy from 0.75 to 0.91, while aggregating all user activity based features the accuracy increases from initial 0.73 to 0.86.

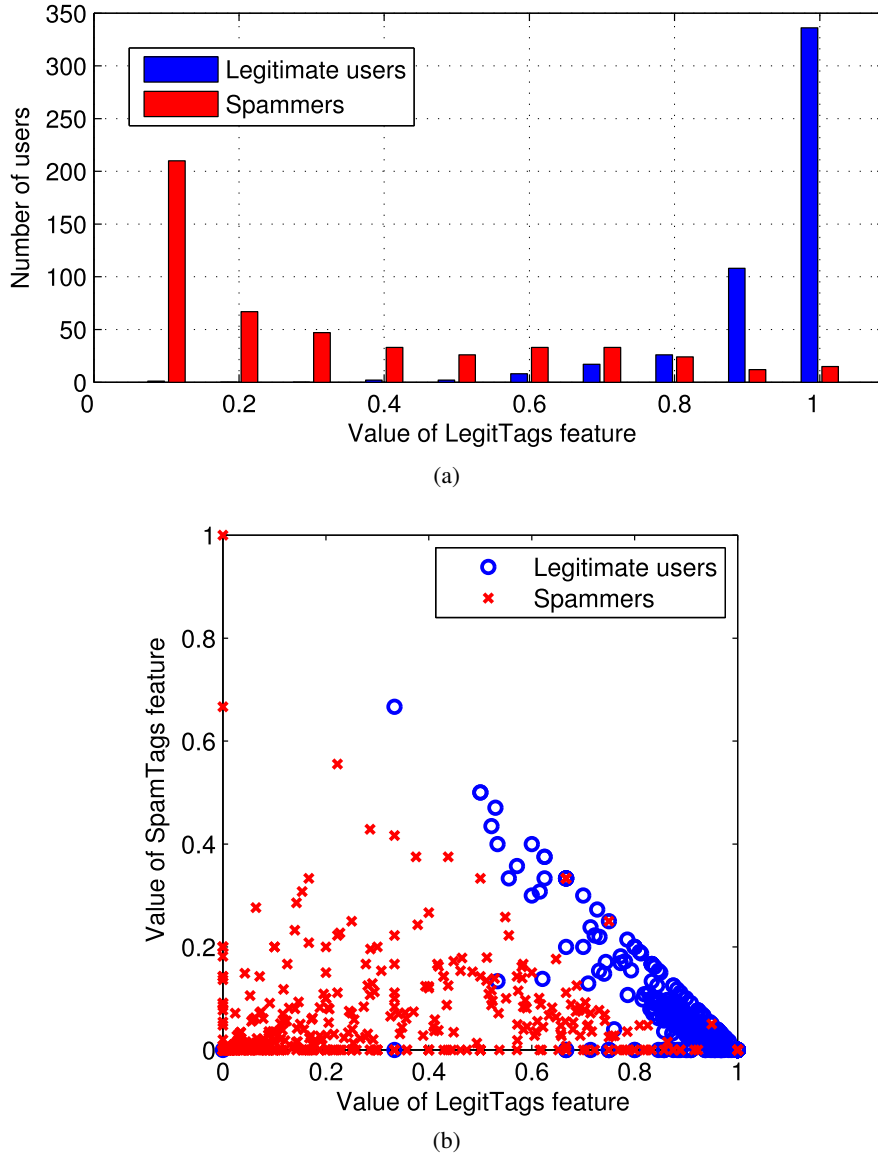


Figure 5.3: Discrimination power of the feature *LegitTags* to separate two types of users, when: (a) used alone, (b) combined with the feature *SpamTags*. Figure (a) represents the histogram of *LegitTags* values, and Figure (b) shows projection of *LegitTags* and *SpamTags* values in the feature space attempting to separate legitimate users (blue circles) from spammers (red crosses).

Finally, the proposed features are fed into more than 40 different classifiers and their performance in classification is evaluated. We used Weka to train classifiers with our features and to measure performance. Diverse classifiers are used, such as decision trees, neural networks and LibSVM, in order to have different perspectives on discriminative functions in feature space. Furthermore, ensemble classifiers [165] such as AdaBoost, bagging and rotation forest, were employed to have a comprehensive evaluation. The top 10 performing classifiers are reported in Table 5.4. Results show that AdaBoost was the best classifier for the evaluation database. It performs well

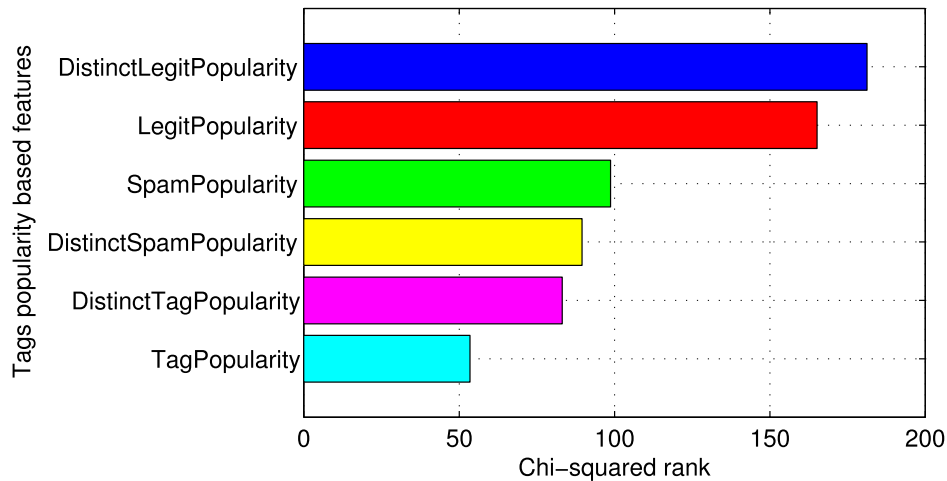


Figure 5.4: Chi-squared ranking for all tags popularity based features.

with 0.987 of accuracy and only 0.013 of *FPR*. LibSVM and rotation forest classifiers have slightly lower accuracy of 0.986 and 0.981, with 0.014 and 0.019 of *FPR*, respectively. As noted by Markines *et al.* [160], in a deployed social spam detection system it is more important that *FPR* is kept low compared to high accuracy, because misclassification of a legitimate user is a more consequential mistake than missing a spammer. Other researchers, who proposed different features from the whole or partial database of the “ECML PKDD Discovery Challenge 2008”, obtained similar results, for example, Markines *et al.* [160] were able to reach 0.979 of accuracy and 0.013 of *FPR*, while Bogers *et al.* [147] got 0.9799 of classification accuracy.

Table 5.4: Top classifiers created in Weka. Evaluation is performed using 10-fold cross-validation. The best performing classifier and metric values are highlighted in **bold**.

Weka classifier	Accuracy	FPR	R	P	F-measure	AUC ROC	MCC
AdaBoostM1	0.987	0.013	0.994	0.980	0.987	0.993	0.974
Libsvm	0.986	0.014	0.978	0.994	0.986	0.993	0.973
RotationForest	0.981	0.019	0.981	0.978	0.980	0.993	0.962
SMO	0.979	0.021	0.979	0.979	0.979	0.991	0.958
RBFNetwork	0.975	0.025	0.965	0.986	0.975	0.993	0.95
Bagging	0.974	0.026	0.974	0.974	0.974	0.996	0.948
Decorate	0.973	0.029	0.970	0.968	0.968	0.990	0.930
FT	0.972	0.028	0.966	0.972	0.970	0.985	0.944
MultiBoostAB	0.971	0.029	0.970	0.972	0.971	0.987	0.942
MLP	0.971	0.029	0.959	0.984	0.971	0.982	0.942

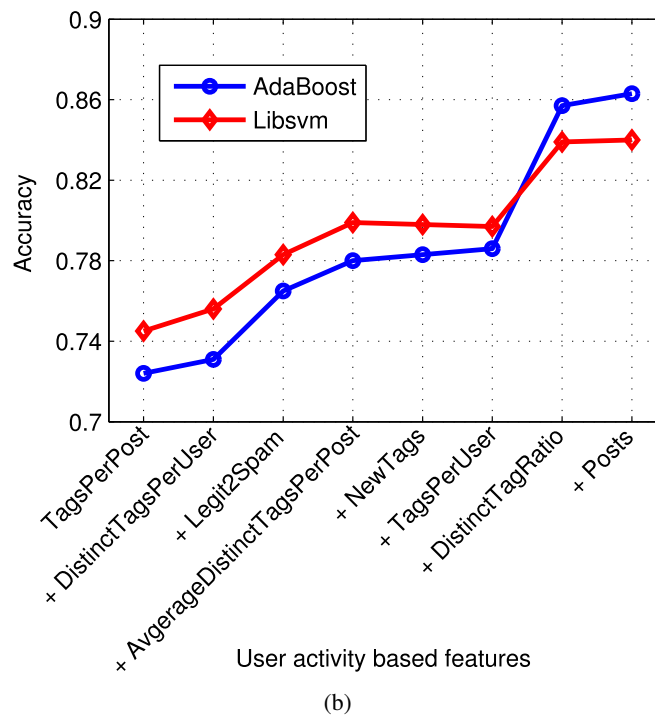
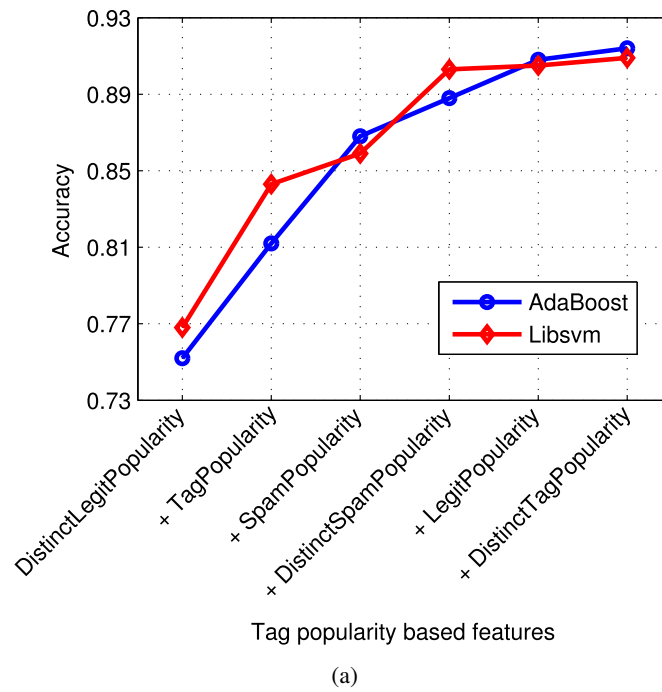


Figure 5.5: Enhancement in the classification performance by aggregating: (a) all tag popularity based features, and (b) all user activity based features.

5.6 Conclusion

In this chapter, we presented different features suitable for fighting spam in social tagging systems. The problem of having trustworthy tags associated to resources is important in social systems, because of their increasing popularity as means of sharing interests and information. Therefore, one of the most important issues in social tagging systems is to identify appropriate tags and at the same time filter or eliminate spam content or spammers.

We proposed 16 distinct features based on user activity in posting and tags popularity. The prominence of the proposed features in distinguishing between legitimate users and spammers is discussed. We measured the performance of each feature solely and showed that *LegitTags* feature, defined as the probability that a particular tag is used only by legitimate users, performed the best. We also showed that aggregation of features leads to the improvement in the classification performance. Finally, performance of different classifiers was compared. The results are promising. The best classifier achieved accuracy of 0.987 with false positive rate of 0.013 in discriminating legitimate users from spammers.

As a future study, one could explore more sophisticated features which are able to deal with dynamics of trust, by distinguishing between recent and old tags. Future work considering dynamics of trust would lead to better modeling of the phenomenon in real-world applications.

Content Enrichment Part III

6 “Epitome” – A Social Game for Photo Album Summarization

With the rapid growth of digital photography, sharing of photos with friends and family has become very popular. When people share their photos, they usually organize them into albums according to events or places. To tell the story of some important events in one’s life, it is desirable to have an efficient summarization tool which can help people to receive a quick overview of an album containing large number of photos. In this chapter, we present and analyze an approach for photo album summarization through a novel social game “Epitome” as a Facebook application. This social game can collect research data and, at the same time, it provides a collage or a cover photo of the user’s photo album, while the user enjoys playing the game. The proof of concept of the proposed method is demonstrated through a set of experiments on several photo albums. As a benchmark comparison to this game, we perform automatic visual analysis considering several state-of-the-art features. We also evaluate the usability of the game by making use of a questionnaire on several subjects who played “Epitome” game. Furthermore, we address privacy issues concerning shared photos in Facebook applications.

Portions of this chapter are published in:

I. Ivanov, P. Vajda, J.-S. Lee, and T. Ebrahimi, “Epitome – A social game for photo album summarization,” in *Proceedings of the ACM International Conference on Multimedia, Workshop on Connected Multimedia*, pp. 33–38, Oct. 2010

P. Vajda, I. Ivanov, L. Goldmann, and T. Ebrahimi, “Social game Epitome versus automatic visual analysis,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, Jul. 2011

P. Vajda, I. Ivanov, J.-S. Lee, and T. Ebrahimi, “Epitomize your photos,” *International Journal of Computer Games Technology*, vol. 2011, no. 706893, pp. 1–11, 2011

6.1 Introduction

Rapid growth of digital photography in recent years has increased the size of personal photo collections. People use their digital cameras or mobile phones equipped with cameras to take photos. Besides storing them on computer hard drives, they often share their digital photos with friends, family and colleagues through social networks. Facebook, Instagram, Flickr and Picasa are examples of such photo sharing web sites. Some people also print their photos on post cards, calendars or photo books, often to give them as presents or to create physical souvenirs. According to a recent study [166], 46 % of adults who use Internet, upload photos that they themselves have created to web sites to share with others online, while 37 % of adults print their photos as photo books for the purpose of showing pictures to others [167].

There is a saying: “A picture is worth a thousand words” [168]. Therefore, people like to use their photos to tell their own stories of some important events in their lives. One’s wedding, birth of a baby, vacation, birthday party or even a long lasting period – from the date of one’s birth till celebration of 18th birthday, are only a few examples of such events. One of the reasons why people share photos is to ask their friends to comment and tag photos.

Photos are often organized into albums (collections) based on places, events or dates, and people. Consumers tend to take several photos from one scene, hoping that one of them will be outstanding, and this leads to large number of similar photos. Therefore, it can be very time-consuming to go through all photos in one of these albums. Summarization is an effective way to help getting a quick overview of a set of photos. In this chapter, album summarization is defined as selecting a set of photos from a larger collection which best represents the visual information of the entire collection. Selected photos can be used to create a collage of a given album or a cover for an album, or to be included in a photo book. However, as already mentioned, manual photo album summarization can be very time-consuming.

Which photos are the most suitable to summarize a photo album? Creation of a photo summary is a very subjective task. There are different criteria upon which a human user would rate digital photos. The color, composition, content, lighting and sharpness of a photo, all contribute to viewer’s response to that photo [169]. These characteristics are used extensively by professionals on web sites, magazine covers and printed advertisements to draw attention, communicate a message and leave a lasting emotional impression. There is a gap between what people think the summary should look like and what we get with an automatic summarization. For example, funny photos are usually chosen within summarized photos, and they are not easy to detect using computer vision techniques. Therefore, including photos containing humans, such as one’s family or friends, in the process of album summarization is needed.

Besides spending a lot of time sharing and consuming content in online social networks, people also use online applications, especially social games. Players pour huge amounts of time and efforts into games. For example, a survey presented in [170] revealed that most players (95 %) play social games several times a week, with 64 % playing daily. The average game session lasts

more than half an hour (that is how long 61 % play), while 10 % may play more than three hours at a time. Work by von Ahn *et al.* [66] showed the tremendous power that networks of people possess to solve problems while playing social games. Therefore, the time and effort in playing a game can be utilized to address some issues in image processing community, i.e. users entertain themselves while playing an enjoyable game, with the added side-effect that they are doing useful work in the process, for example, summarizing one’s photo album. This is one of our motivations to develop a novel approach for photo album summarization through gaming.

In this chapter, we present and evaluate an approach for photo album summarization through a novel social game “Epitome”. It has been implemented as a Facebook application⁵⁷ and as an application for mobile phones on the Android OS platform. The main idea of this approach is to show a reduced set of photos from a Facebook album, ask users to play the game and then integrate results of several users in order to produce a summarization for the whole album. There are two games involved in this approach: “Select the Best!” and “Split it!”. In the first game, a user has to select the most representative photo of a reduced set of images from one Facebook album. The goal of the second game is to mimic separation of one album into different events or scenes, by splitting the reduced set of images into two distinct parts. The results achieved in the two games are compared with those of other users, and every user receives a score based on his/her performance. A sequence of photos which gets the largest number of users’ votes represents a summarization sequence of the album. The proof of concept of the proposed method is demonstrated through a set of experiments on several photo albums. We compare results obtained by this game with an automatic image selection, by making use of visual and temporal features. Furthermore, the usability of the game is evaluated by making use of a questionnaire (a user study) on several subjects who played the “Epitome” game. We also address privacy issues concerning shared photos in Facebook applications.

The chapter is organized as follows. We introduce related work in Section 6.2. The proposed social game application is presented in Section 6.3. Evaluation methodologies and results are discussed in Section 6.4. Finally, Section 6.5 concludes the chapter with a summary and some perspectives for future study.

6.2 Related Work

The proposed game is related to different research fields including visual analysis, automatic photo album summarization and gaming. Therefore, the goal of this section is to review the most relevant work in these fields.

⁵⁷ <http://apps.facebook.com/epitome>

6.2.1 Automatic Photo Album Summarization

State-of-the-art techniques for automatic photo album summarization are based on time separated events, spatial information using GPS coordinates and content-based image similarities. Harada *et al.* [171] developed an interface for automatic personal photo structuring, considering the time difference between two consecutive photos in order to determine different events. Naaman *et al.* [172] developed a system which automatically organizes digital photographs considering their geographic location or event-based description extracted from user tags. For photo collection clustering, combination of spatial, temporal and content-based similarity is then used. This clustering can be used for photo navigation for different categories, such as elevation, season, time of the day, location, weather status, temperature and time zone. Once photos are clustered, different page layouts are shown. Atkins [173] proposed a photo collection page layout generation method, considering hierarchical partition of the page, which provides explicit control over the aspect ratios and relative areas of photos. This approach attempts to maximize page coverage without having overlapping photos. Geigel and Loui [174] emphasized the aesthetic side of a page layout for image collections. They used a genetic algorithm to optimize aspects such as balance and symmetry for a good placement of images in the personalized album pages. Rabbath *et al.* [175] combined content analysis of text and images to automatically select photos that match a specific query (where, when, what, who) in the user's social network and intelligently arrange and compose them into a printable photo book. In general, however, automatic summarization has its limitations. There is usually a gap between what people think the summary should look like and what automatic summarization produces. A promising solution to narrow the gap is to incorporate human knowledge and preference into the summarization process.

Regarding content-based image similarity, various visual features have been used in automatic photo album summarization. The bag-of-words (BoW) model is based on the histogram of local features [176]. Zhang *et al.* [177] presented a comparative study on the performance of different local features on texture and object recognition tasks based on global histogram of features. BoW method gives a robust, simple, and efficient solution for measuring image similarity without considering the spatial information in images. The BoW mostly uses local feature descriptors, the scale-invariant feature transform (SIFT) [80], which is based on an approximation of the human visual perception. A faster version of the SIFT descriptor with comparable accuracy, called the speeded up robust feature (SURF), is proposed in [79]. Another popular feature is the histogram of oriented gradients (HOG) [178]. It is a grid-based histogram on gradient information of the image. This feature was first proposed for human detection, while the recent literature also considers it for general image retrieval. More details about mentioned features is given in Section 3.4.2. Another feature showing good performance in scene recognition, object detection and segmentation, is "tiny" image, a very low-resolution 32×32 color image, resized from the original image [179]. The use of this feature was motivated by psychophysical results showing the remarkable tolerance of the human visual system to degradations in image resolution.

6.2.2 Crowdsourcing Through Games

Ames and Naaman [65] showed that providing incentives to users in form of entertainment or rewards, e.g. games, can motivate them to tag photos in online and mobile environments. Gaming also provides a new way of motivating people by making the subjective data acquisition interesting and enjoyable. The most famous examples of these kind of games are the games with a purpose (GWAPs), such as the ESP game and the Peekaboom, developed for collecting information about image content. In the ESP game⁵⁸ [66], two players, who are not allowed to communicate with each other, are asked to enter a textual label which describes a shown image. The task of each user is to enter the same word as his/her partner in the shortest possible time. In the Peekaboom game [67], one player is given a word related to the shown image, and the aim is to communicate that word to the other player by revealing portions of the image, while the second player sees an empty black space in the beginning. This idea served as a basis for several other games available online⁵⁹ [180], such as video tagging, music description and tagging, tag description, object segmentation, visual preference and image similarities. Foldit [181] is a game that presents simplified three-dimensional protein chains to players, and provides a score according to the predicted quality of the folding done by the player. All actions by the player are performed in a three dimensional virtual world. It requires training to solve complex open protein puzzles which in turn requires a lot of commitment by the players.

Following the presented state-of-the-art techniques, a game-based approach for photo album summarization called “Epitome” is developed. The game provides a collage or a cover photo of the user’s photo album, while, at the same time, the user enjoys playing the game. It can also collect research data. In this way, both users and research community can benefit. Therefore, this concept is novel compared to previously mentioned games.

6.3 Algorithms for Photo Album Summarization

In this section two algorithms for photo album summarization are described. First, the proposed “Epitome” game is described, which takes advantage of the plenty of casual gamers to solve this complex problem of album summarization. Then, an automatic visual algorithm is presented as a comparison benchmark to the task.

6.3.1 Social Game “Epitome”

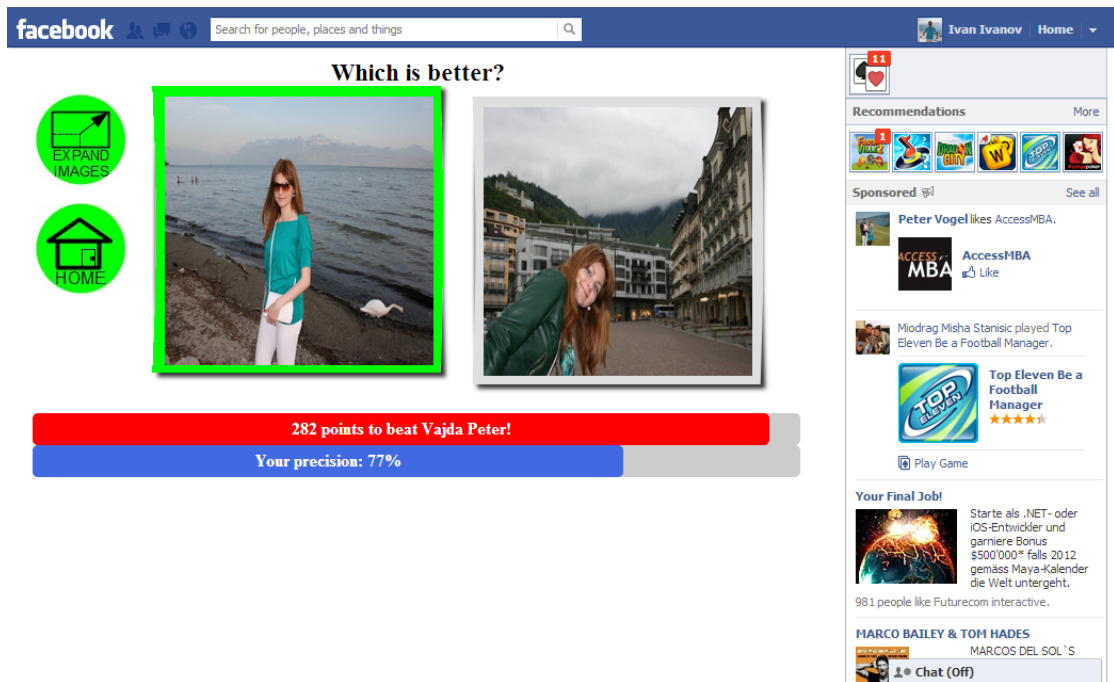
A social game “Epitome” provides an intuitive and enjoyable user interface as a Facebook application, as shown in Figure 6.1. The main purpose of the game is to create photo collages for Facebook photo albums considering the feedback of the owners’ Facebook friends.

The scenario of the game is as follows. A Facebook user, denoted as a player in this chapter,

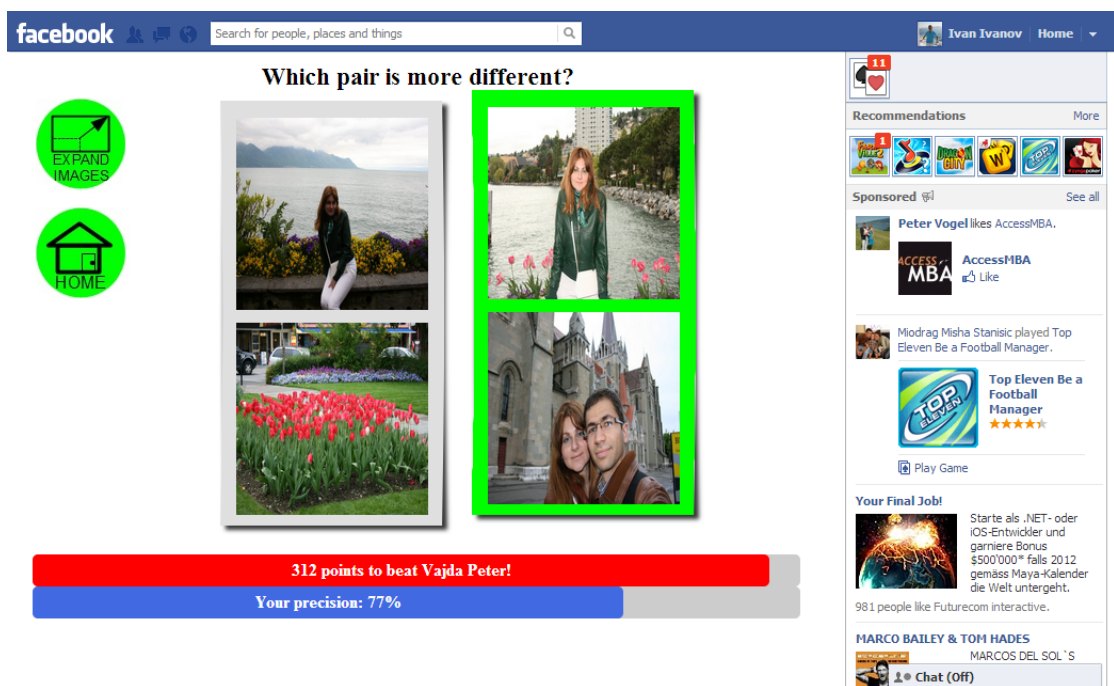
⁵⁸ <http://www.gwap.com/gwap/gamesPreview/espgame>

⁵⁹ <http://www.gwap.com/gwap>

6.3. Algorithms for Photo Album Summarization



(a)



(b)

Figure 6.1: Screenshots from “Epitome” game: (a) “Select the Best!” game, and (b) “Split it!” game (retrieved in January 2013).

installs the game in his/her Facebook applications page and allows access to his/her photo gallery, as shown in Figure 6.11 (a). Then, the player can select between two games. In both games, K consecutive photos, with respect to the time order by which photos were uploaded to Facebook, are randomly selected from one of the player’s friends Facebook albums. The number K is selected in such a way that these photos easily fit into one screen and they are usually placed in a grid. For example, Figure 6.1 shows an example where 2 or 4 (2×2) photos are shown to the player. In “Epitome” game, K was initially set to nine (3×3 photos) for both games, and after receiving users’ feedback this value was reduced to two and four, respectively for the first and the second game. The usability study based on users’ feedback is presented in Section 6.4.2. Without any loss of generality, we assume that the number K is the same for both games. In the first game, called “Select the Best!”, K images are shown to the player from one of his/her friends’ photo albums chosen randomly and he/she has to choose the best representative photo, which the player likes the most. By clicking on one of these K images, the player can see the expanded version of the image in a resolution that fits into the entire application screen. If the player chooses the photo which is the most frequently selected, then the player’s score increases. Since albums are chosen at random, it may happen that a new album without any subjective data appears in the game. In that case, the player automatically wins and his/her score increases. The second game is called “Split it!”. In this game, the player should split photos into two distinct parts. For example, the first part of the set can be photos taken in a down town, while the second part of the set can be represented by photos taken at a lake. The results of the two games by many players are combined to produce the summarization of a photo album. In this way, the summarization is conducted based on the feedback of the album owner’s friends. The game has appealing look using different visual and audio effects, as shown in Figure 6.1.

The social game developed as a Facebook application consists of three parts, as shown in Figure 6.2: the *client side* which deals with the user interface, the *server side* which performs the analysis, and the *Facebook* which provides user data, photos and authentication. These three parts communicate over an HTTP network using JSON data structure. The client side application is a weak interface for server information, where most of the processing is done on the server, therefore it makes our game easily portable to different client platforms. Authentication and photo albums are handled by Facebook through Facebook API. Finally, the scores, the information about photo albums and the results are stored on the external server side.

In order to perform summarization using players’ inputs, the application calculates three different values: *Importance*, *Segmentation* and *UserScore*.

Importance value is determined in the “Select the Best!” game for each photo album separately. The goal of this game is to select the most representative photo in the player’s opinion of the particular Facebook album of K photos, given the fact that the players can select only one representative photo among K photos. These K consecutive photos are chosen randomly from the album every time a user plays the game. Photos are then shown to the user for selection. A feature vector $Selected_n^{best}$, $n \in [1 \dots N]$, is calculated for each player, n among N players, as

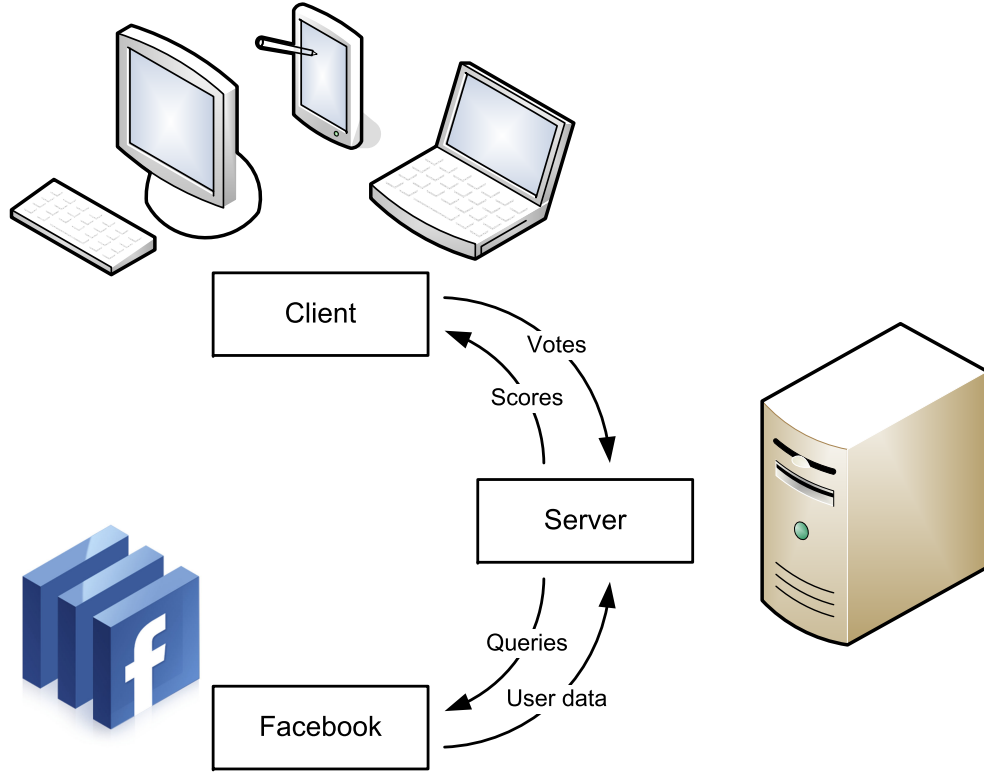


Figure 6.2: System architecture of the “Epitome” game as a Facebook application. It consists of three parts: the *client side* which deals with the user interface, the *server side* which performs the analysis, and the *Facebook* which provides necessary data and authentication.

follows:

$$Selected_n^{best}[i] = \delta_{i,s}, \quad (6.1)$$

$$Appeared_n^{best}[i] = \sum_{j \in I_K} \delta_{i,j}, \quad (6.2)$$

$$\delta_{i,j} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j, \end{cases} \quad (6.3)$$

where indices $i, j, s \in [1 \dots M]$, M is the size of a particular Facebook album, I_K is the set of indices of photos shown to the player and s is index of the selected photo. The vector $Appeared_n^{best}$ of dimension M stores the frequency of all photos that appear in the game. At the end, we perform

normalization on vector $Selected_n^{best}$ by element-wise division in order to obtain *Importance*:

$$Importance[i] = \frac{\sum_n Selected_n^{best}[i]}{\sum_n Appeared_n^{best}[i]}, \quad (6.4)$$

which is an M -dimensional vector showing the distribution of the most representative photos within one Facebook album.

Segmentation vector is calculated in the “Split it!” game for each photo album separately in an analogous way to that for *Importance*. In this game, players are asked to perform partitioning of K consecutive photos from a Facebook album into two distinct parts, by selecting a starting photo in the second segment. These K consecutive photos are chosen randomly from the album every time a user plays the game. A feature vector $Selected_n^{segm}$, $n \in [1 \dots N]$, is calculated for each player, n among N players, as follows:

$$Selected_n^{segm}[i] = \delta_{i,s}, \quad (6.5)$$

$$Appeared_n^{segm}[i] = \sum_{j \in I_K} \delta_{i,j}, \quad (6.6)$$

where indices $i, j, s \in [1 \dots M]$, M is the size of a particular Facebook album, I_K is the set of indices of photos shown to the player, and s is index of the selected photo. The vector $Appeared_n^{segm}$ of dimension M stores the frequency of all photos that appear in the game. At the end, we perform normalization on vector $Selected_n^{segm}$ by element-wise division in order to obtain *Segmentation*:

$$Segmentation[i] = \frac{\sum_n Selected_n^{segm}[i]}{\sum_n Appeared_n^{segm}[i]}, \quad (6.7)$$

which is an M -dimensional vector showing the frequency with which each photo in one Facebook album is selected as a starting photo in a new segment.

Finally, vectors *Importance* and *Segmentation* are used to automatically select L most representative photos within one Facebook photo album, as shown in Figure 6.3. In this game L was arbitrarily set to five. First, $L - 1$ maximum values from the vector *Segmentation* of the album are determined in order to segment the album into L most probable segments. For each of these segments, a photo with the highest score in the vector *Importance* is chosen. These L photos represent a collage of the album, which is shown to the owner of the album, if he/she reaches a certain level of *UserScore*. The screenshot of “My collage!” is shown in Figure 6.4.

6.3. Algorithms for Photo Album Summarization

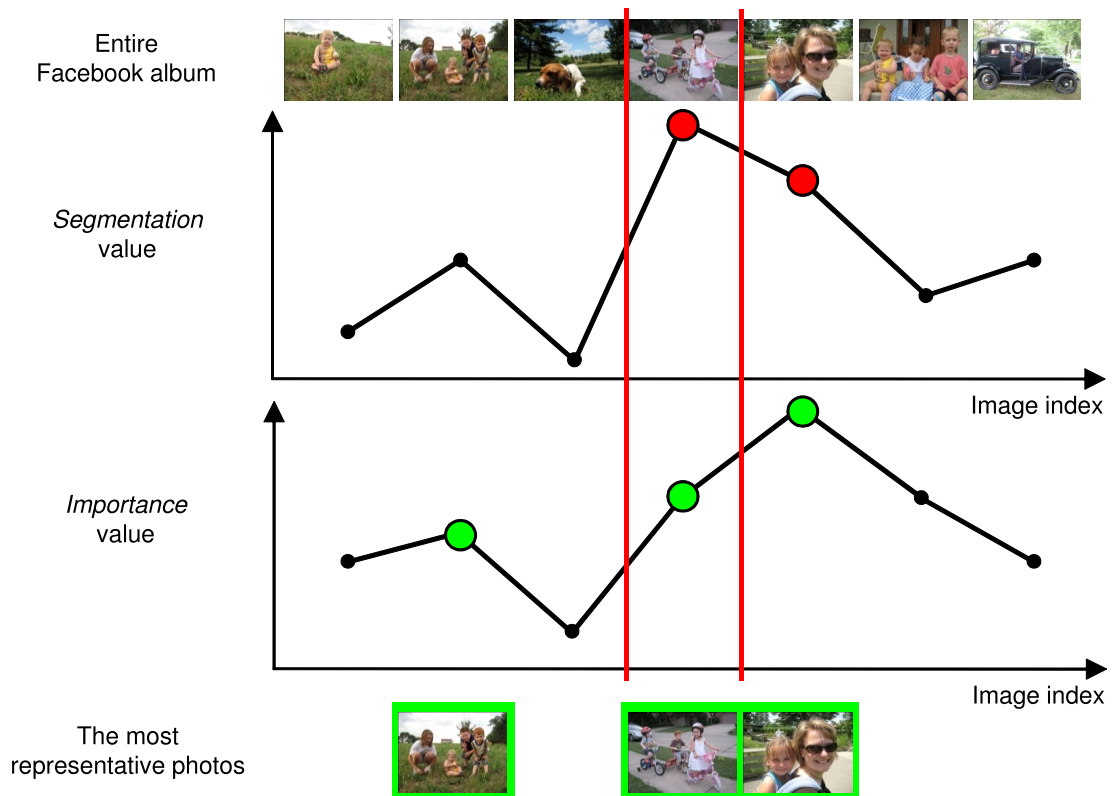


Figure 6.3: An example of selecting the three most representative photos within one Facebook album through “Epitome” game.

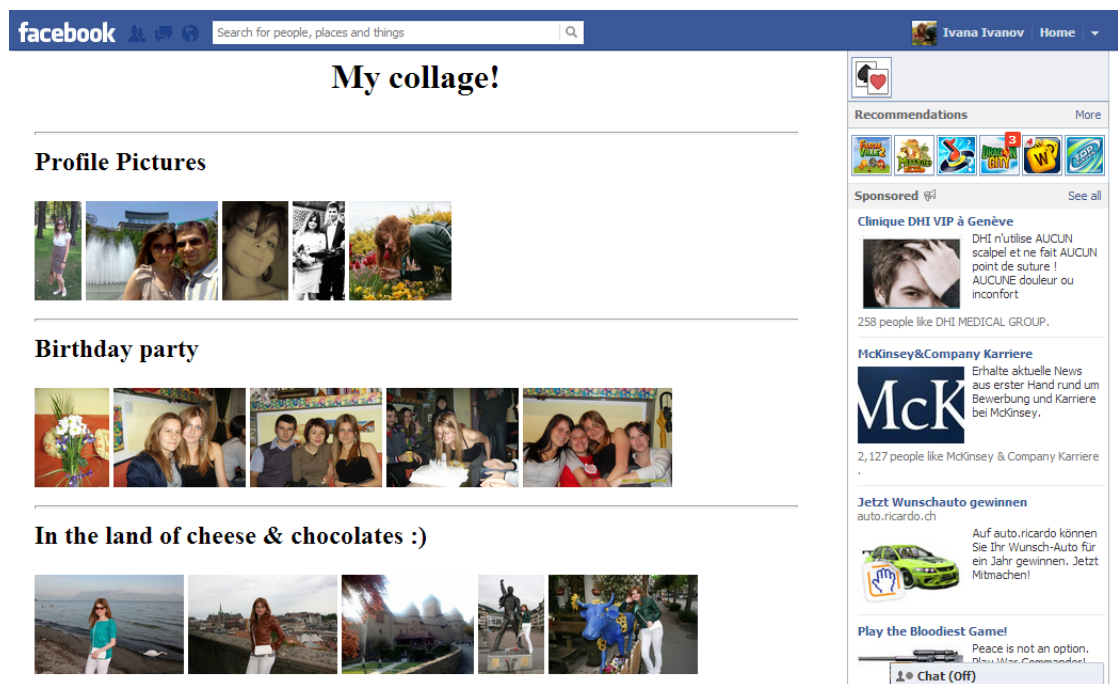


Figure 6.4: Screenshot from “Epitome” game: “My collage!” page (retrieved in January 2013).

UserScore value is defined to motivate players to play this game frequently. For example, in the “Select the Best!” game, the player increases his/her own *UserScore* if he/she selects the photo which has the highest *Importance* value among K photos. The same approach is used in “Split it!” game, where the player increases his/her *UserScore* if he/she selects the separation place where *Segmentation* value is the highest among $K - 1$ separation places. Initial *UserScore* is set to zero. *UserScore* values for all players are sorted to show ranking of players in “Epitome” game.

6.3.2 Automatic Photo Album Summarization

Automatic photo album summarization is performed considering different visual and temporal features. After extracting these features, the album is segmented into five parts by calculating the four highest Euclidean distances of the consecutive photos’ features. For each image in a particular segment, we calculate the sum of the Euclidean distances between that feature of the photo and the rest of the image features in the segment. The image with the lowest sum is then selected as the most representative photo in that segment.

Different features can be used for segmentation and selection of the most representative photo in the segments. We considered the following features: BoW method based on SURF features, HOG, HSV color histogram, “tiny” features and time stamp, as described below.

The *bag-of-words* method in computer vision was derived from BoW methods in natural language processing (NLP) [176]. A similar method in computer vision documents represents images or objects, and visual clusters of local features are considered as a word. In our case, SURF features were used as local features [79]. BoW is a vector which represents the histogram of visual features. Therefore, this method does not consider spatial information or order of visual features. 1000 feature clusters were calculated by a hierarchical k-means algorithm. Each image is represented by 1000 normalized values.

The *histogram of oriented gradients* [130] calculates the histogram of gradients in the region around one salient point in an image. It is evaluated on a dense grid of uniformly spaced cells representing salient points and uses overlapping local contrast normalization for improved accuracy. Using gradient information for feature description is very robust to different illumination conditions. The dimensions of the HOG features are around 1000.

The *color histogram* descriptor is extracted from photos in the HSV color domain. Color descriptors often fail in image retrieval in different lighting conditions, however in our case photos from one Facebook album are compared and assumed to have similar lighting conditions. The dimensions of the HSV color features are around 1000.

The “*tiny*” feature represents a scaled 32×32 grayscale version of the original color image.

The *time stamp* is extracted from EXIF data for further analysis. It corresponds to the time order



Figure 6.5: Some example photos from the database used for performance evaluation of the “Epitome” game. More sample images of this database are provided in Appendix A.6.

by which photos were uploaded to Facebook.

6.4 Experiments and Results

Evaluation of the “Epitome” game can be performed in two ways: performance and usability of the game. Each of these two evaluation methods is presented in the following sections.

6.4.1 Performance Evaluation

The performance of summarizing albums with “Epitome” game is evaluated with respect to the ground truth given by humans. In this section, we first describe the evaluation database and the conducted experiments, and then we discuss the results.

6.4.1.1 Database

The database of photos used for performance evaluation is the official database from “HP Challenge 2010: High Impact Visual Communication” at the “Multimedia Grand Challenge 2010”, which is described in details in Appendix A.6. It consists of six albums, each with 20 photos. Photos from this database depict different landmarks and famous sightseeing places, family photos, and photos of cars, flowers and sea animals. Some example photos are shown in Figure 6.5.

6.4.1.2 Evaluation Methodology and Results

We first constructed a ground truth by asking different people to subjectively perform summarization and then tested our algorithm against the ground truth data. We recruited $N = 63$ participants, among whom 61 % were males and 39 % were females, aged 18–65 (average age was 31), with different backgrounds and cultural differences. In the collection of the ground truth data, participants were shown 20 photos belonging to the same album. The task of the participants was to select the five most representative photos of the whole album, while looking at all photos of that album.

For simplicity of the explanation on how the designed photo selection tool (social game) was evaluated, let us consider only one album with $M = 20$ photos. First, a ground truth data is collected. Every user n among N users is asked to select the five most representative photos. After his/her participation in collecting the ground truth data, the corresponding feature vector $Selected_n$, $n \in [1 \dots N]$, is formed as follows:

$$Selected_n[i] = \sum_{j \in [1 \dots 5]} \delta_{i,s_j}, \quad (6.8)$$

where $i, s_j \in [1 \dots M]$ and $\forall j, l \in [1 \dots 5] : s_j \neq s_l$. s_j for $j \in [1 \dots 5]$ are the five indexes of the photos which were chosen as the representative ones. The selected indexes are distinctive. Feature vectors of the users n and m , $n, m \in [1 \dots N]$, are then compared to each other and the score of their matching $S_{n,m}$ is calculated as:

$$S_{n,m} = Selected_n \cdot Selected_m^T. \quad (6.9)$$

In other words, the higher the number of identical photos that are chosen by two users, the better will be the score of the match between them. Note that the maximum score of the match is 5. Finally, to each user n , $n \in [1 \dots N]$, a value $Performance_n$ is assigned as:

$$Performance_n = \sum_{i=1}^N S_{n,i}. \quad (6.10)$$

The maximum value in the vector $Performance$ shows the best performing participant who has the highest number of selected photos which are matched with all other users. The maximum possible value of the performance is $5 \cdot N$, which in our case becomes 315. These results are considered as the ground truth data and compared with the results obtained from the games and from the automatic album summarization algorithms, in order to prove the concept of the approach. All computations are repeated in a same way for all albums.

Then, the participants are asked to play our game with the selected database. The vectors *Importance* and *Segmentation* of dimension M are determined for each album, which are described in Section 6.3.1. These values are used to automatically select the $L = 5$ most representative photos within each album in the database. These L photos are then represented as a choice of the proposed method. Then, the complete procedure of measuring similarity between the choice of the proposed method and all other users is repeated and the final scores are computed according to Equations 6.9 and 6.10.

Furthermore, the performance of this game is compared to the performance of an automatic image selection which considers different visual and time features described in Section 6.3.2. We calculated the performance of 20 different feature pairs, considering five features for segmentation and four features for choosing the representative images, as shown in Figure 6.6. The result shows

that the best performance (around 100) is achieved by the pair of HSV color histogram, denoted as “colHSV” in Figure 6.6, for album segmentation and best photo selection in the segment. In the following the performance of automatic visual analysis, represented by color histogram, is compared with the “Epitome” game.

Figure 6.7 shows the distribution of the participants’ performance, including the choice of the proposed method, the automatic visual analysis and the random selection of photos. Performance of the random photos selection is measured as the average value of the performance of 50 users who randomly selected 5 photos in each of the 6 albums. All performances are sorted in a descending order. As one can see, the performance of the proposed method is better than the automatic visual analysis, since it is closer to the best performance of users for ground truth generation. On average, this approach achieves 80 % of the performance of the best user for each album, which proves the concept of the game. It also outperforms the automatic visual analysis, which can achieve performance of 64 %. For albums three and five, this value is even higher, i.e. about 95 %. Our approach performs significantly better than random photos selection. The most representative photos for one of the albums selected by the proposed method are shown in Figure 6.9. Figure 6.8 shows the comparison of performance in summarizing photo albums performed by “Epitome” game, automatic photo selection using color histogram and by users who participated in creating the ground truth data.

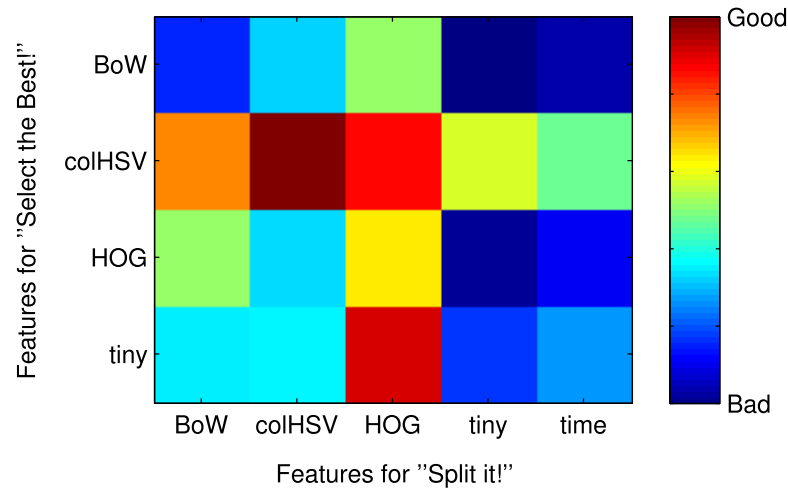


Figure 6.6: Comparison between different visual and time features. The best performance is achieved with HSV color histogram (“colHSV”) for both “Split it!” and “Select the Best!” tasks. Dark red color indicates the best (*Performance* ≈ 110) and dark blue color indicates the worst performing algorithm (*Performance* ≈ 70). For example, using “time” feature for segmenting an album and “BoW” feature for selecting the most representative images gives poor results on evaluation.

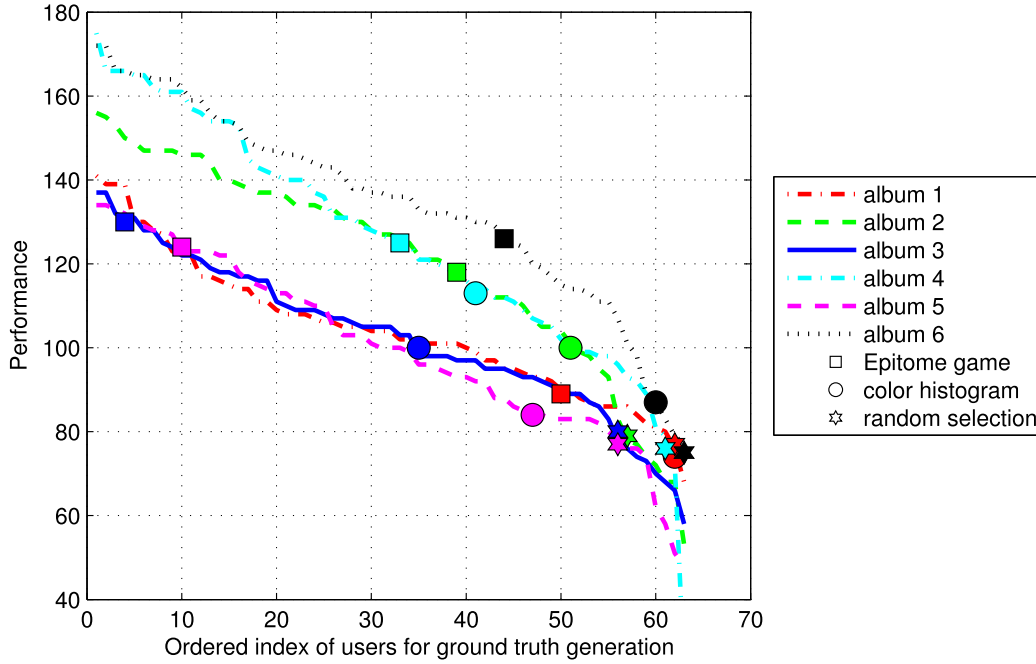


Figure 6.7: The distribution of the participants’ performance. The results of the “Epitome” game are shown with square markers, the results of automatic visual analysis with circle marker and the results of random photos selection with star markers. Different colors of the markers correspond to different albums.

6.4.2 Usability Evaluation

The usability of the “Epitome” game is evaluated through a user study. We asked participants (users) to play the game with different Facebook photo albums and to provide us their feedback on the game in the form of questionnaire.

We recruited 40 participants, among whom 63 % were males and 37 % were females, aged 23–46 (average age was 28), with different cultural backgrounds. First, all participants were introduced to “Epitome” game by showing them basic rules on how to play the game. Then, all participants spent sufficient time to play the game. After a participant played with “Epitome” game with different Facebook photo albums, a questionnaire was used to obtain the feedback from the participant. The questionnaire consists of three groups of questions:

- general questions about motivation to play the game and enjoyment;
- questions to assess different platforms for playing the game (mobile, Facebook or simple web page), e.g. satisfaction with visual presentation for each of them;
- questions about privacy issues regarding showing one’s photos to his/her friends, friends of friends, everybody or nobody.

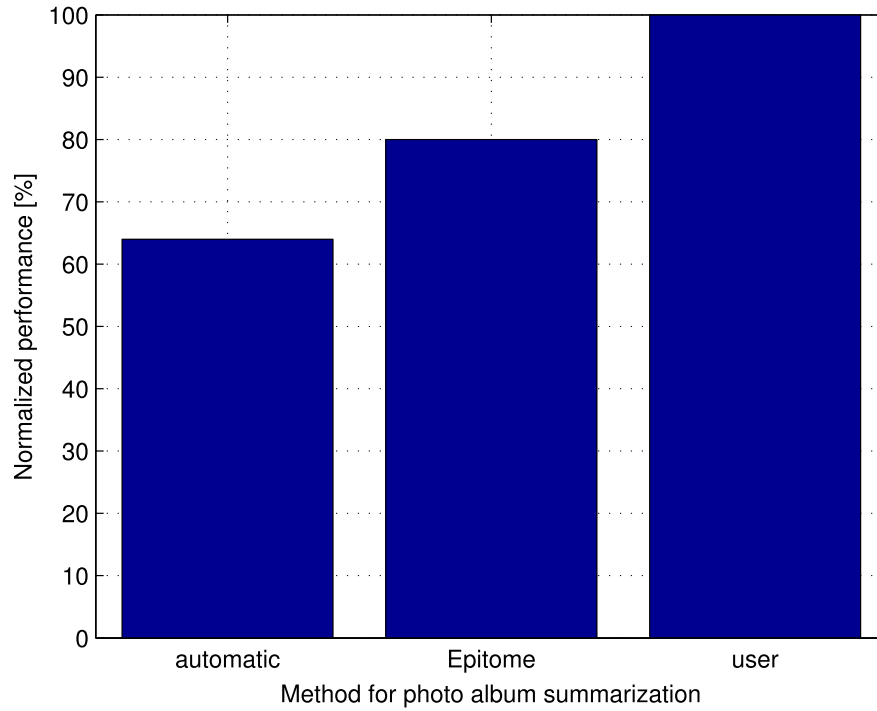


Figure 6.8: The comparison of normalized performance in summarizing photo albums performed by “Epitome” game, automatic photo selection using color histogram and by users who participated in creating the ground truth data.

In this study, we used discrete rating scales with adjective description of each level. Depending on the question, participants had to choose one of the answers or to rank answers according to their preferences. For each of the questions we calculated mean of the participants’ responses.

In this chapter, we do not describe the whole questionnaire and results, but we rather discuss some of the interesting outcomes from our study. All questions with choices are listed in Appendix C along with results for each of the questions.

6.4.2.1 Motivation to Play the Game

Questions 1–6, 9–14 and 21 listed in Appendix C.1 belong to questions about motivation to play the game. Results showed that 70 % of the players are very satisfied with the game. We further asked players of “Epitome” game what motivated them most to play the game. Results are shown in Figure 6.10. Players enjoyed the most to watch their Facebook friends’ photos, which was even more preferred than the original goal of the game, i.e. getting their own albums summarized. We observed another interesting value of the game that people like the idea of watching (browsing) friends’ photos through “Epitome” game. Players were not motivated to play “Epitome” by the fact that they participate in collecting research data. This shows that fun and enjoyment are important aspects of the game that should be considered. In another question about motivation,

Chapter 6. “Epitome” – A Social Game for Photo Album Summarization

players prefer more to see their friends’ photos compared to photos of some unknown people. This promotes the importance of the social part of the game.

One of the questions was about preferred patterns of playing the game. Like other casual games, players would like to play our game several times a month, and around five minutes every time.

6.4.2.2 Platform

An important question we discuss here is about different platforms for playing “Epitome” game (questions 7 and 8 listed in Appendix C.1), such as a simple web page, a mobile phone and a Facebook application. Average ranks for these platforms are 2.3, 2.2 and 1.5, respectively, which shows that players prefer Facebook the most. Surprisingly, players have similar preferences for mobile phone application and simple web page. One of the reasons for this could be that the mobile phone had limited bandwidth in wireless connection and the game was faltering while loading some images from Facebook.



Figure 6.9: Photos from the album 3. The most representative photos selected by the proposed method are marked with green bounding box, while the red bounding box denotes photos selected by making use of color histogram.

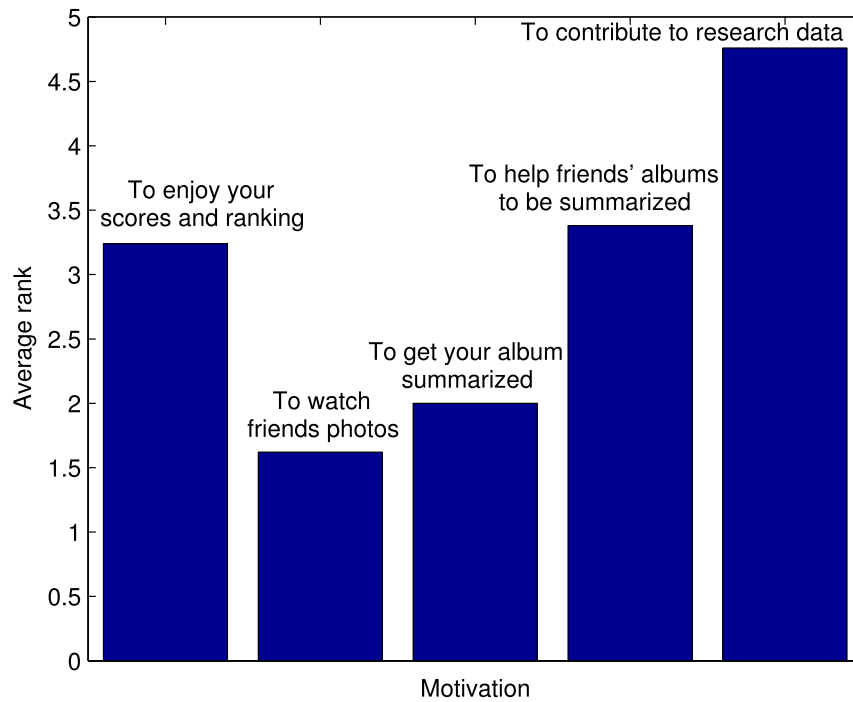


Figure 6.10: Average rank of different motivations to play “Epitome” game. Lower average ranks are better.

6.4.2.3 Privacy Issues

We also addressed Facebook permission issues. To play the third party applications, like “Epitome”, users should accept an agreement with an application on accessing users’ data stored in Facebook. But, we note that users’ privacy settings in Facebook are different from what the third party application actually access. If the user allows the third party applications to access his/her photos, they get right to distribute and modify (resize, rotate, change in color perception, etc.) photos of the user. Before using this application for the first time, Facebook shows to the player a permission page informing him/her what kind of data will be retrieved from his/her Facebook account if he/she allows access to the application, as shown in Figure 6.11 (a). People are usually not concerned about this issue and easily allow access to data by the application. In order to address the privacy issue regarding this process of allowing access to the data, we created new visually intuitive permission pages, as shown in Figure 6.11 (b)–(d). For example, since “Epitome” is dealing with photos, we created three permission pages for Facebook applications concerning: *user_photos* (if the user allows the application to modify photos he/she has uploaded or to distribute them to anybody), *user_photo_video_tags* (if the user allows the application to modify photos he/she has been tagged in or to distribute them to anybody), and *friends_photos* (if the user allows the application to modify photos his/her friends have uploaded or to distribute them to anybody). Modification of player’s photos can be creation of a collage for his/her Facebook album in which photos are resized, rotated, changed in color representation, etc.

In our experiments, users were asked whether they allow access to their data using the default Facebook permission page either on the mobile phone or in Facebook, and were separately asked if they would allow access to any of the three new permission pages. We measured how many players allow “Epitome” game to access photos they have uploaded, photos they have been tagged in and photos their friends have uploaded. Results are depicted in Figure 6.12. Clearly, the players understand the risk better by viewing our illustrative permission pages than the default Facebook permission page, and more than 90 % of the players did not allow the application to retrieve their photos that can be further modified or distributed within “Epitome”. This shows that the default Facebook permission page is neither sufficiently intuitive nor informative.

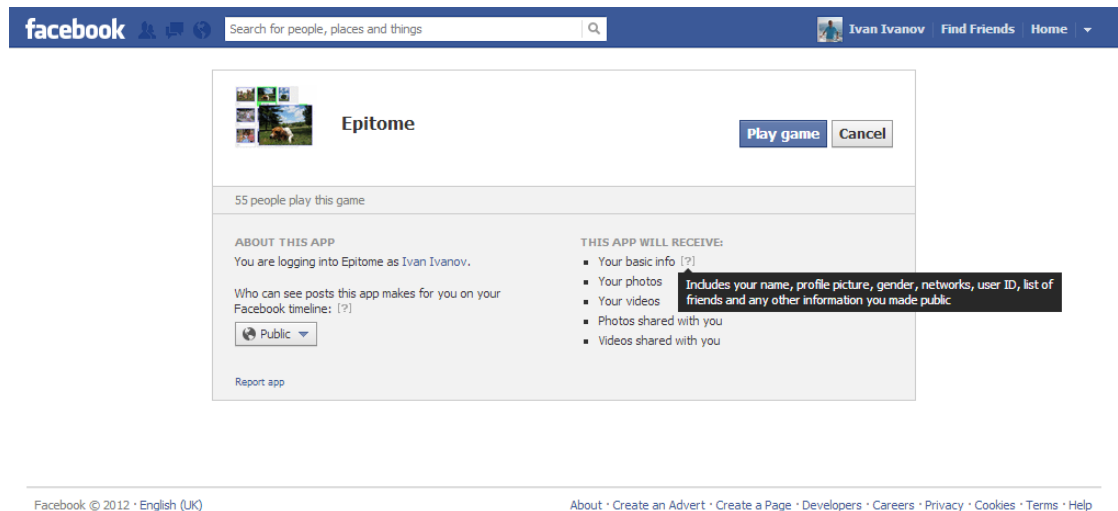
The users do not have sufficient control over details about permission in Facebook applications. From the questions related to permission settings of shared photos (questions 15–20 listed in Appendix C.1), we conclude that players would not like to give more permissions to the application compared to the permission they already set for their photos in Facebook. For example, 86 % of the users would like to share their private photos through this application only with their friends.

6.4.3 Implementation Challenges

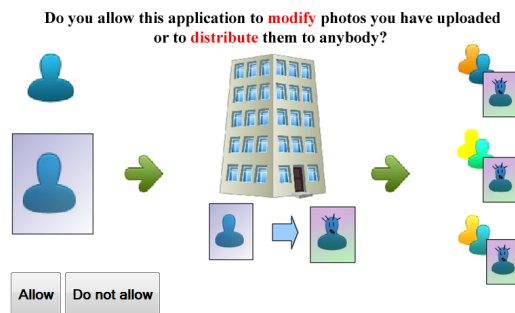
We showed in Section 6.4.2.1 that fun and enjoyment are important aspects of our game. To ensure enjoyment and to keep the game user-friendly, patience of the players should not be challenged at any time while they play the game. For example, long loading of different game content naturally annoys players. This can be disappointing from players’ perspective and it can prevent players to enjoy the game. Therefore, the flow of the game should be smooth. This is especially important in the initialization of our game, when we need lists of albums and photos to play with. These lists have to be up-to-date if we want to keep good reputation of the game.

A natural way to have up-to-date lists of albums and photos is to query Facebook through API every time a player logs into the game. However, this approach is very time-consuming. For example, in the beginning of the game, we must execute $N + 3$ Facebook queries to obtain list of player’s friends and list of photo albums for the player and all of his N friends, and one additional Facebook query for the list of photos in which the player is tagged by other Facebook users. In average, the execution time of one Facebook query to receive list of photo albums of one user is approximately 2.5 seconds. This means that if a user has 20 friends who play Epitome, then the initialization time of our game will be more than 50 seconds and the player will have to wait this amount of time in the beginning of our game. Therefore, to achieve high scalability and performance, the key point is to reduce the number of Facebook queries.

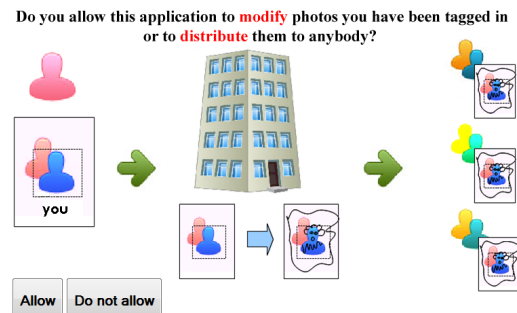
An effective approach to keep lists of albums and photos up-to-date is to use database caching. The database connected to the “Epitome” game will store the URLs of all albums and photos in the game and instead of querying Facebook, the game will query the local database stored on the server shown in Figure 6.2. Queries to a local database are faster than Facebook queries, and in



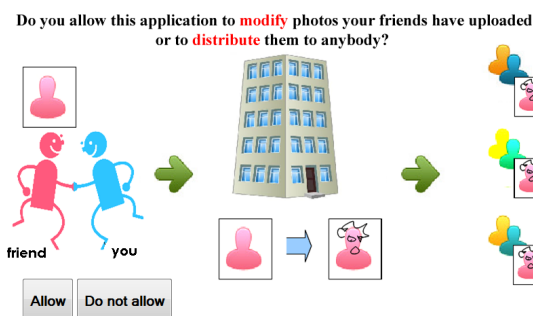
(a)



(b)



(c)



(d)

Figure 6.11: Different permission pages used in our study: (a) default Facebook permission page (retrieved in January 2013), (b) *user_photos* permission page, (c) *user_photo_video_tags* permission page, and (d) *friends_photos* permission page.

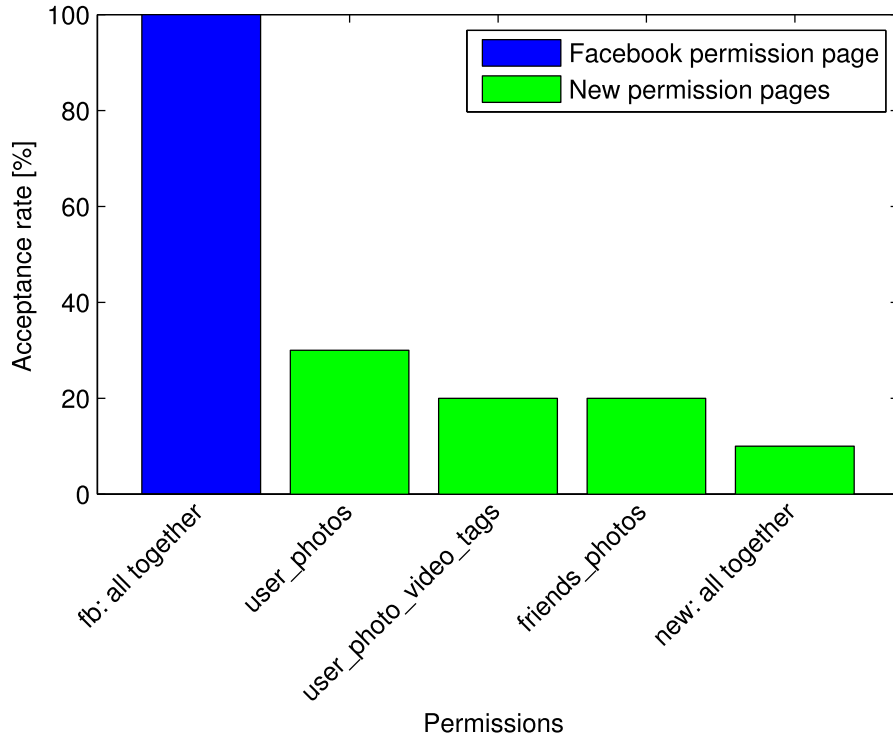


Figure 6.12: Acceptance rate for the default Facebook and permission pages used in our study.

addition, lower number of Facebook queries is required with this solution. In the beginning of the game, two Facebook queries are executed to obtain list of player’s friends and list of photo albums for the player. Then, the received number of photos in each of the player’s albums is compared with the corresponding number in the local database. If there are M albums that are modified or added since the last time this user played the game, then M additional Facebook queries are executed and retrieved URLs are copied into the local database to keep consistency with online Facebook data. Photos in which the player is tagged by other Facebook users are retrieved in one additional Facebook query and updated in the database. Finally, albums of the player’s N friends are obtained from the local database. In total, $M + 3$ Facebook queries and around N fast database queries are executed, meaning that the number of Facebook queries does not depend on the number of player’s friends. The changes in number of photo albums can be considered as relatively small compared to the number of friends, which reduces the initialization time of our game.

The two presented approaches for retrieving or updating lists of albums and photos are compared in terms of computational time. We measured computational time with respect to the number of friends, where the average number of albums per person is 3 and the average number of photos per person is 90. The results are shown in Figure 6.13. Facebook querying with database caching benefits from the reduced number of Facebook queries, when compared to the method based solely on Facebook querying. This is especially noticeable in the case when a user has many friends who play Epitome. The computational time for solely Facebook querying increases

almost linearly with the increase in number of friends, whereas the time for the method with database caching remains almost constant regardless of the number of friends.

Analogously, the approach with database caching leads to significant improvement in terms of computational time in the phase of showing photo album summarization. In this case, again, the game queries local database for the indices of the summarization sequence of photos, and not Facebook.

Although caching lists of albums and photos allows for lower computational time in the game, this method has some disadvantages. Challenges appear if a user does not play the “Epitome” game regularly, and thus the consistency between the local caching database and Facebook is broken. For example, issues in accessing albums or photos appear if a user deletes photos or albums in Facebook, where as deletion is not yet performed in the local database.

The “Epitome” game is written in PHP, while the introductory session when the system queries Facebook is written in Ajax (Asynchronous JavaScript and XML).

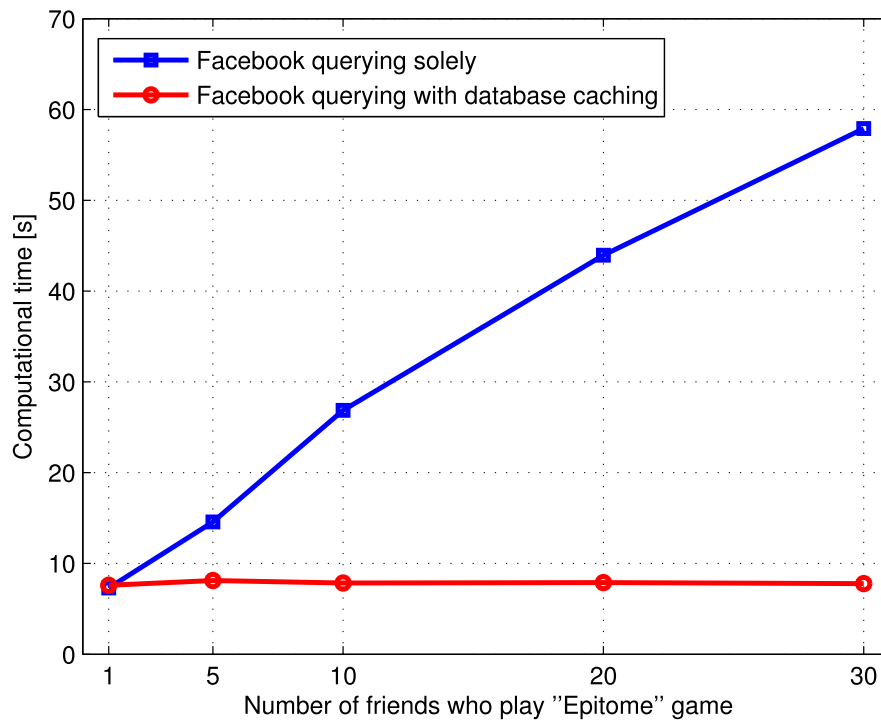


Figure 6.13: Comparison of the computational time of the initialization phase in “Epitome” game for two approaches. The computational time for solely Facebook querying increases almost linearly with the increase in number of friends, whereas the time for the method with database caching remains almost constant regardless of the number of friends.

6.4.4 Summarizing Complete Facebook with the Game

As we pointed out in a few places in this thesis, Facebook is the largest photo-sharing web site. It is interesting to estimate how long it would take players to summarize complete Facebook using the “Epitome” game, assuming that all photos on Facebook belong to one photo album. Facebook stores around 219 billion photos on its servers [36]. Given the task to summarize all Facebook photos, we assume that around 10 million players of the most popular game “Texas HoldEm Poker” [182] would participate and play the game in average 30 minutes a day [170]. A rough estimation can be achieved if we calculate the time necessary for all photos to appear in the game and to be played with. To achieve this, two of the “Epitome” games have to be played at least 438 billion times. Taking into account that the average time to play one “Epitome” game is 5 seconds, it can be estimated that we would need at least 121.67 days ($= \frac{438 \cdot 10^9 \text{ games}}{360 \cdot 10^7 \text{ games/day}}$) to summarize all photos on Facebook. The processing time to generate fine summarization sequence is a disadvantage of the “Epitome” game.

6.4.5 Statistics of the Game

The “Epitome” game was published on Facebook in June 2011, and during the first two months, 49 users played it 5870 times on a database of 21780 photos. Distribution of players’ score is shown in Figure 6.14. A few players played the game frequently and thus had higher scores than the others. Many new users started recently playing this game and therefore they still have low scores. Figure 6.15 shows the distribution of the photos’ score, i.e. the number of votes per appearance of each photo. Again, since “Epitome” game was recently published online, there are much more photos available, especially from new users, than those photos users played with in the game, and therefore many photos are not shown yet to users. This is the reason for many extreme values (score of zero and one) in Figure 6.15. Figure 6.16 shows the number of pictures changed in collages over time. It can be concluded from this figure that it converges.

6.4.6 Advantages and Disadvantages

In summary, the “Epitome” game has the following advantages:

- (1) Performance of the game-based album summarization is better than using only computer vision approaches, which was shown in Section 6.4.1.2.
- (2) People like to watch their friends’ photos through this game, which also encourages social interaction between them, as shown in Section 6.4.2.1.
- (3) The game itself is interesting and people can have fun through the game.

However, a disadvantage is the processing time for generating fine album summarization.

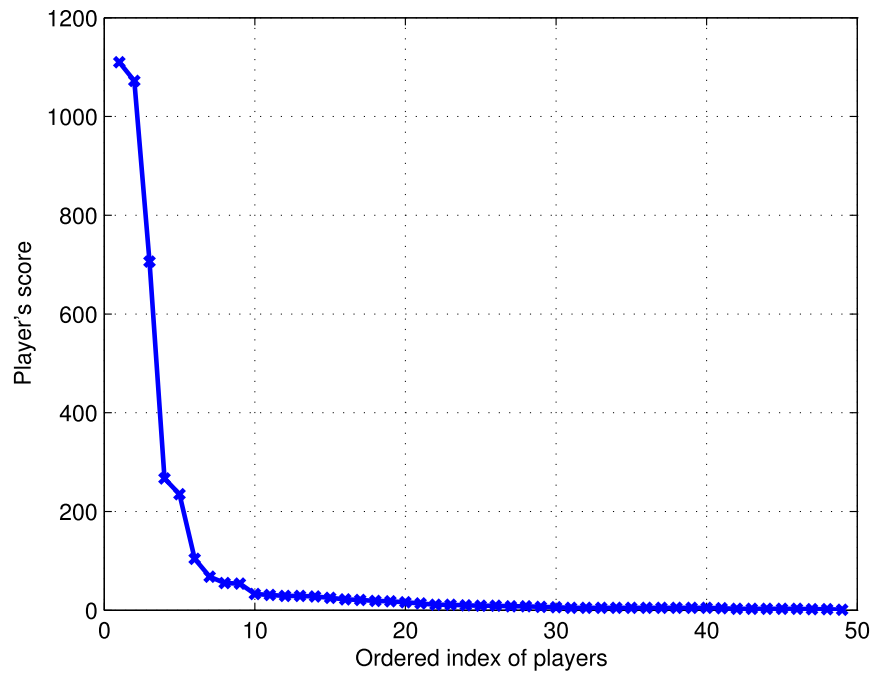


Figure 6.14: The distribution of players' score in "Epitome" game. Scores are sorted in descending order.

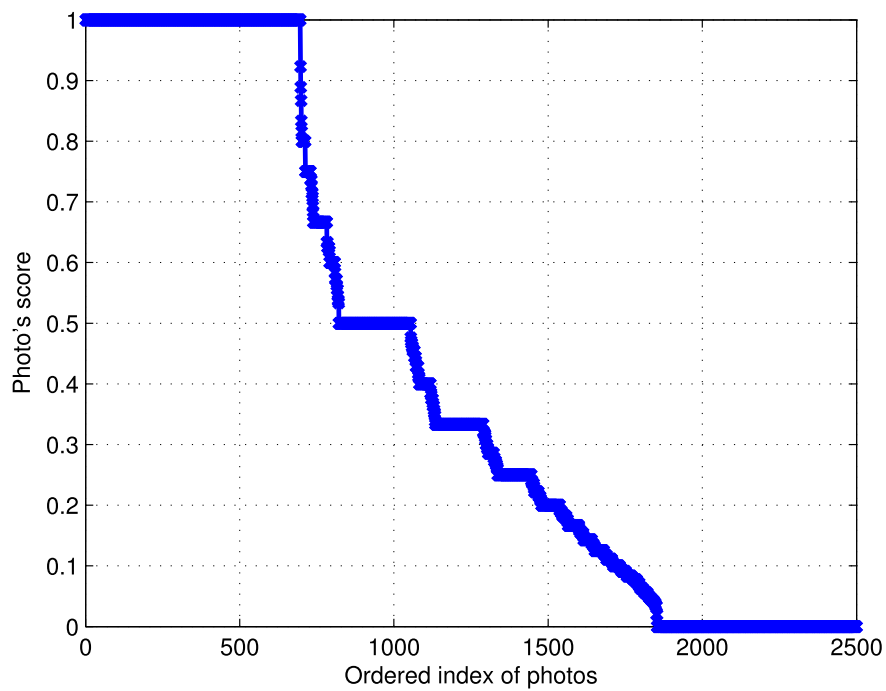


Figure 6.15: The distribution of photos' scores in "Epitome" game, i.e. the number of votes per appearance of each photo. Scores are sorted in descending order. The rest of the photos did not yet appeared or nobody voted for them.

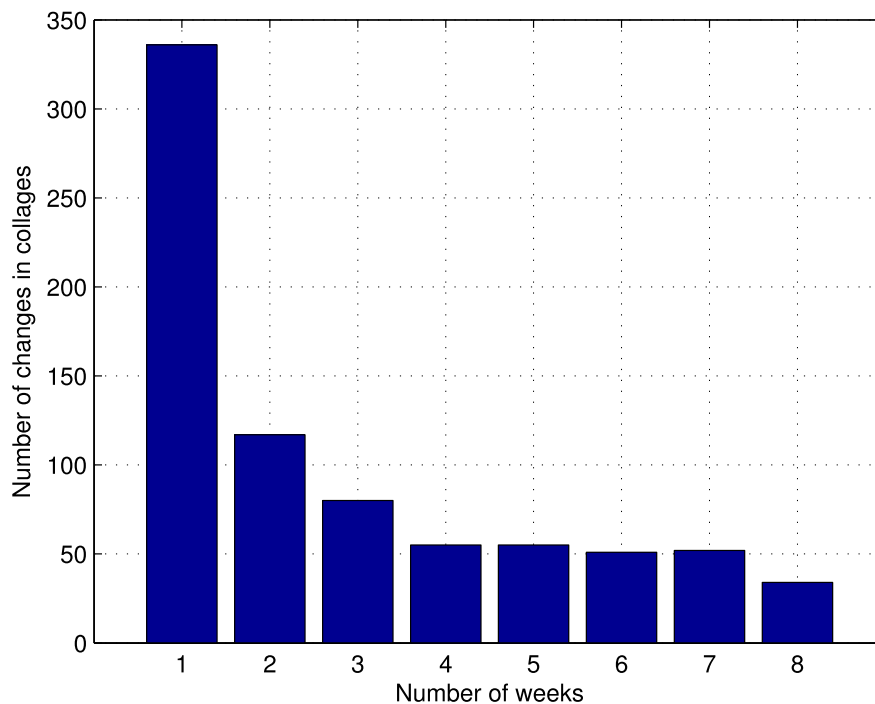


Figure 6.16: The number of photos changed in collages over time for the first two months after “Epitome” game was launched.

6.5 Conclusion

In this chapter, we described and analyzed a social game “Epitome” for photo album summarization on Facebook. The game is a social application to enjoy photos of one’s Facebook friends, while contributing to summarization of their photo albums and collecting research data. The proof of concept of the game was demonstrated and validated through a set of experiments on several photo albums. The results of the experiments showed that the summarization game achieves 80 % of the best performance of different participants and significantly outperforms automatic visual summarization methods (64 %). The usability of this game was validated by making use of a questionnaire. The results of our user study showed that the main motivation for a player of the game is to watch his/her friends’ photos and obtain his/her album summarization. Finally, a default Facebook permission page was analyzed and considered as not sufficiently intuitive nor informative.

As a future study, we will make the game more attractive for users and also consider including in this approach more sophisticated visual analysis. We also plan to improve the game by reducing the bandwidth which is necessary to load all images. The game can be adjusted and ported to any other photo-sharing web site.

Conclusions **Part IV**

7 Conclusions and Future Prospects

7.1 Summary of Achievements

This thesis addresses diverse research challenges related to social media. We have explored existing and identified new techniques to efficiently enrich each of the three key components in social media, namely metadata, users, and shared content. According to the target of analysis in social media, we have clustered our contributions into three groups: metadata, user, and content enrichment. In the following, we provide the main achievements of our work.

7.1.1 Metadata Enrichment

In Chapter 2, we have presented an efficient system for semi-automatic object tagging in images. Tagging improves “findability” of photos, while automatic tagging reduces time-consuming manual annotation. After marking desired object in an image, the proposed system performs object duplicate detection in the whole database and returns the search results with images containing similar objects. Then, the annotation can be performed through a tag propagation process, when the user enters his/her tag for the object and it is propagated to the images in the search results. The performance of the system has been assessed by evaluating the performance of the image matching and object duplicated detection steps, since tag propagation relies on their outcomes. First, we have evaluated the proposed system on a database of 3200 images associated with ground truth, and showed that the detection works reliably for salient objects such as trademarks, books, newspapers, and gadgets. Then, we have extended the evaluation to include a large-scale database of more than 1 million images. The results show that there are types of objects, where the result is satisfying. Usually text based objects work better. However, for objects, which have few number of features or they are shiny such as cars, the algorithm performs worse.

In Chapter 3, we have gone one step further towards automatic detection of informative content (objects) in images, and used detected salient objects for visual search and tag propagation. The goal of this work is to explore whether we can reduce manual work in the previous semi-

automatic tag propagation system to minimum by automatically detecting salient objects. We have performed an objective comparison of the accuracy of the saliency maps for three state-of-the-art visual attention approaches for object detection. We have showed that the performance of the considered visual attention models vary across different objects in images and object sizes. There is also trade-off between the performance and complexity of the methods. For the applications that require good precision of detected objects, the best would be to make use of the frequency-tuned approach, as it is precise enough to estimate the position of the most interesting objects in images and fast when combined with adaptive thresholding for salient object detection to be used in a real-time tag propagation scenario. We have also evaluated features for visual search extracted from different object classes. The results show that local features perform significantly better than global features for majority of the object classes. We compared performance across different types of a query image: an entire image, a ground truth object, and an automatically selected object by making use of visual attention model. Automatically selected objects in most cases perform worse than ground truth objects and entire images, or rarely reach the same level of performance as entire images. Therefore, to achieve the best performance in the object-based tag propagation task, additional user input, e.g., adjusting the borders of the bounding box around the predicted object, is necessary.

Within the work on metadata enrichment, we collected a general purpose image database containing 3200 images of 8 object classes, such as books, buildings, cars, gadgets, newspapers, shoes, text, and trademarks. All images from the database are associated with human annotated ground truth in terms of tags, and the most salient object in every image is outlined with a bounding box. We have used this database to assess the robustness of the tag propagation method with respect to different object classes, as well as, to examine the performance of different visual attention models. The database is presented in Appendix A.

7.1.2 User Enrichment

In Chapter 4, we have presented different techniques for user trust modeling that are suitable for geotagging and can be used in geotag propagation systems. The problem of having trustworthy geotags associated to content is important in social media, because geotags in form of geographical locations provide efficient information for grouping or retrieving images. We have proposed a system for automatic geotag propagation by associating locations with distinctive landmarks and using object duplicate detection for tag propagation. The adopted graph-based approach reliably establishes the correspondence between a small set of tagged images and a large set of untagged images. Based on these correspondences and a trust value of the model derived for each user, only reliable geotags are propagated, which leads to a decrease of tagging efforts. We have analyzed the performance of the tag propagation alone which leads to a promising average accuracy of 71 % over all the landmarks. We have also shown that the performance varies considerably among different landmark types depending on their visual characteristics. We have analyzed the influence of wrongly annotated tags, which causes even more wrongly propagated tags in the database. Furthermore, we have presented our socially-driven approach to model users' tagging

behavior, and compared it with four other techniques for trust modeling in social tagging systems. The results show that by propagating tags based on the trust modeling relying on users' tagging behavior, the larger number of tags (more than twice) can be propagated with the same accuracy compared to using other trust models that simply rely on the user contributed tags or if using no trust modeling at all.

In Chapter 5, we have proposed and presented 16 distinct features suitable for fighting spam in social tagging systems. The problem of having trustworthy tags associated to resources is important in social systems to identify appropriate tags and at the same time to filter or eliminate spam content or spammers. The proposed features are based on user activity in posting and tags popularity. The prominence of the features in distinguishing between legitimate users and spammers is discussed. We measured the performance of each feature solely and showed that *LegitTags* feature, defined as the probability that a particular tag is used only by legitimate users, performed the best. We also showed that aggregation of features leads to the improvement in the classification performance. Finally, performance of different classifiers was compared. The results are promising. The best classifier achieved accuracy of 0.987 with false positive rate of 0.013 in discriminating legitimate users from spammers.

Within the work on user enrichment, we collected a database of images depicting famous landmarks. The database consists of 1320 images in total from 22 cities and 66 geographically unique landmarks. All images from the database are associated with human annotated ground truth in terms of tags. We have used this image database to demonstrate the effectiveness of our method for modeling the user trust (reliability) in geotagging. The database is also suitable for image retrieval applications. The database is presented in Appendix A.

7.1.3 Content Enrichment

In Chapter 6, we have investigated how social gaming can be used to address a challenge of photo album summarization. To tell the story of some important events, it is desirable to have an efficient summarization tool which can help people to get a quick overview of an album containing a huge number of photos. We have described and analyzed a social game "Epitome" for photo album summarization on Facebook. The game is a social application to enjoy photos of one's Facebook friends, while contributing to summarization of their photo albums and collecting research data. The proof of concept of the game is demonstrated and validated through a set of experiments on several photo albums. The results of the experiments show that the summarization game achieves 80 % of the best performance of different participants and significantly outperforms automatic visual summarization methods (64 %). The usability of this game is validated by making use of a questionnaire. The results of our user study show that the main motivation for a player of the game is to watch his/her friends' photos and obtain his/her album summarization. Finally, a default Facebook permission page is analyzed and considered as not sufficiently intuitive nor informative.

Appendix B addresses an issue of interoperability between different image repositories and applications. JPSearch - Part 4 standard ensures the longevity and portability of metadata by embedding them in image files themselves. We have demonstrated the use of JPSearch in an advanced image management platform for online use, called “Cheese”. It offers user exciting features like visual similarity based search, as well as all standard features such as image upload, object-based tagging and keyword based search. Since the manual annotation of images is quite time-consuming, a semi-automatic tag propagation based on visual similarity offers a very interesting solution that is implemented in “Cheese”. For improved interoperability between different image repositories and applications, the platform accelerates the reuse of metadata by supporting the export and import of image files with embedded metadata in JPSearch - Part 4 compliant format. By making use of these features, the update of the metadata is facilitated, e.g. by adding, replacing, removing all or part of the metadata.

7.2 Future Prospects

We identify here future directions of the research performed in this thesis and state a few thoughts about the future of social media by the author of the thesis.

We have presented our semi-automatic tag propagation system in Chapter 2. The system has the potential to be improved in many ways. As a future study, one could extend it to support other classes of objects and consider evaluation of the system in the view point of the database size and latency in the system because it is important for the system to be interactive. The system can be extended to support tag recommendation, which will assist the user by suggesting probable tags for the marked object.

In the study presented in Chapter 3, we have focused on automatic detection of salient object in an images, and used detected salient object for visual search and tag propagation. We have experimented only with a single salient object in an image. However, the study could be extended to include detection of multiple salient objects in images. The future work can also focus on more sophisticated visual search algorithms or other object classes. It would be interesting to see if visual attention models could predict the most representative photos of one photo album.

The comparative study of trust modeling presented in Chapter 4 is a pioneering work, therefore a future study may consider a more careful selection of participants, for example, equal distribution of participated users in terms of group sizes and background. We have compared trust modelling for automatic tagging considering closed set problem, as we could precisely measure number of tags in the system. However, we expect that the open set problem would also work fairly good, granted that we have a “good” thresholding method for object duplicate detection step. Most of the current techniques for noise and spam reduction focus only on textual tag processing and user profile analysis, while visual features of multimedia content can also provide useful information about the relevance of the content and content-tag relationship. In the future, a promising research direction would be to combine multimedia content analysis with conventional tag processing and

user profile analysis.

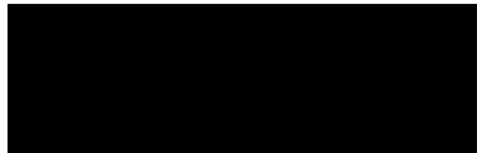
In Chapter 5, we have presented a set of distinct features to distinguish between legitimate users and spammers. As a future study, one could explore more sophisticated features which are able to deal with dynamics of trust, by distinguishing between recent and old tags. Future work considering dynamics of trust would lead to better modeling of the phenomenon in real-world applications.

A game for photo album summarization is presented in Chapter 6. As a future study, we could make the game more attractive and engaging for users. We can also consider including in this approach more sophisticated visual analysis. We also plan to improve the game by reducing the bandwidth which is necessary to load all images.

In Appendix B, we have described our system to increase the interoperability between different image repositories. The presented system has potentials for future extensions, for example to include other JPSearch compliant parts. For the time being, the platform supports importing photos from Facebook, however we can consider other social web sites, such as Flickr or Picasa. Also, a newly proposed extension of JPSearch by including ontology related technologies would be an interesting solution to consider implementing in “Cheese”. For example, adoption of a visual ontology which is aimed to be integrated as metadata description and to allow the cross linking of information. By this, any visual information and its metadata can be part of the web of things and be interlinked with semantic concepts.

The author of this thesis brings here his own thoughts on the future of social media. Social media will become an essential — not optional — form of communication. Therefore, it will allow cross-platform information exchange and usage. Of course, reliable mechanisms to preserve privacy are an essential prerequisite. From large companies and start-ups to police forces and local governments, each domain will find its own uses for social media, such as marketing, product development, customer service, or intelligence gathering. To say that the future of social media belongs to the mobile will come as no surprise to most of the readers. We will most likely see a further evolution of how we produce and consume social media content. For example, users will play an active role as content producers, broadcasting the hottest local news on TV. This trend will lead to an increase in the value of the information produced via social media. The consumption of the future social media will be performed through a mixture of interactive, immersive and effective engagement. For example, watching a movie “with” friends who are physically separated, while being able to chat with them and exchange opinions at the same time, will make future TV sets social, too. This gives us a belief that everything (products, household items, buildings, animals, etc.) will be connected to the Internet and somewhat involved in our social media world, while having the ability to communicate. Be it in the way presented here or any similar way, social media will continue to be a dynamic research challenge!

Appendices



A Database Overview

A.1 Introduction

Comprehensive databases are the key factor for the development and evaluation of the proposed algorithms in social media analysis, like in any other research field. In order to develop robust methods they have to be as realistic as possible and include all possible challenges one may face in relevant application scenarios. Furthermore, for the comparative evaluation of different algorithms usually some sort of human annotated ground truth is required to compare it to the machine generated predictions. Although the acquisition and annotation of a high quality database is a very time-consuming and resource intensive task, the availability of publicly available databases is crucial for the advancement of the field.

Databases used in social media analysis may provide different forms (type of media), such as images, bookmarks, users, tags, comments, or ratings. For example, image databases are usually used for image retrieval application scenarios in photo sharing websites. Databases of users are used for modeling and understanding communities in online social media. Initially, research on social media concentrated on individual forms due to specific application scenarios. However, as majority of these forms intersect among each other, e.g., users post some images and at the same time annotate them with tags, databases usually have data of multiple forms. Indeed, for thorough social media analysis it is better to have data of multiple forms, since different scenarios can be evaluated. Majority of the databases used in our research work are of multiple forms.

The scope of this chapter is to describe in more detail the databases that have been used throughout this work. Summary of the databases is presented in Table A.1. The databases vary in the type of media, size, and availability of the ground truth. Each of the following sections will focus on description of one of the databases.

Table A.1: Summary of the databases used throughout this work. Majority of the databases are of multiple forms and vary in the type of media, size, and availability of the ground truth. Databases created on our own are highlighted.

Database	Media (resources)	Size			Ground truth	Application	URL
		#users	#resources	#metadata			
General purpose image database	images	-	3200	3200	yes	image retrieval, tag propagation, visual attention	http://cheese.epfl.ch
MIRFLICKR-1M image database	images	-	1 M	-	yes (par- tially)	image retrieval, tag propagation	http://press.liacs.nl/mirflickr
Personal image database	images	-	≈ 16 k	≈ 0	no	image retrieval, tag propagation	http://cheese.epfl.ch
Famous landmarks image database	images	47	1320	3295	yes	image retrieval, tag propagation, trust modeling	http://cheese.epfl.ch
HP Challenge 2010 image database	images	-	120	-	yes	photo album summarization, collage creation	http://comminfo.rutgers.edu/ conferences/mmchallenge/2010/02/10/ hp-challenge-2010
ECML PKDD Discovery Challenge 2008 bookmarks database	bookmarks	≈ 32 k	≈ 2.5 M	≈ 14 M	yes	trust modeling, spam detection, tag recommendation	http://www.kde.cs.uni-kassel.de/ws/ rsdc08

A.2. General Purpose Image Database

Table A.2: Summary of the classes and some example objects of the general purpose image database. The database contains 3200 images in total, split into 8 classes of objects and 20 objects for each of the class.

Classes	Example objects	Number of images
Cars	BMW Mini Cooper, Citroen C1, Ferrari Enzo, Jeep Grand Cherokee, Lamborghini Diablo, Opel Ampera, Peugeot 206, Rolls Royce Phantom	400
Books	“Digital Color Image Processing”, “Image Analysis and Mathematical Morphology”, “JPEG2000”, “Pattern Classification”, “Speech Recognition”	400
Gadgets	Canon EOS 400D, iPhone, Nokia N97, Sony Playstation 3, Rolex Yacht-Master, Tissot Quadrato Chronograph	400
Buildings	Sagrada Familia (Barcelona), Brandenburg Gate (Berlin), Tower Bridge (London), Golden Gate Bridge (San Francisco), Eiffel Tower (Paris)	400
Newspapers	MobileZone, Le Matin Bleu, 20 Minutes, EPFL Flash	400
Text	Titles, paragraphs and image captions in newspapers	400
Shoes	Adidas Barricade, Atomic Ski Boot, Converse All Star Diego, Grubin Sandals, Merrell Moab, Puma Unlimited	400
Trademarks	Coca Cola, Guinness, Heineken, McDonald’s, Starbucks, Walt Disney	400

A.2 General Purpose Image Database

We created a general purpose image database in 2009 within the scope of this work and reported it in [50]. The database is publicly available. Part of the database is obtained from Google Image Search⁶⁰, Flickr and Wikipedia by querying the associated tags for different classes of objects. The rest of the database is formed by manually taking photos of particular objects using digital camera Canon EOS 400D.

The database consists of 3200 images: 8 classes of objects, and 20 objects for each of them. For each object, 20 sample images are collected. Summary of the considered classes and some example objects are shown in Table A.2. Figure A.1 shows a single image for a single object from some of the 160 objects, while Figure A.2 provides several images for 3 selected objects (e.g., Merrell Moab hiking shoes, Golden Gate Bridge, and Starbucks trademark). As can be seen from these samples, images with a large variety of view points and distances, as well as with different background environments, are considered for each object.

All images from the database are associated with human annotated ground truth in terms of tags that define class and object name. In addition, the most salient object in every image, which

⁶⁰ <http://images.google.com>

Appendix A. Database Overview

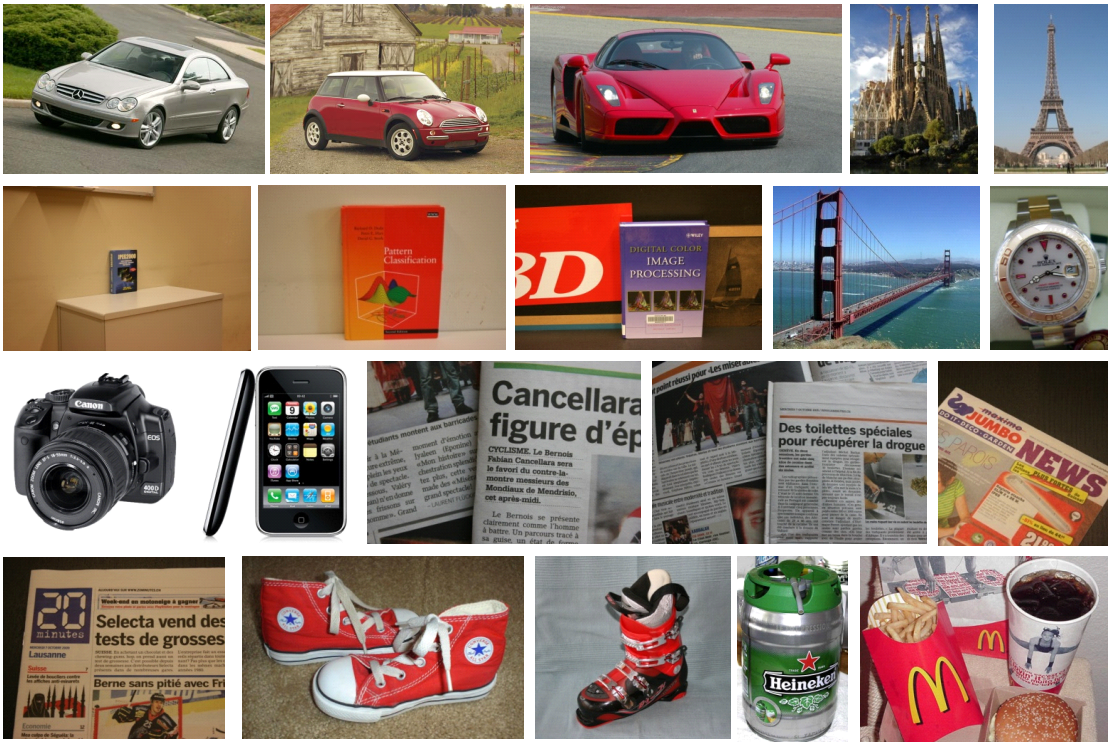


Figure A.1: Sample images from the 160 objects within the general purpose image database.



Figure A.2: Selected objects for 3 different objects from the general purpose image database: Merrell Moab hiking shoe, Golden Gate Bridge (San Francisco), and Starbucks trademark.

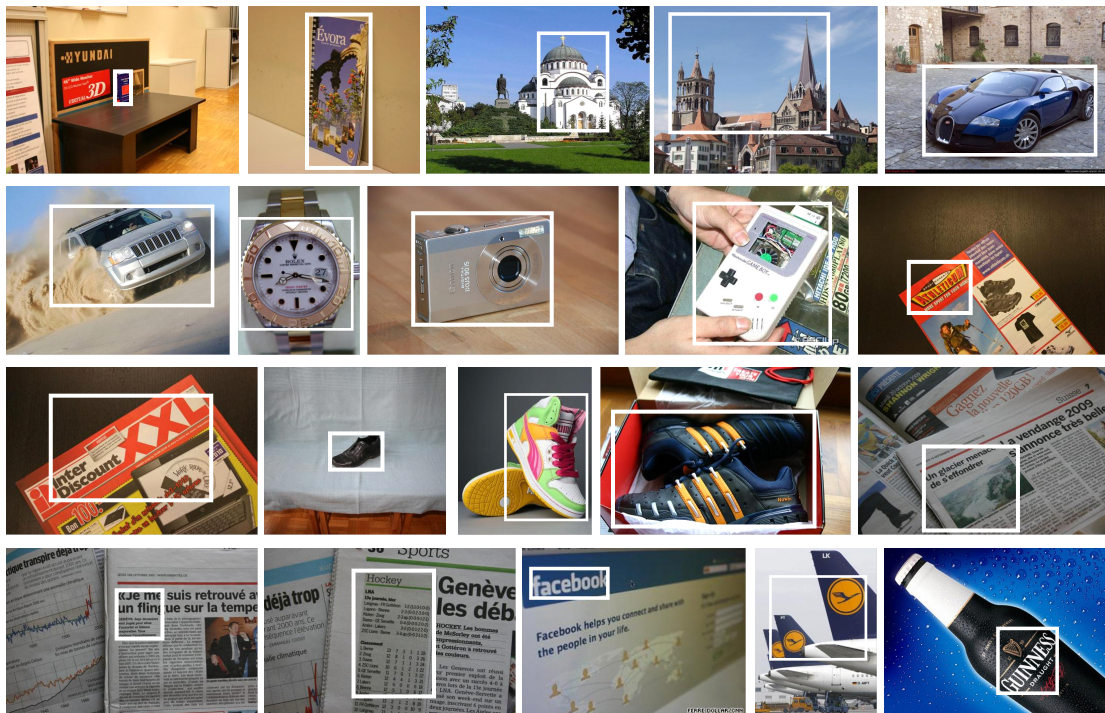


Figure A.3: Samples of salient objects from the general purpose image database along with their locations shown with white bounding boxes.

determines the class of that image, is located and its borders are outlined with a bounding box (rectangle). Here we assume that each image contains exactly one unambiguous salient object. For labeling the ground truth, we first asked three subjects to draw a bounding box in each of 3200 images to specify a salient object in the image. Then, for each image, the mean value of the position of each bounding box side provided by users is calculated to create the final location of the bounding box around the salient object in the image. As addressed in [183], one advantage is that it is much easier to provide ground truth annotation for bounding boxes than, e.g., for pixelwise segmentations. At the same time, the rectangle representation of the salient object satisfies many applications, such as adaptive image display on small devices and image collage [184]. Some examples of salient objects from our database along with their locations are shown in Figure A.3. As it can be seen, salient objects differ in color, size, position within an image, etc.

Within this work, the general purpose image database is used to assess the robustness of the tag propagation method with respect to different object classes (see Chapter 2), as well as, to examine the performance of different visual attention models (see Chapter 3). The database is also suitable for image retrieval applications, and object detection or recognition.

A.3 MIRFLICKR-1M Image Database

A MIRFLICKR-1M image database was created in 2010 by Huiskes *et al.* [185]. It is an extension in terms of the number of images from a MIRFLICKR-25000 image database published in 2008 [186]. The MIRFLICKR-1M database became very fast a popular large-scale public test benchmark database, with over 15 thousand downloads from research groups worldwide as of January 2011.

The entire database contains 1 million images of everyday scenes downloaded from the social photo sharing website Flickr. Images are selected based on their high interestingness rating according to Flickr. As a result, the image database is representative for the domain of original and high-quality photography. In particular, it is suitable for the research community dedicated to improving image retrieval, and thus it has been used in the CLEF Cross Language Image Retrieval Track (ImageCLEF⁶¹) from 2009-2012 for the visual concept detection and annotation task. Some sample images from this database are shown in Figure A.4.

All images are accompanied by the following textual features: tags that users assigned to the photos they uploaded to Flickr, EXIF metadata (if available) including information about the camera that took the photo (e.g., brand and manufacturer) and the parameters used (e.g., exposure, aperture, focal length, ISO speed, and time and date), information about the user that took the photo and the license associated with it. All photos in this database were released by their users under a Creative Commons attribution license, allowing for image use as long as the photographer is credited for the original creation. Of the entire collection, 25000 images (that initially belonged to the MIRFLICKR-25000 image database) were manually annotated with a limited number of concepts, such as sky, water, people, nights, animals, man-built structures, and indoor. A number of content-based visual descriptors for the entire database of images is supplied, as well. Visual descriptors are the MPEG-7 edge histogram, homogeneous texture and color descriptors.

We used the MIRFLICKR-1M database as a distractor database to assess the performance of the tag propagation method with respect to different object classes, as described in Chapter 2.

A.4 Personal Image Database

We collected a database of personal images from a few researchers who agreed to make publicly available images selected from their personal collections. The database is publicly available, as presented in Table A.1. This database contains personal images captured while traveling around the world, for example when attended conferences or summer schools, or during vacations or people gathering.

The database contains 16018 images (as of September 2012), and it continuously grows, thanks to the contribution from active users. Figure A.5 shows some example images from the database.

⁶¹ <http://www.imageclef.org>

A.4. Personal Image Database



Figure A.4: Sample images from the MIRFLICKR-1M database.



Figure A.5: Sample images from the personal image database.

Appendix A. Database Overview

The majority of images in this database depict famous landmarks, nature or indoor scenes, and people reunion.

The database does not contain any ground truth data. Only a small number of images are accompanied with tags provided by the creators of those images, as well as, with geotags to keep track of the locations where images were captured.

The personal image database is suitable for image retrieval applications, as images were usually taken by different people at the same event and location. Therefore, similar images from different perspectives are captured. We used this database as a distractor database to assess the performance of the tag propagation algorithm with respect to different object classes (see Chapter 2).

A.5 Image Database of Famous Landmarks

A public database of images depicting famous landmarks was created by us in 2010 as part of the reported work in [51] in order to evaluate the proposed geotag propagation method with trust modeling. We are interested in images that depict geographically unique landmarks. For instance, pictures taken by tourists are ideal because they often focus on the unique and interesting landmarks of a place. The database is obtained from Google Image Search, Flickr and Wikipedia by querying the associated tags for famous landmarks.

The database consists of 1320 images: 22 cities (such as Amsterdam, Barcelona, London, Moscow, and Paris) and 3 landmarks for each of them (objects or areas in those cities, such as Bird's Nest Stadium, Sagrada Familia, Reichstag, Golden Gate Bridge, and Eiffel Tower). Each landmark has 20 image samples. Figure A.6 shows a single image for a single landmark from each of the 22 considered cities, while Figure A.7 provides several images for 3 selected landmarks (e.g., Berlin - Reichstag, San Francisco - Golden Gate Bridge and Paris - Eiffel Tower). As can be seen from these samples, images with a large variety of view points and distances are considered for each landmark. Figure 4.5 summarizes all cities and landmarks contained in the database.

All images from the database are associated with human annotated ground truth in terms of tags that define city (name of the city where image was taken) and sublocation (area or name of the landmark), and several other tags describing landmarks depicted in images. In addition, for a set of 66 images we collected variety of user-contributed tags, while performing experiments on user trust modeling in social tagging systems described in Chapter 4. We recruited 47 participants, among whom 66 % were males and 34 % were females, aged 16-63 (average age was 29), with different backgrounds (architects, researchers, engineers, doctors, high school students) and cultural differences (from 8 different countries located mostly in Europe). The participants were asked to tag a set of images from the database, putting the name of the landmark depicted in the image, and we collected 3295 tags (658 of them were unique tags).

Within this work, the image database of famous landmarks is used to demonstrate the effectiveness

of our method for modeling the user trust (or reliability) in geotagging, as described in Chapter 4. The database is also suitable for image retrieval applications.

A.6 HP Challenge 2010 Image Database

The official dataset from “HP Challenge 2010: High Impact Visual Communication” was developed for the “Multimedia Grand Challenge 2010” and is publicly available.

The database consists of 120 images, divided into 6 albums, each with 20 images. Some example images are shown in Figure A.8. Albums cover images that are usually taken during a vacation, describing a variety of topics: images depicting different landmarks and famous sightseeing places, images with parents and kids, and images of cars, flowers and sea animals. Figure A.8 provides example images within one of the albums. Even though images of each album are pinned under the same topic, we can see that their content is rather heterogeneous, presenting different objects or scenes (mainly outdoor scenes) with large variances in color representation, presence of people, etc.

By performing experiments described in Chapter 6, we collected scores provided by different participants and selected a few the most representative photos within each of the albums in this database. The same database can be also used to generate a collage for each of the albums.

A.7 ECML PKDD Discovery Challenge 2008 Bookmarks Database

A public database is released by BibSonomy as a part of the “ECML PKDD Discovery Challenge 2008” on spam detection in social bookmarking systems.

This database consists of 31715 users who are manually labeled either as spammers (29248) or legitimate (2467) users, user-contributed tags (in total, 14074956) and resources (bookmarks; in total, 2461957) which can be either web pages or BibTeX files. Web pages are associated with, for example, URL, description, and creation date, while BibTeX files have metadata, such as author, title, journal, and publisher. However, an important skewness is present in this database since a majority of the users are spammers. This means that if a classifier labels all users as spammers, we would achieve a classification accuracy of over 0.92. Therefore, for serious evaluation, this skewness has to be taken into consideration and a balance with respect to the number of spammers and legitimate users has to be achieved. Some statistical data from this database (e.g., number of bookmarks and tags, average number of posts per user, average number of tags per user) are shown in Table 5.3 in Chapter 5.

We used this database to learn a model which effectively predicts whether a user is a spammer or not, as described in Chapter 5. This database can also be used to create a model which predicts the tags a user will use to describe a bookmark posted in BibSonomy.

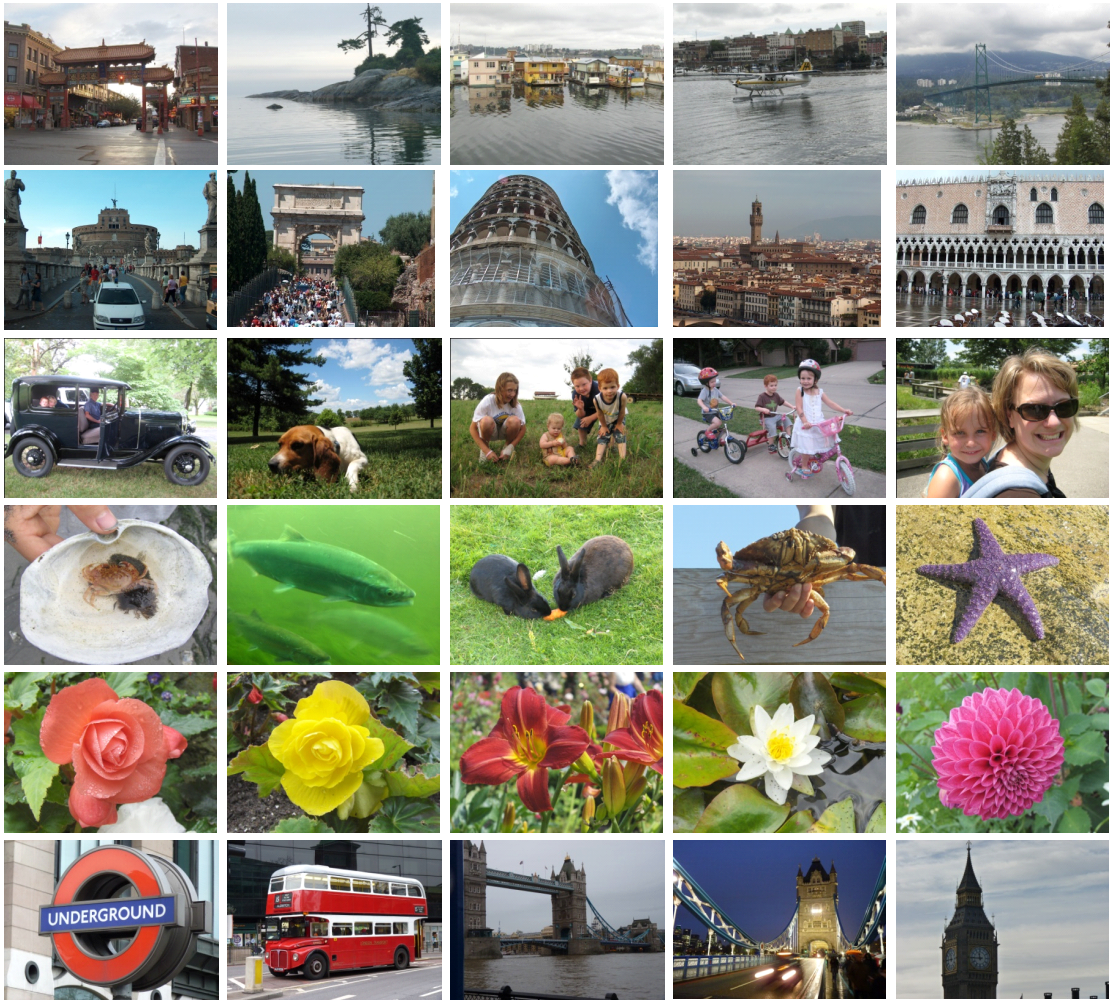


Figure A.8: Some example images for each of six albums in the HP Challenge 2010 image database. Images in each row belong to the same album. The albums cover a large variety of objects and scenes usually taken during a vacation.

B Portability of Metadata Across Image Repositories – JPSearch Standard

B.1 Introduction

Nowadays, digital images are being generated, distributed and stored worldwide. As a result, an impressive growth in personal, professional and shared image collections has been observed in recent years. For example, 4 billion photos have been shared on Instagram since its initial launch in October 2010 [187]. However, searching and managing these large distributed collections represents a considerable challenge, as discussed in Chapter 1. There are multiple social networks, repositories, systems and applications for management of multimedia content, however, almost every one provides a different search interface and metadata description format. This fact prevents users from experiencing a unified access to the multimedia repositories [188].

Users also put a lot of effort into creating metadata through tagging and commenting their own and friends' photos, which results to collaboratively generated and rich folksonomies. For example, in a large-scale analysis of users' behavior in Flickr performed over 1.83 million of photos, van Zwol [189] found that 10 % of the most frequently viewed photos receive in average between 7 and 8 comments. We assume that this number is even higher in Facebook where social networking aspect is much stronger among users. Also, Huiskes *et al.* [185] reported that the average number of tags per image is 8.94 in a MIRFLICKR-1M image database. These metadata are beneficial for both social and research communities, as they allow users and researchers to categorize images by use of tags and to easily find images concerning a certain topic.

Moreover, all metadata including tags and comments can be lost if a user decides to transfer his/her photos from one social network (or platform or device) to another. Metadata is usually stored aside image files, and if they need to be transferred to a new platform it would require additional time to separately import or enter them into the new platform. It would be more practical if metadata go together with image files and are automatically imported in a new platform. Because the portability of metadata of photos is not guaranteed, the user is de facto locked into one multimedia platform. If users want to exchange photos between different platforms, the essential requirement is a common representation of data.

Appendix B. Portability of Metadata Across Image Repositories – JPSearch Standard

The MPEG-7 standard was a first step towards addressing interoperability between photo repositories, by proposing a common syntax to describe multimedia content either with low-level or high-level features and metadata. JPSearch standard, a recent addition to a series of standards that have been developed by the Joint Photographic Experts Group (JPEG), provides a solution for the exchange of image collections and associated metadata between compliant platforms [190]. JPSearch defines a common query language with a set of precise input parameters to specify search criteria for search and retrieval over one or more local or distributed image repositories, a common specification language to allow users/systems to describe the aggregated return result sets for user presentation or machine consumption. Also, JPSearch - Part 4 aims to provide additional functionality carrying associated metadata within existing photo file formats (JPEG and JPEG 2000) to support reuse of metadata [191]. It supports two functionalities, namely the portability of metadata and the persistent association of metadata with an image.

We have developed an advanced image management platform for online use⁶², called “Cheese”. To the best of our knowledge, “Cheese” is the first platform in the world to be JPSearch - Part 4 compliant [192]. In this chapter, we present the platform. Beside standard features such as image upload, tagging and keyword based search, it offers the user search for visually similar images, object-based tagging and semi-automatic tag propagation. For improved interoperability between different image repositories and applications, the platform supports the export and import of image files with embedded metadata in JPSearch - Part 4 compliant format. Since the visual search part of this platform is described in details and evaluated in Chapter 2, we only summarize it here and put more focus on how the use and reuse of metadata is established through the JPSearch - Part 4 standard.

The remaining sections of this chapter are organized as follows. We introduce related work on standardization efforts towards making multimedia repositories interoperable in Section B.2. Then, we provide an overview of the JPSearch standard in Section B.3. Section B.4 describes our platform. Finally, Section B.5 concludes the chapter with a summary and some perspectives for the future work.

B.2 Related Work

As illustrated in the previous section, the large amount of different multimedia description formats is an obstacle in accessing simultaneously multiple media repositories. This section briefly introduces some important and well known multimedia description format standards facing interoperability during media exchange.

The Motion Pictures Expert Group (MPEG) introduced MPEG-7 standard, as an Extensible Markup Language (XML) based standard for the description and search of multimedia content [193]. MPEG-7 is titled Multimedia Content Description Interface and is formally standardized as ISO/IEC 15938 [194] in 2002. MPEG-7 proposes a rich set of descriptions and description

⁶² <http://cheese.epfl.ch>

schemes that enables a user to describe the structure, as well semantics of multimedia content, like pictures, audio, speech, and video. Part 12 of this standard (titled MPEG Query Format or MPQF) is an XML-based language which defines the format of queries and replies to be interchanged between components in a multimedia information search and retrieval system [195].

Another standard is Dublin Core [196], published in 2009 as ISO Standard 15836. It is based on the Resource Description Framework (RDF) [197]. The main goal of this standard is to define a set of metadata which is widely interchangeable among different web resources, such as video, images, and web pages, and physical resources, such as books and artworks. Focusing on simplicity, it contains only fifteen attributes, like title, creator, subject, description, date, type, format, and rights.

Targeting applications of MPEG-7 and Dublin Core include efficient search and retrieval, browsing, filtering, and universal media access [192], therefore MPEG-7, Dublin Core and JPSearch obviously target common application domains. However, their specific objectives differ greatly. MPEG-7 and Dublin Core mainly focuses on rich multimedia content annotations, which can represent both low- and high-level features of the content. JPSearch, on the other hand, does not concentrate on the annotation itself but rather on metadata interoperability by standardizing the interfaces and protocols for data exchange, thus making metadata more valuable and extending its lifespan. Therefore, MPEG-7 and Dublin Core standards on one hand, and JPSearch on the other hand, nicely complement each other.

Although the JPSearch standard has been published recently (in 2007), there are already several systems and applications which parts have been implemented in accordance with this standard. For example, the image search system developed by the Distributed Multimedia Applications Group at the Universitat Politècnica de Catalunya searches from a central point for images on different servers such as Panoramio, Picasa, or Flickr, simultaneously [198]. Another system, developed by the same group, is the BIOPSEARCH [199], a content-based medical image retrieval application which allows users navigating and searching over an image database containing optical biopsies of the human colon. The Mobile Museum Guide is a mobile application developed at the Vrije Universiteit Brussel, which allows visitors of a museum to perform content based querying by taking a picture of a painting and to receive additional information about the painting [200]. All these systems deal with the common query format and metadata interoperability, and are compliant with Parts 2 and 3 of the JPSearch standard. We go a step further and focus on reusability of metadata by embedding them into image files. We have developed an advanced image management platform for online use which is JPSearch - Part 4 compliant.

B.3 JPSearch Overview

The JPEG committee, formally known as ISO/IEC JTC1 SC29 WG1, recognized the need of a standard for interoperability among image search and retrieval systems, and provided a set of standardized interfaces for digital image management and retrieval systems called JPSearch, also

Appendix B. Portability of Metadata Across Image Repositories – JPSearch Standard

known more formally as ISO/IEC 24800. It was standardized in 2007.

There are many systems which provide image search and retrieval functionality on desktop computers, mobile phones, on the Internet, on imaging devices, and in other consumer and professional applications. Existing systems are implemented in a way that tightly couples many components of the search process. JPSearch provides an abstract framework as well as a modular and flexible search architecture that decouples the components of image search and provides a standard interface between these components. Thus, aligning image search system design to this standard framework facilitates interoperability between them. Interoperability can be defined in many ways, for example between components within one image search system that interact to provide search results so that these components could be supplied by different best-of-breed vendors, or at the metadata level such that different systems may add, update or query metadata for images and image collections [201]. Therefore, using JPSearch allows image repositories to be independent of particular system implementations and for users to easily move or upgrade their image management applications or to move to a different device or upgrade to a new computer [192].

JPSearch facilitates the use and reuse of metadata [201]. A user makes a heavy investment when annotating a collection of images. With JPSearch, the portability of the metadata is guaranteed, hence allowing a user to subsequently migrate to applications or systems which best suit his/her needs. In community based image sharing systems, such as Flickr or Facebook, this portability enables the owner of an image collection to merge community metadata back into his/her own management system, hence helping to overcome the manual annotation bottleneck.

Similarly, JPSearch makes possible the use and reuse of ontologies (formal representations of a set of concepts within a domain and the relationships between those concepts) to provide a common language for contexts. Indeed, searching for images always takes place in a context, either implicit or explicit. A common format for handling context allows a user to carry his/her context with him/her to different search engines. It also allows the context to be owned by the user and not by the system, hence protecting the user's privacy [202].

JPSearch also provides a common query language, giving search providers a reference standard to remove ambiguity in the formulation of a query, and to make searching over multiple repositories easier and consistent [203]. The common query language also defines query management process such as relevance feedback. Finally, by providing a solution for the carriage of image collections and associated metadata between compliant devices and systems, JPSearch enables image search and retrieval functionality across multiple repositories. Therefore, it allows leveraging the generally high cost of creating metadata.

JPSearch is designed as a multi-part specification. Three main processes are standardized in the specification: search and retrieval by ISO/IEC 24800-3 (query format) [204], the creation or maintenance of metadata by ISO/IEC 24800-4 (file format for metadata embedded in image data) [191] and the synchronization or migration of repositories by ISO/IEC 24800-5 (data interchange

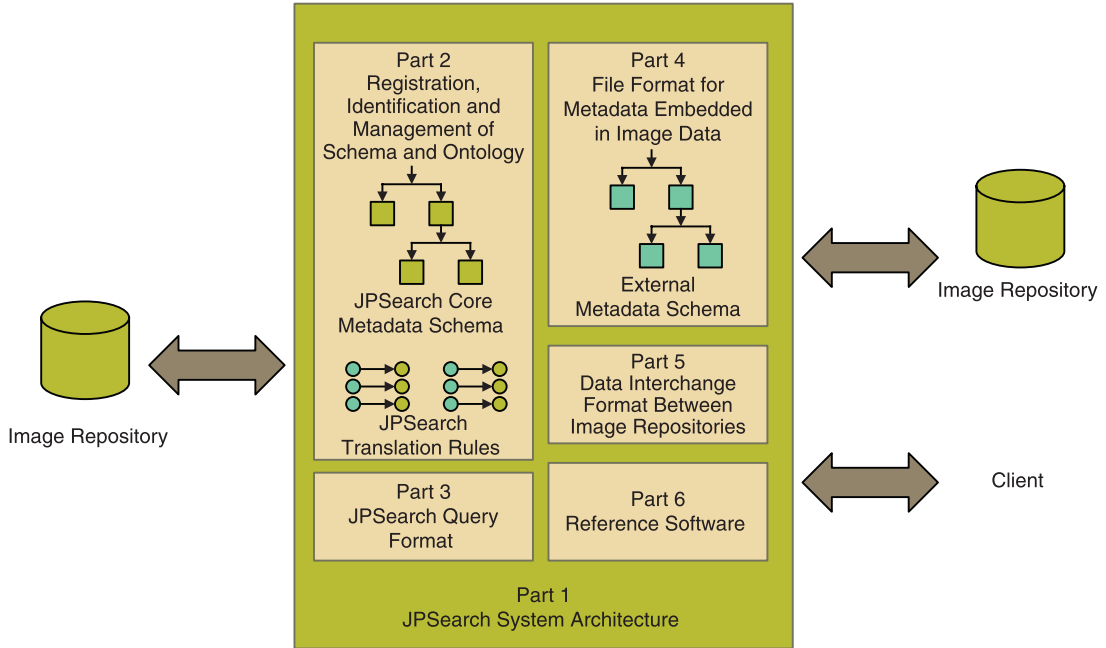


Figure B.1: The global architecture of the JPSearch system (image source: [192]).

format between image repositories) [205]. On the other hand, ISO/IEC 24800-2 (registration, identification and management of schema and ontology) links all the other parts to a common metadata interoperability model [206], which plays a key role in ISO/IEC 24800. ISO/IEC TR 24800-1 (JPSearch overview) [190] and ISO/IEC 24800-6 (reference software) [207] are intended to help understanding and developing JPSearch compliant systems. The JPSearch system architecture is constructed such that it integrates smoothly in typical image processing and management architectures enabling bilateral exchange of information between content producers, consumers, and/or aggregators, as shown in Figure B.1. In-depth presentations of JPSearch can be found in [208] and [203].

B.4 “Cheese” – JPSearch - Part 4 Complaint Platform

In this section, we present our platform for reusability of metadata associated with images. The system also supports visual search for tag propagation. Since the visual search part of this system is described in details in Chapter 2, we only summarize it here and put more focus on how the use and reuse of metadata is established by conforming the system to the JPSearch - Part 4 standard.

B.4.1 System Overview

We have developed an advanced image management platform for online use, called “Cheese”. Beside standard features such as image upload, tagging and keyword based search, it offers the

Appendix B. Portability of Metadata Across Image Repositories – JPSearch Standard

user visual similarity based search, object-based tagging and semi-automatic tag propagation. For improved interoperability between different image repositories and applications, the platform supports the export and import of image files with embedded metadata in JPSearch - Part 4 compliant format. A few screenshots from this platform are given in Figures 2.3, 2.5 and B.4 (a), where object selection and tagging, as well as tag propagation process are shown.

A user can upload as many photos as he/she wants to a database of images available online. Images are organized into variety of galleries, for example users' personal collections, social events, travelling, etc. The user can annotate any photo in the database, which is either uploaded by himself/herself or by any other user. Images are annotated on object level, with the possibility to annotate multiple objects in one image. Annotations are done by associating tags, geotags or any other sort of keywords to images. Tag propagation to similar objects is also supported in this platform, where the user confirms to which photos initial tags should be propagated, as we already presented in Chapter 2.

By uploading a new image or any metadata, the user allows others to copy, distribute and transmit their work as long as they credit the user for the original creation, but others may neither alter, transform, or build upon the work, nor use the work for commercial purposes (according to the Creative Commons license: Attribution - Noncommercial - No Derivative Works 3.0 – CC BY-NC-ND 3.0 [209]).

B.4.2 Object-based Visual Search for Tag Propagation

We summarize here our method for object-based tag propagation implemented in “Cheese”, which was previously described in Chapter 2 and reported in [50]. The system architecture is split into offline and online part, as illustrated in Figure 2.1.

The goal of the *offline processing* is to preprocess uploaded images in order to allow efficient and interactive object tagging. Salient regions in images are detected using the Fast-Hessian detector [79]. The detected regions are described using speeded up robust features (SURF) [79]. Hierarchical k-means clustering is applied to group the features according to their similarity. The vocabulary tree [82] is derived. Since the importance of the individual nodes (visual words) in the tree structure may differ among the images in the database, different term frequency - inverse document frequency (TF-IDF) weights are assigned to each of the corresponding nodes, as explained in Chapter 2. The basic idea behind IDF is that the importance of a visual word is higher if it is contained in fewer images.

The goal of the *online processing* is to propagate tag of an object in a given image to other images containing the same object. The user chooses a photo which he/she wants to annotate and marks a desired object in the image by selecting a bounding box around it. Search for similar objects is done through two-level detection approach: image matching and geometric validation by object duplicate detection. Image matching is used to select a reduced set of candidate images which are most likely to contain the target object. Based on weighting vectors, the query image

is matched to all the images in the database and the individual matching scores are computed. Image whose scores exceed a predefined threshold are discarded (see Chapter 2). Then, object duplicate detection is applied by making use of the generalised Hough transform (GHT) to detect and to localize the target object within the reduced set of images. Once target objects are detected, the user can ask the system to automatically propagate initial tags to other images within the database containing target objects.

B.4.3 Portability of Metadata

To the best of our knowledge, “Cheese” is the first platform in the world to be JPSearch - Part 4 compliant. JPSearch aims at defining the interfaces and protocols for data exchange between devices and systems, while restricting as little as possible how each component performs its task. By providing a solution for the exchange of image collections and associated metadata between compliant applications, JPSearch allows leveraging the generally high cost of creating metadata.

JPSearch facilitates the use and reuse of metadata. In particular, JPSearch - Part 4 [191], introduced as a standard in 2010, aims at providing a compatible mechanism to exchange image data and its associated metadata using existing file formats (JPEG [210] and JPEG 2000 [211]). It supports two functionalities, namely the portability of metadata and the persistent association of metadata with an image. For example, a user wants to migrate from one social network or photo management application, e.g. Facebook, to another one, e.g. Google+, and to automatically transfer all Facebook images to Google+ without losing comments, tags or other metadata associated to images by his/her family, friends or colleagues. This example is shown in Figure B.2, where screenshots of two different applications are at the top and at the bottom of the figure. In this case, a tool compliant with JPSearch - Part 4 standard, like “Cheese”, can be used to embed all Facebook comments and metadata into image files themselves, and then to automatically extract all metadata using appropriate tool on Google+. Additional metadata can be added on Google+ and again embedded inside image files, hence preserving metadata even when the user switches from one social network or photo-sharing web site to another one.

In order to provide this additional functionality of making metadata portable, JPSearch proposes a specific syntax. It is an extension of JPEG and JPEG 2000 file format which carries associated metadata within an image file. In the following, we will describe how metadata are embedded inside JPEG file format, as presented in Figure B.3. Embedding metadata inside JPEG 2000 file format is done in a similar way as described in [191]. The JPEG specification defines a set of segments called application marker (APP) segments that allow for the addition of application specific information. For example, Exchangeable Image File Format (EXIF) [141] uses APP1 segment to insert some of camera data information. Metadata in JPSearch file format are stored in APP3 segment. Multiple APP3 segments can co-exist inside one JPEG file, if it is necessary to store metadata of different types or if the size of metadata exceeds the size limitation of APP segment defined in JPEG specification. Each APP3 segment has one JPSearch Metadata block which is a container of JPSearch metadata and each of them has one or more Elementary Metadata

Appendix B. Portability of Metadata Across Image Repositories – JPSearch Standard

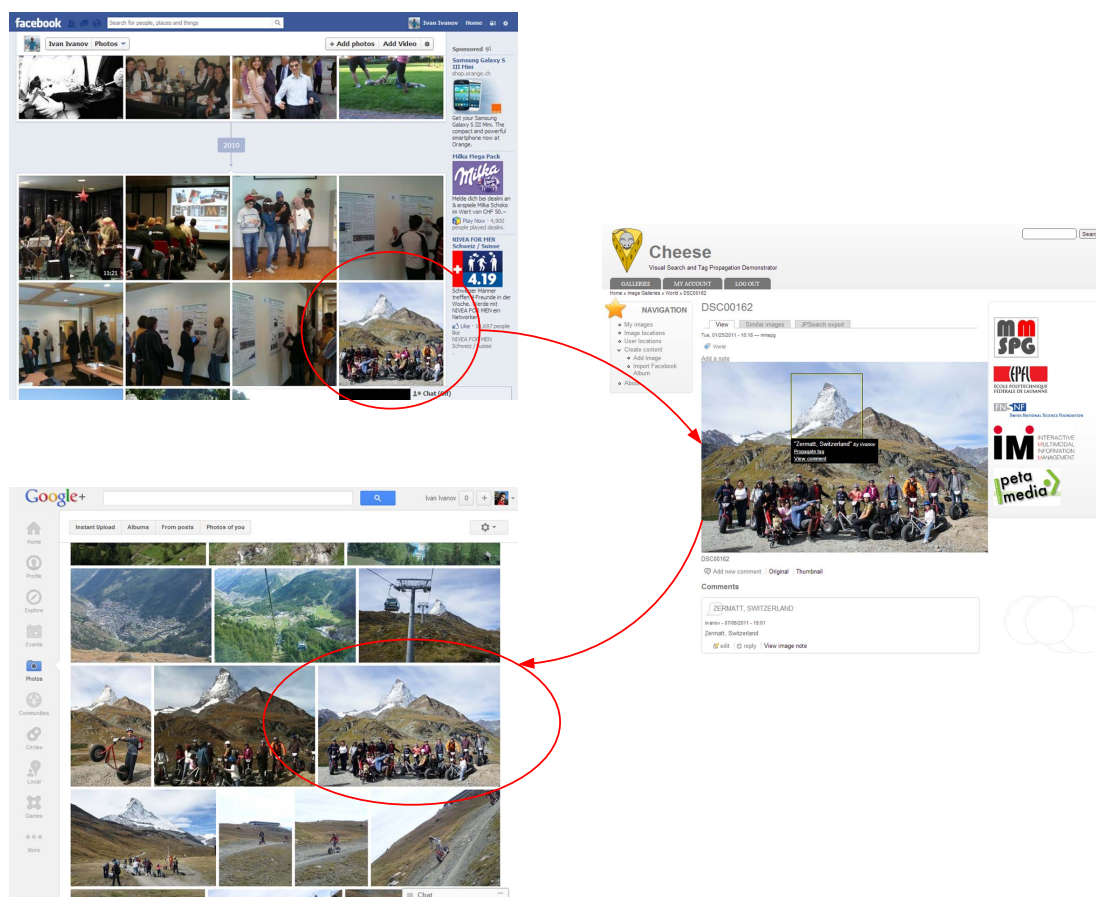


Figure B.2: “Cheese” platform is JPSearch - Part 4 compliant. It supports the export (e.g., from Facebook) and import (e.g., to Google+) of image files with embedded metadata in JPSearch - Part 4 compliant format (screenshots retrieved in January 2013).

blocks inside. Elementary Metadata block is a basic segment of JPSearch file format and it stores one instance of certain metadata schema by certain author. Data block inside Elementary Metadata stores binary stream of metadata. Several Elementary Metadata blocks using the same JPSearch metadata schema can be also instantiated simultaneously to implement social tagging functionality. Types of metadata schemes, as well as coding methods (e.g., XML representation), are defined in ISO/IEC 24800-2 [206]. JPSearch Metadata and Elementary Metadata blocks have additional fields to specify the start code, the size of the block, the author’s name, the last update date/time, etc. Detailed information on JPSearch - Part 4 file format structure is provided in ISO/IEC 24800-4 [191].

“Cheese” platform uses textual XML representation of metadata. JPSearch Core Metadata schema is described in ISO/IEC 24800-2 [206]. It defines different types, elements and attributes that can be used to describe the information about an image. An example presented in Figure B.4 (c) shows an XML representation of metadata associated with the image in Figure B.4 (a). It has a unique ID, creator information, GPS coordinates where the image was taken. In addition,

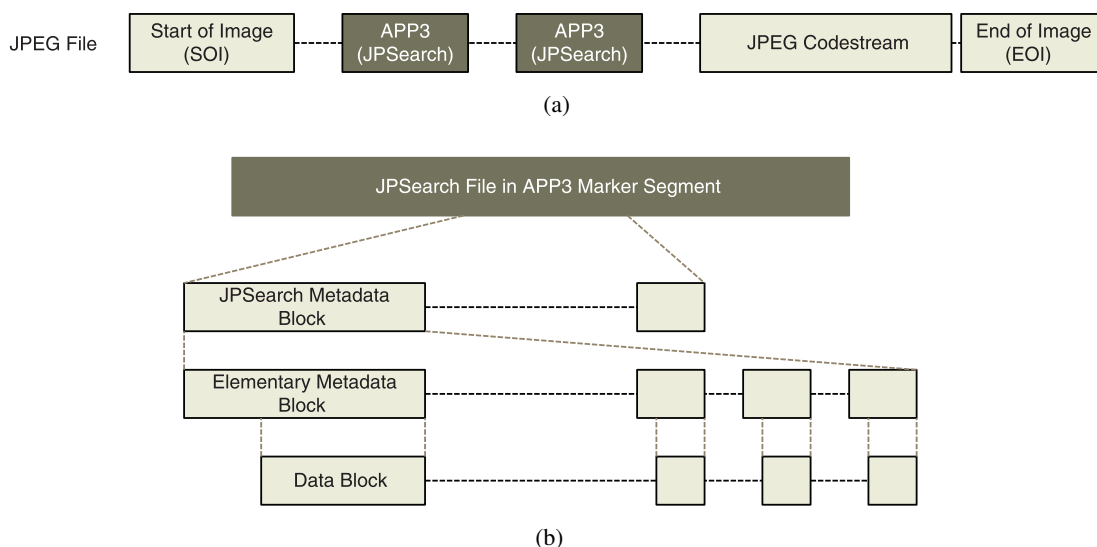


Figure B.3: An example of JPSearch files embedded in (a) JPEG file format, with the potential to have multiple JPSearch files carried by one image file. (b) Each JPSearch file can contain on its turn multiple JPSearch Core Metadata schema and registered schema. Each Data block is an instance of a JPSearch Core Metadata schema (at least one) or JPSearch registered schema each associated with a specific, potentially different author (image source: [192]).

a certain region (object) in the image has been annotated with bounding box and accompanied with a description (tags) and keyword (tagger’s username). The corresponding JPSearch file embedded in the JPEG image file is shown in Figure B.4 (b).

“Cheese” platform also allows importing images from a user’s Facebook photo albums together with associated metadata, like comments and tags. To make sure that privacy issues are addressed within “Cheese”, the user has the possibility to select photos that he/she wants to import into the platform, and make visible to the other users. Metadata linked to each of the photos can be additionally updated in “Cheese” by making use of JPSearch standard, as a way to preserve longevity of metadata. Metadata are stored in image files using the representation presented above.

B.4.4 Implementation Details

The implementation of the modules that support JPSearch - Part 4 standard inside the presented system to read and write XML representation out of the metadata, and to read and write its binary representation inside JPEG file format are done in PHP and Java, respectively. Some modules use external libraries:

- (1) JDOM⁶³, an open source Java-based solution for accessing, manipulating, and outputting

⁶³ <http://www.jdom.org>

Appendix B. Portability of Metadata Across Image Repositories – JPSearch Standard



(a)

<pre>SOI ... APP3 Marker APP3 Length JPSearchMetadata { StartCode = 'JPS' 0x00; VersionID = 0x01; NumberOfElementaryMetadata = 1; [ElementaryMetadata { StartCode = 'SEM' 0x00; LengthOfBlock = ...; SchemaIdentifier = "jpcore"; Annotation {...} LengthOfData = ...; Encoding = 'T'; Data = 'XML data in (c)'; } ...] } ... JPEG Codestream EOI</pre>	<pre><?xml version="1.0" encoding="UTF-8"?> <ImageDescription xmlns="JPSearch:schema:coremetadata" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="JPSearch:schema:coremetadata jpcore.xsd"> <Identifier> DSC00162:mmspg:unique:identifier:1:2:3 </Identifier> <Creators> <GivenName> Touradj </GivenName> <FamilyName> Ebrahimi </FamilyName> </Creators> <GPSPositioning latitude="46.016667" longitude="7.75" altitude="1608"/> <RegionOfInterest> <RegionLocator> <Region dim="2"> 200 30 360 180 </Region> </RegionLocator> <Description> Zermatt, Switzerland </Description> <Keyword> iivanov </Keyword> </RegionOfInterest> ... </ImageDescription></pre>
--	---

(b)

(c)

Figure B.4: An example of (a) an annotated object in an image (screenshot retrieved in January 2013), (b) the corresponding JPSearch file embedded in the JPEG image file, and (c) the corresponding XML representation (JPSearch - Part 4 compliant) of the metadata.

XML data from Java code.

- (2) Java Image I/O API, a Java plug-in for handling JPEG images, also specialized to manipulate the content of APP segments. This plug-in is available online⁶⁴.

B.5 Conclusion

Social networks are gaining popularity for sharing interests and information. Especially photo sharing and tagging is becoming increasingly popular. Photos are usually associated with metadata, such as tags, geotags, comments and ratings.

In this work, we demonstrated the use of JPSearch in an advanced image management platform for online use, called “Cheese”. It offers user exciting features like visual similarity based search, as well as all standard features such as image upload, object-based tagging and keyword based search. Since the manual annotation of images is quite time-consuming, a semi-automatic tag propagation based on visual similarity offers a very interesting solution that is implemented in “Cheese”. For improved interoperability between different image repositories and applications, the platform accelerates the reuse of metadata by supporting the export and import of image files with embedded metadata in JPSearch - Part 4 compliant format. By making use of these features, the update of the metadata is facilitated, e.g. by adding, replacing, removing all or part of the metadata.

The presented system has potentials for future extensions, for example to include other JPSearch compliant parts. For the time being, the platform supports importing photos from Facebook, however we can consider other social web sites, such as Flickr or Picasa. Also, a newly proposed extension of JPSearch by including ontology related technologies would be an interesting solution to consider implementing in “Cheese”. For example, adoption of a visual ontology which is aimed to be integrated as metadata description and to allow the cross linking of information. By this, any visual information and its metadata can be part of the web of things and be interlinked with semantic concepts [208].

⁶⁴ <http://docs.oracle.com/javase/6/docs/technotes/guides/imageio/index.html>

C A User Study of the Social Game “Epitome”

The usability of the “Epitome” game is evaluated through a user study, which is previously described in Chapter 6. We asked participants (users) to play the game with different Facebook photo albums and to provide us their feedback on the game in the form of questionnaire. In this appendix, we first list all questions with choices, and then present results for each question of the questionnaire.

C.1 Questionnaire

The usability of the “Epitome” game is evaluated by making use of a questionnaire on several subjects who played the game. The questionnaire consists of three groups of questions:

- (1) general questions about motivation to play the game and enjoyment (questions 1–6, 9–14 and 21),
- (2) questions to assess different platforms for playing the game (questions 7 and 8), and
- (3) questions about privacy issues regarding showing one’s photos to his/her friends, friends of friends, everybody or nobody (questions 15–20).

The entire questionnaire is given below.

About the participant

Name: _____

Gender: _____ Age: _____

Occupation: _____

Country: _____

About the “Epitome” game

1a. Are you satisfied with the “Epitome” game?

(Please select one answer)

- ☐ Completely satisfied
- ☐ Very satisfied
- ☐ Fairly well satisfied
- ☐ Somewhat unsatisfied
- ☐ Very unsatisfied

1b. If not completely satisfied, what is the main reason for that?

(Please select one answer)

- ☐ Time-consuming
- ☐ Complicated
- ☐ Boring or unattractive
- ☐ You do not see the point of the game
- ☐ Other: _____

2. Please rank the following motivations to play the “Epitome” game according to your preferences in order to make it more enjoyable?

(1 – most preferred, 5 – least preferred)

- ___ To enjoy your scores and ranking
- ___ To watch friends photos
- ___ To get your album summarized
- ___ To help friends’ albums to be summarized
- ___ To contribute to research data

3. Please rank the following improvements of the “Epitome” game according to your preferences in order to make it more enjoyable?

(1 – most preferred, 5 – least preferred)

- ___ Convenience (less time-consuming, easy to play)
- ___ Different levels, championships
- ___ Graphic and sound effects, animations
- ___ Less images to show in one screen
- ___ Interaction with other players

4. How often would you play the “Epitome” game?

(Please select one answer)

- ☐ Several times per day
- ☐ Once per day
- ☐ Once in a week
- ☐ Once in a month
- ☐ Not at all

5. How long would you play the “Epitome” game at once?
(Please select one answer)

- ☐ 5 seconds
- ☐ 30 seconds
- ☐ 1 minute
- ☐ 2 minutes
- ☐ 5 minutes

6. Would you prefer to play only one integrated game?
(Please select one answer)

- ☐ Yes
- ☐ No, I want to play “Select the Best!” and “Split It!”
- ☐ No, I want to play several games
- ☐ I don’t mind

7. Please rank these three platforms for playing the “Epitome” game according to your preferences?

(1 – most preferred, 5 – least preferred)

- ___ Mobile
- ___ Facebook
- ___ Simple web page

8a. How do you like the mobile interface?

(Please select one answer)

- ☐ Very good
- ☐ Good
- ☐ Not so good
- ☐ Not good at all
- ☐ I don’t mind

8b. And how do you like the Facebook interface?

(Please select one answer)

Appendix C. A User Study of the Social Game “Epitome”

- ☐ Very good
- ☐ Good
- ☐ Not so good
- ☐ Not good at all
- ☐ I don’t mind

9. Would you enjoy the “Epitome” game more, if you play with less than 9 images? If yes, how many images should be displayed?

(Please select one answer)

- ☐ Yes, and how many?
 - ☐ 8
 - ☐ 6
 - ☐ 4
 - ☐ 2

☐ No

10. How much do you prefer to watch your friends’ photos compared to the photos of unknown people?

(1 – most preferred, 5 – least preferred)

(Please select one answer)

- ☐ 1
- ☐ 2
- ☐ 3
- ☐ 4
- ☐ 5

11. Is it good to show your rank and compare it with your friends’ ranks for the enjoyment of “Epitome” game?

(Please select one answer)

- ☐ Very good
- ☐ Good
- ☐ Not so good
- ☐ Not good at all
- ☐ I don’t mind

12. Is it good to have your summarization sequence as a result of the “Epitome” game?

(Please select one answer)

- ☐ Very good
- ☐ Good
- ☐ Not so good
- ☐ Not good at all
- ☐ I don't mind

13. How many images in album summarization sequence of photos would you prefer?

(Please select one answer)

- ☐ 1
- ☐ 2–4
- ☐ 5–10
- ☐ 11–15
- ☐ I don't mind

14. There are two statements:

1st statement – to have perfectly summarized Facebook album, but waiting for it long time period,

2nd statement – to have preliminarily summarized Facebook album after a short time.

Which of these statements is more important for you?

(Please select one answer)

- ☐ 1st statement is the most important for me
- ☐ I prefer the 1st statement
- ☐ Both statements have the same importance for me
- ☐ I prefer the 2nd statement
- ☐ 2nd statement is the most important for me

15. To whom would you allow “Epitome” to show your private photos which are not shared even with your friends in order to receive a good summarization of your Facebook albums?

(Please select one answer)

- ☐ Everybody
- ☐ Friends of your friends
- ☐ Only your friends
- ☐ Nobody

16. To whom would you allow “Epitome” to show your private photos which are shared just with your friends in order to receive a good summarization of your Facebook albums?

(Please select one answer)

- ☐ Everybody

Appendix C. A User Study of the Social Game “Epitome”

- ☐ Friends of your friends
- ☐ Only your friends
- ☐ Nobody

17. To whom would you allow Epitome to show your private photos which were shared with friends of friends in order to receive a good summarization of your Facebook albums?

(Please select one answer)

- ☐ Everybody
- ☐ Friends of your friends
- ☐ Only your friends
- ☐ Nobody

18. To whom would you allow Epitome to show photos in which you were tagged in order to receive a good summarization of your Facebook albums?

(Please select one answer)

- ☐ Everybody
- ☐ Friends of your friends
- ☐ Only your friends
- ☐ Nobody

19. To whom would you allow Epitome to show photos of your friends in order to receive a good summarization of your Facebook albums?

(Please select one answer)

- ☐ Everybody
- ☐ Friends of your friends
- ☐ Only your friends
- ☐ Commonly shared friends
- ☐ Nobody

20. Do you want to play with photos of your friends even if they do not play Epitome?

(Please select one answer)

- ☐ Yes
- ☐ No

21. Any suggestions to improve the game?

Thank you very much for participating in this survey.

C.2 Results of the Questionnaire

In this section we present results for each question in the questionnaire. The results are shown in Figure C.1, which spans over the next three pages. For each question, the distribution of the participants' answers is shown either in the form of pie charts or bar plots. Different answers are distinguished with different colors. Some of the interesting outcomes from this study are discussed in Section 6.4.2.

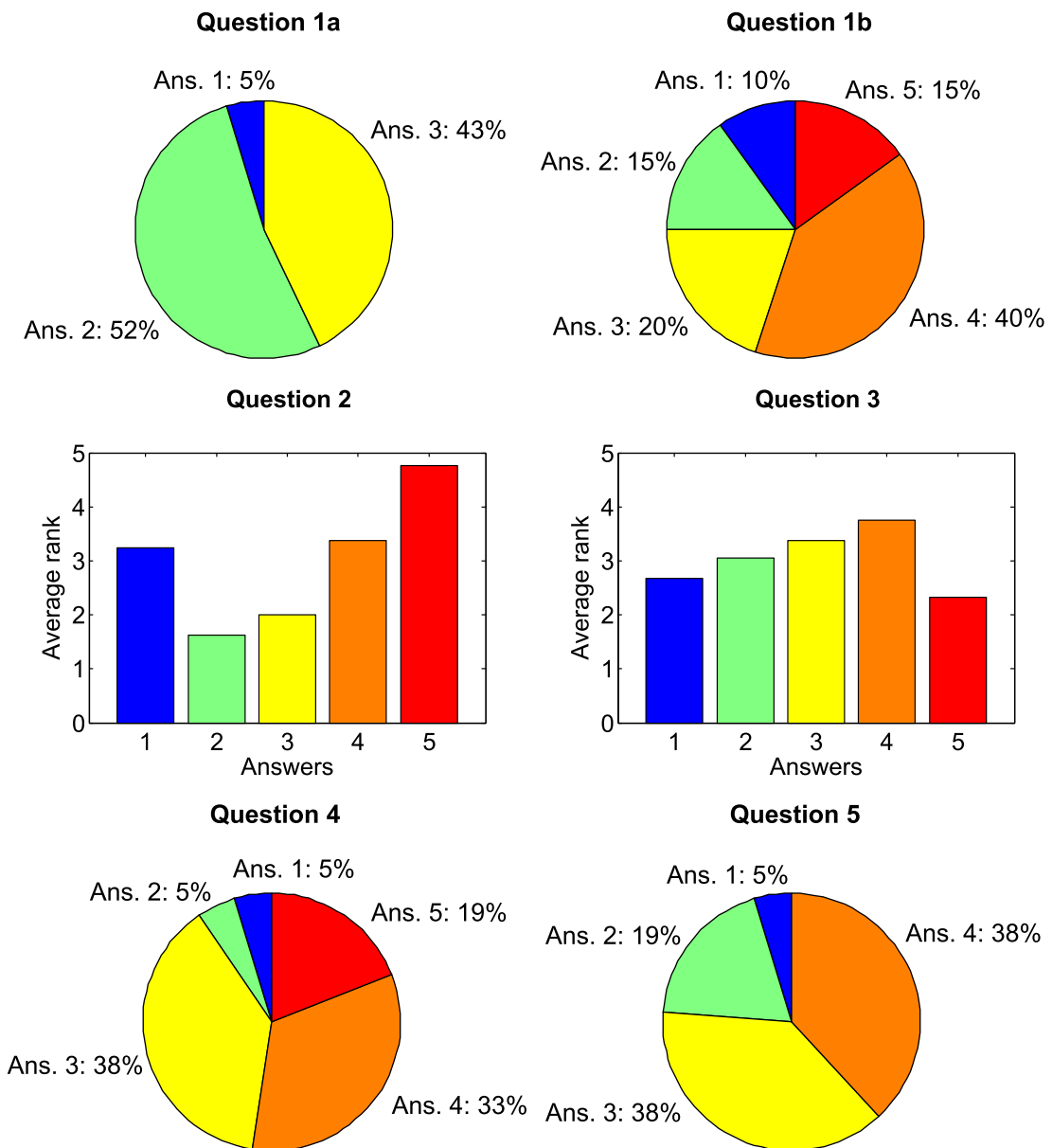


Figure C.1: Results for the survey questions 1–5.

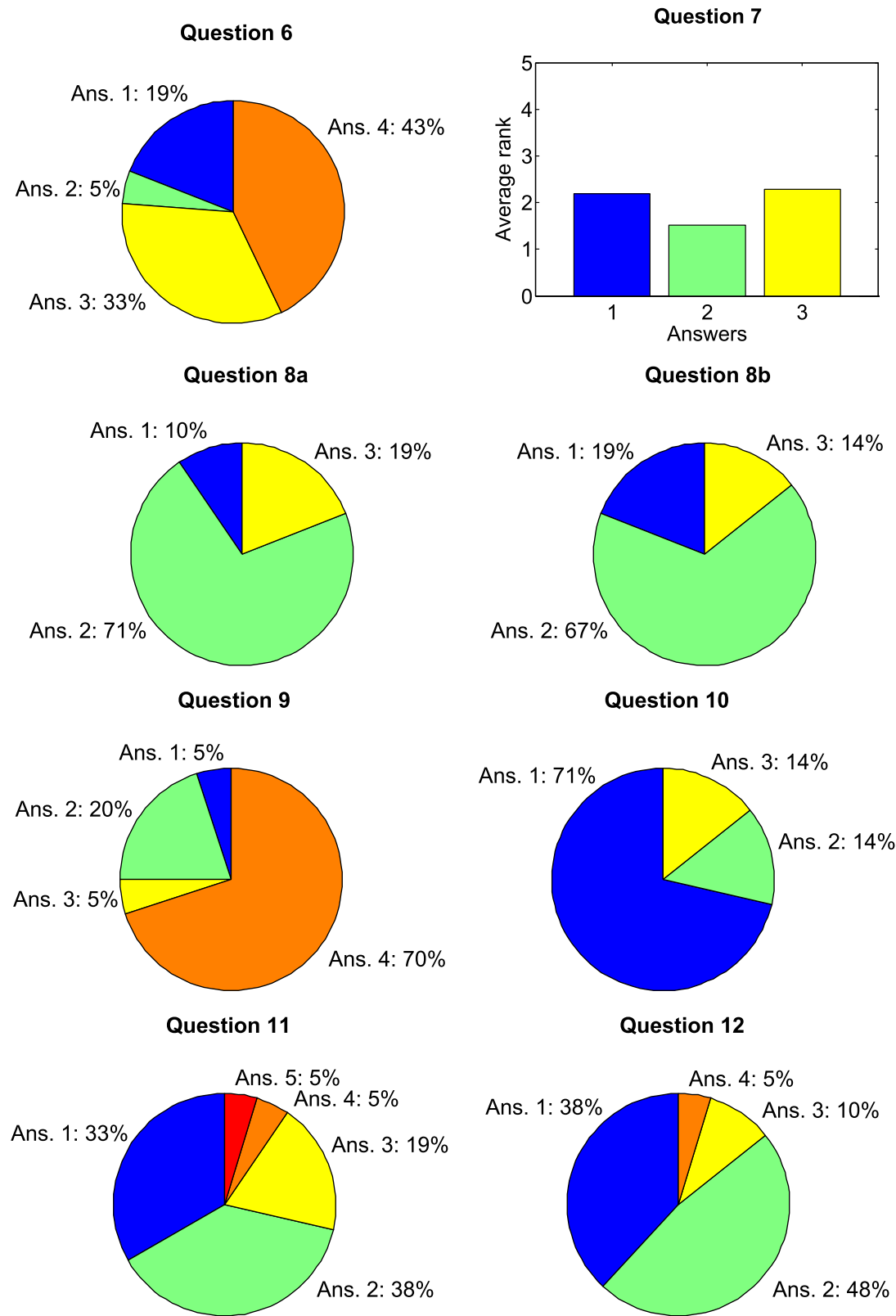


Figure C.1: Results for the survey questions 6–12.

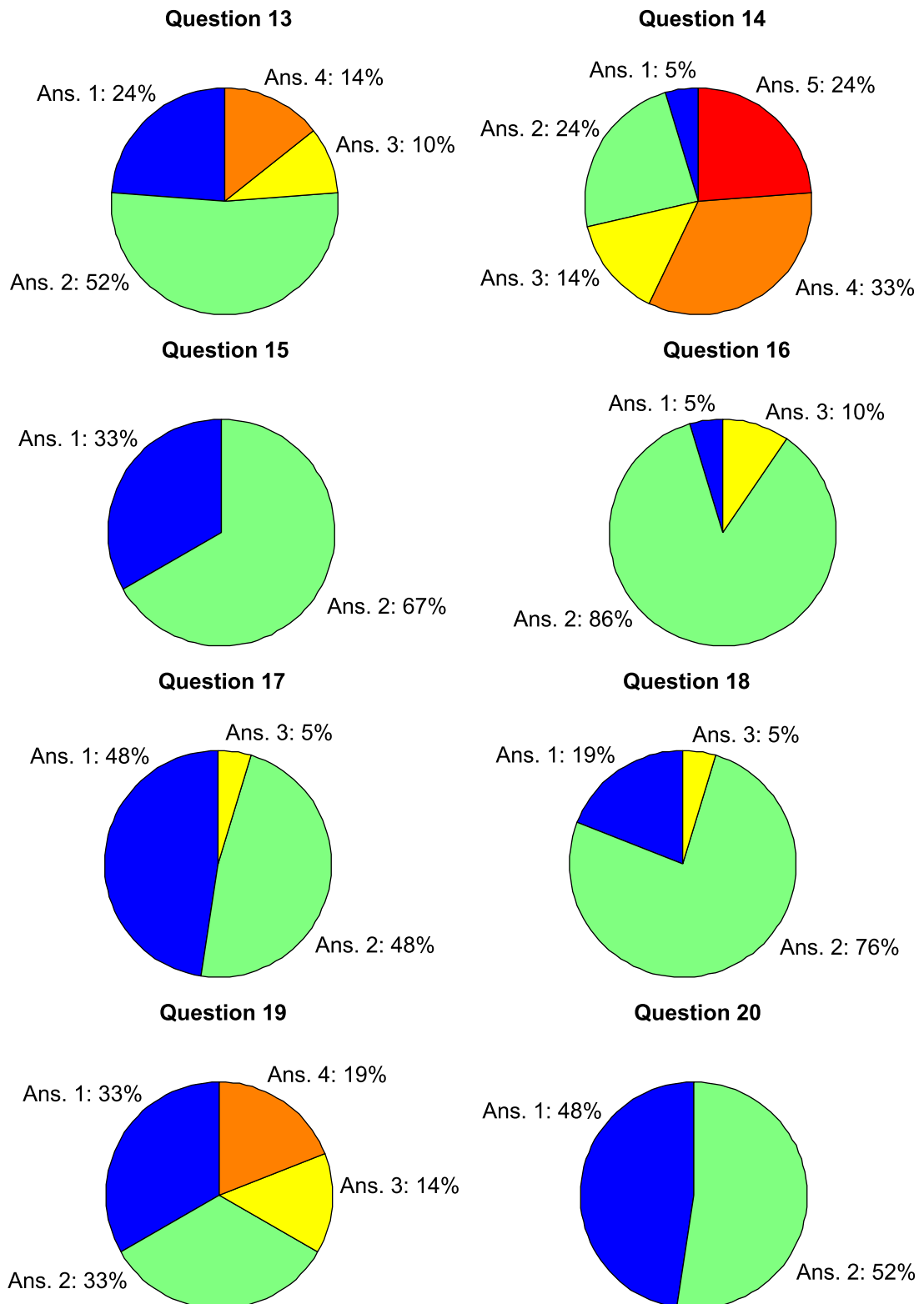


Figure C.1: Results for the survey questions 13–20.

Bibliography

- [1] National Public Radio Inc., “Thomas Friedman on ‘How America Fell Behind’,” Sep. 2011. Available at: <http://www.npr.org/2011/09/06/140214150/thomas-friedman-on-how-america-fell-behind>. (Cited on page 1.)
- [2] D. Nations, “What is social media?,” Feb. 2013. Available at: <http://webtrends.about.com/od/web20/a/social-media.htm>. (Cited on page 1.)
- [3] J. Brenner, “Pew Internet: Social networking (full detail),” Nov. 2012. Available at: <http://pewinternet.org/Commentary/2012/March/Pew-Internet-Social-Networking-full-detail.aspx>. (Cited on page 1.)
- [4] M. Madden and K. Zickuhr, “65 % of online adults use social networking sites,” Aug. 2011. Available at: <http://pewinternet.org/Reports/2011/Social-Networking-Sites.aspx>. (Cited on page 2.)
- [5] S. Laird, “Instagram users share 10 Hurricane Sandy photos per second,” Oct. 2012. Available at: <http://mashable.com/2012/10/29/instagram-hurricane-sandy>. (Cited on page 2.)
- [6] M. Ewing, “71 % more likely to purchase based on social media referrals,” Jan. 2012. Available at: <http://blog.hubspot.com/blog/tabid/6307/bid/30239/71-More-Likely-to-Purchase-Based-on-Social-Media-Referrals-Infographic.aspx>. (Cited on page 2.)
- [7] M. Rhodes, “1 in 4 UK consumers use TripAdvisor before they book their holiday,” Nov. 2010. Available at: <http://www.freshnetworks.com/blog/2010/11/1-in-4-uk-consumers-use-tripadvisor-before-they-book-their-holiday>. (Cited on page 2.)
- [8] L. McDonald, “Infographic: The social network landscape,” Nov. 2012. Available at: <http://www.silverpop.com/blogs/email-marketing/social-network-growth-infographic.html>. (Cited on page 3.)
- [9] Wikimedia Foundation Inc., “Facebook Graph Search,” Feb. 2013. Available at: http://en.wikipedia.org/wiki/Facebook_Graph_Search. (Cited on page 2.)
- [10] Wikimedia Foundation Inc., “Social media,” Jan. 2013. Available at: http://en.wikipedia.org/wiki/Social_media. (Cited on page 4.)

Bibliography

- [11] Wikimedia Foundation Inc., “Web 2.0,” Oct. 2012. Available at: http://en.wikipedia.org/wiki/Web_2.0. (Cited on pages 4 and 123.)
- [12] Nielsen Holdings N.V., “State of the media: The social media report 2012,” Dec. 2012. Available at: <http://blog.nielsen.com/nielsenwire/social/2012/>. (Cited on page 5.)
- [13] C. Marlow, M. Naaman, D. Boyd, and M. Davis, “HT06, tagging paper, taxonomy, Flickr, academic article, to read,” in *Proceedings of the ACM International Conference on Hypertext and Hypermedia*, pp. 31–40, Aug. 2006. (Cited on pages 7 and 96.)
- [14] K. Liu, B. Fang, and Y. Zhang, “Detecting tag spam in social tagging systems with collaborative knowledge,” in *Proceedings of the IEEE International Conference on Fuzzy Systems and Knowledge Discovery*, pp. 427–431, Aug. 2009. (Cited on pages 7, 92, 96, 97, 100, 114, and 116.)
- [15] T. Vander Wal, “Folksonomy coinage and definition,” Feb. 2007. Available at: <http://vanderwal.net/folksonomy.html>. (Cited on page 7.)
- [16] I. Peters and W. G. Stock, “Folksonomy and information retrieval,” *Proceedings of the American Society for Information Science and Technology*, vol. 44, no. 1, pp. 1–28, 2007. (Cited on page 7.)
- [17] D. Benz, A. Hotho, R. Jäschke, B. Krause, F. Mitzlaff, C. Schmitz, and G. Stumme, “The social bookmark and publication management system BibSonomy,” *International Journal on Very Large Data Bases*, vol. 19, no. 6, pp. 849–875, 2010. (Cited on pages 7 and 126.)
- [18] S. Milgram, “The small world problem,” *Psychology Today*, vol. 2, pp. 60–67, 1967. (Cited on page 12.)
- [19] M. E. J. Newman, A.-L. Barabási, and D. J. Watts, eds., *The structure and dynamics of networks*. Princeton studies in complexity, Princeton, Oxford, UK: Princeton University Press, 2011. (Cited on page 12.)
- [20] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature*, vol. 393, no. 6684, pp. 440–442, 1998. (Cited on page 12.)
- [21] M. McGlohon, L. Akoglu, and C. Faloutsos, “Statistical properties of social networks,” in *Social Network Data Analytics* (C. C. Aggarwal, ed.), pp. 17–42, Springer, 2011. (Cited on page 12.)
- [22] X. Xie, “Potential friend recommendation in online social network,” in *Proceedings of the IEEE/ACM International Conference on Cyber, Physical and Social Computing*, pp. 831–835, Dec. 2010. (Cited on page 13.)
- [23] H. Wu, V. Sorathia, and V. Prasanna, “When diversity meets speciality: Friend recommendation in online social networks,” *Human Journal*, vol. 1, no. 1, pp. 52–60, 2012. (Cited on page 13.)

- [24] R. Xiang, J. Neville, and M. Rogati, "Modeling relationship strength in online social networks," in *Proceedings of the ACM International Conference on World Wide Web*, pp. 981–990, Apr. 2010. (Cited on page 13.)
- [25] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Structural Analysis in the Social Sciences, Cambridge, UK: Cambridge University Press, Nov. 1994. (Cited on page 13.)
- [26] B. Prakash, M. Seshadri, A. Sridharan, S. Machiraju, and C. Faloutsos, "EigenSpokes: Surprising patterns and scalable community chipping in large graphs," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 290–295, Dec. 2009. (Cited on page 13.)
- [27] R.-A. Negoescu and D. Gatica-Perez, "Modeling Flickr communities through probabilistic topic-based analysis," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 399–416, 2010. (Cited on page 13.)
- [28] G. A. Fowler, "Facebook: One billion and counting," Oct. 2012. Available at: <http://online.wsj.com/article/SB10000872396390443635404578036164027386112.html>. (Cited on page 13.)
- [29] R. Lawler, "YouTube is launching a redesign to reduce clutter and put videos front and center," Dec. 2012. Available at: <http://techcrunch.com/2012/12/06/youtube-redesign-i-like-videos/>. (Cited on page 13.)
- [30] S. Fiegerman, "Twitter now has more than 200 million monthly active users," Dec. 2012. Available at: <http://mashable.com/2012/12/18/twitter-200-million-active-users/>. (Cited on page 13.)
- [31] A. Efrati, "Google+ announces 135 million users, debuts Instagram competitor," Dec. 2012. Available at: <http://blogs.wsj.com/digits/2012/12/06/google-announces-135-million-users-debuts-instagram-competitor/>. (Cited on page 13.)
- [32] D. Nishar, "LinkedIn: 200 million members!," Jan. 2013. Available at: <http://blog.linkedin.com/2013/01/09/linkedin-200-million/>. (Cited on page 13.)
- [33] A. C. Squicciarini, M. Shehab, and J. Wede, "Privacy policies for shared content in social network sites," *International Journal on Very Large Data Bases*, vol. 19, no. 6, pp. 777–796, 2010. (Cited on page 13.)
- [34] Z. Wang, W. Zhu, P. Cui, L. Sun, and S. Yang, "Social media recommendation," in *Social Media Retrieval* (N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, and X.-S. Hua, eds.), Computer Communications and Networks, pp. 283–304, Springer-Verlag London, Jan. 2013. (Cited on page 13.)
- [35] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–5:60, 2008. (Cited on page 13.)

- [36] C. Taylor, “The most important Facebook number: 140.3 billion,” Oct. 2012. Available at: <http://mashable.com/2012/10/05/the-most-important-facebook-number-140-billion>. (Cited on pages 14, 25, 91, and 164.)
- [37] Grabworthy, “300,000,000 photos uploaded to Facebook daily!,” Jan. 2013. Available at: <http://grabworthy.com/300-million-photos-uploaded-to-facebook-daily/>. (Cited on page 14.)
- [38] Google Inc., “YouTube: Statistics,” Jan. 2013. Available at: http://www.youtube.com/t/press_statistics. (Cited on page 14.)
- [39] S. Sakr, A. Liu, D. Batista, and M. Alomari, “A survey of large scale data management approaches in cloud environments,” *IEEE Communications Surveys Tutorials*, vol. 13, pp. 311–336, Aug.-Sep. 2011. (Cited on page 14.)
- [40] D. Agrawal, S. Das, and A. El Abbadi, “Big data and cloud computing: Current state and future opportunities,” in *Proceedings of the International Conference on Extending Database Technology*, pp. 530–533, Mar. 2011. (Cited on page 14.)
- [41] Wikimedia Foundation Inc., “BigTable,” Jan. 2013. Available at: <http://en.wikipedia.org/wiki/Bigtable>. (Cited on page 14.)
- [42] Wikimedia Foundation Inc., “Apache Cassandra,” Jan. 2013. Available at: http://en.wikipedia.org/wiki/Apache_Cassandra. (Cited on page 14.)
- [43] Wikimedia Foundation Inc., “Apache Hadoop,” Jan. 2013. Available at: http://en.wikipedia.org/wiki/Apache_Hadoop. (Cited on page 14.)
- [44] S. Lohmann, J. Ziegler, and L. Tetzlaff, “Comparison of tag cloud layouts: Task-related performance and visual exploration,” in *Proceedings of the International Conference on Human-Computer Interaction*, pp. 392–404, Aug. 2009. (Cited on page 14.)
- [45] M. Gupta, R. Li, Z. Yin, and J. Han, “An overview of social tagging and applications,” in *Social Network Data Analytics* (C. C. Aggarwal, ed.), pp. 447–497, Springer, 2011. (Cited on page 14.)
- [46] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme, and G. Stumme, “Tag recommendations in folksonomies,” in *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 506–514, Sep. 2007. (Cited on pages 15 and 123.)
- [47] X. Li, C. G. M. Snoek, and M. Worring, “Learning tag relevance by neighbor voting for social image retrieval,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 180–187, Oct. 2008. (Cited on pages 15 and 123.)
- [48] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine,” *Computer Networks and ISDN Systems*, vol. 30, no. 1-7, pp. 107–117, 1998. (Cited on pages 16 and 123.)

-
- [49] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen, "Combating Web spam with TrustRank," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 576–587, Aug. 2004. (Cited on pages 16, 101, and 123.)
- [50] I. Ivanov, P. Vajda, L. Goldmann, J.-S. Lee, and T. Ebrahimi, "Object-based tag propagation for semi-automatic annotation of images," in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 497–506, Mar. 2010. (Cited on pages 16, 17, 18, 41, 56, 123, 179, and 194.)
- [51] I. Ivanov, P. Vajda, J.-S. Lee, L. Goldmann, and T. Ebrahimi, "Geotag propagation in social networks based on user trust model," *Multimedia Tools and Applications*, vol. 56, no. 1, pp. 155–177, 2012. (Cited on pages 17, 18, 92, 97, 98, 100, 107, 112, 116, 118, and 184.)
- [52] I. Ivanov, P. Vajda, J. S. Lee, and T. Ebrahimi, "In tags we trust: Trust modeling in social tagging of multimedia content," *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 98–107, 2012. (Cited on pages 17, 96, and 125.)
- [53] I. Ivanov, P. Vajda, J.-S. Lee, P. Korshunov, and T. Ebrahimi, "Geotag propagation with user trust modeling," in *Social Media Retrieval* (N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, and X.-S. Hua, eds.), Computer Communications and Networks, pp. 283–304, Springer-Verlag London, Jan. 2013. (Cited on page 17.)
- [54] I. Ivanov, P. Vajda, P. Korshunov, and T. Ebrahimi, "Comparative study of trust modeling for automatic landmark tagging," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, 2013. (Cited on pages 17 and 92.)
- [55] S. Yazdani, I. Ivanov, M. Analoui, R. Berangi, and T. Ebrahimi, "Spam fighting in social tagging systems," in *Proceedings of the International Conference on Social Informatics*, pp. 448–461, Dec. 2012. (Cited on page 17.)
- [56] I. Ivanov, P. Vajda, J.-S. Lee, and T. Ebrahimi, "Epitome – A social game for photo album summarization," in *Proceedings of the ACM International Conference on Multimedia, Workshop on Connected Multimedia*, pp. 33–38, Oct. 2010. (Cited on page 17.)
- [57] P. Vajda, I. Ivanov, L. Goldmann, and T. Ebrahimi, "Social game Epitome versus automatic visual analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, Jul. 2011. (Cited on page 17.)
- [58] P. Vajda, I. Ivanov, J.-S. Lee, and T. Ebrahimi, "Epitomize your photos," *International Journal of Computer Games Technology*, vol. 2011, no. 706893, pp. 1–11, 2011. (Cited on page 17.)
- [59] P. Vajda, I. Ivanov, and T. Ebrahimi, "'Cheese' – Towards JPSearch compliant image database," Tech. Rep. JPEG2011/N5739, ISO/IEC JTC1/SC29/WG1, Jan. 2011. (Cited on page 17.)

- [60] I. Ivanov, P. Vajda, L. Goldmann, and T. Ebrahimi, ““Cheese” – A JPSearch Part 4 compliant image database,” Tech. Rep. JPEG2011/N5826, ISO/IEC JTC1/SC29/WG1, Jul. 2011. (Cited on page 17.)
- [61] Yahoo! Inc., “Flickr – All time most popular tags,” Oct. 2012. Available at: <http://www.flickr.com/photos/tags>. (Cited on page 25.)
- [62] B. Sigurbjörnsson and R. van Zwol, “Flickr tag recommendation based on collective knowledge,” in *Proceeding of the ACM International Conference on World Wide Web*, pp. 327–336, Apr. 2008. (Cited on pages 25, 28, and 91.)
- [63] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000. (Cited on page 25.)
- [64] P. Vajda, *Object duplicate detection*. PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Oct. 2011. (Cited on pages 26 and 104.)
- [65] M. Ames and M. Naaman, “Why we tag: Motivations for annotation in mobile and online media,” in *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, pp. 971–980, Apr. 2007. (Cited on pages 26 and 146.)
- [66] L. von Ahn and L. Dabbish, “Labeling images with a computer game,” in *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, pp. 319–326, Apr. 2004. (Cited on pages 26, 144, and 146.)
- [67] L. von Ahn, R. Liu, and M. Blum, “Peekaboom: A game for locating objects in images,” in *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, pp. 55–64, Apr. 2006. (Cited on pages 26 and 146.)
- [68] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, “LabelMe: A database and web-based tool for image annotation,” *International Journal of Computer Vision*, vol. 77, no. 1–3, pp. 157–173, 2008. (Cited on page 26.)
- [69] D. Morrison, S. Marchand-Maillet, and E. Bruno, “TagCaptcha: Annotating images with CAPTCHAs,” in *Proceeding of the ACM International Workshop on Human Computation*, pp. 44–48, Jun. 2009. (Cited on page 26.)
- [70] N. Sawant, J. Li, and J. Z. Wang, “Automatic image semantic interpretation using social action and tagging data,” *Multimedia Tools and Applications*, vol. 51, pp. 213–246, Jan. 2011. (Cited on page 27.)
- [71] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, D. A. Forsyth, and U. C. Berkeley, “Names and faces in the news,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2004)*, pp. 848–854, 2004. (Cited on page 27.)

- [72] S. Ahern, S. King, M. Naaman, R. Nair, and J. H.-I. Yang, “ZoneTag: Rich, community-supported context-aware media capture and annotation,” in *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, pp. 1–4, Apr. 2007. (Cited on page 27.)
- [73] T. Quack, B. Leibe, and L. Van Gool, “World-scale mining of objects and events from community photo collections,” in *Proceedings of the ACM International Conference on Content-based Image and Video Retrieval*, pp. 47–56, Jul. 2008. (Cited on pages 27, 94, and 95.)
- [74] S. Gammeter, L. Bossard, T. Quack, and L. Van Gool, “I know what you did last summer: Object level auto-annotation of holiday snaps,” in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 614–621, Oct. 2009. (Cited on pages 27 and 95.)
- [75] S. Lindstaedt, V. Pammer, R. Mörzinger, R. Kern, H. Mülner, and C. Wagner, “Recommending tags for pictures based on text, visual content and user context,” in *Proceedings of the International Conference on Internet and Web Applications and Services*, pp. 506–511, Jun. 2008. (Cited on page 27.)
- [76] D. Liu, X.-S. Hua, and H.-J. Zhang, “Content-based tag processing for Internet social images,” *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 723–738, 2011. (Cited on page 28.)
- [77] Wikimedia Foundation Inc., “TinEye,” Oct. 2012. Available at: <http://en.wikipedia.org/wiki/TinEye>. (Cited on page 28.)
- [78] H. Jégou, M. Douze, and C. Schmid, “Improving bag-of-features for large scale image search,” *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010. (Cited on page 28.)
- [79] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded-up robust features,” in *Proceedings of the European Conference on Computer Vision*, pp. 404–417, May 2006. (Cited on pages 30, 36, 39, 40, 67, 68, 145, 152, and 194.)
- [80] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. (Cited on pages 30, 36, 66, 67, 103, and 145.)
- [81] K. Mikolajczyk and C. Schmid, “An affine invariant interest point detector,” in *Proceedings of the European Conference on Computer Vision*, pp. 128–142, May 2002. (Cited on pages 31 and 103.)
- [82] D. Nister and H. Stewenius, “Robust scalable recognition with a vocabulary tree,” in *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2161–2168, Jun. 2006. (Cited on pages 32, 35, 103, and 194.)

- [83] B. Girod, V. Chandrasekhar, D. M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. A. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham, “Mobile visual search: Linking the virtual and physical worlds,” *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 61–76, 2011. (Cited on page 32.)
- [84] J. Sivic, F. Schaffalitzky, and A. Zisserman, “Object level grouping for video shots,” *International Journal of Computer Vision*, vol. 67, no. 2, pp. 189–210, 2006. (Cited on page 36.)
- [85] P. Vajda, I. Ivanov, L. Goldmann, J.-S. Lee, and T. Ebrahimi, “Robust duplicate detection of 2D and 3D objects,” *International Journal of Multimedia Data Engineering and Management*, vol. 1, pp. 19–40, Jul.–Sep. 2010. (Cited on pages 36, 102, 104, and 110.)
- [86] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Jun. 2007. (Cited on page 36.)
- [87] D. H. Ballard, “Generalizing the Hough transform to detect arbitrary shapes,” *Pattern Recognition*, vol. 13, no. 2, pp. 111–122, 1981. (Cited on pages 37 and 104.)
- [88] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. (Cited on page 42.)
- [89] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes (VOC) Challenge,” *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010. (Cited on pages 43 and 70.)
- [90] G. Koutrika, F. A. Effendi, Z. Gyöngyi, P. Heymann, and H. Garcia-Molina, “Combating spam in tagging systems: An evaluation,” *ACM Transactions on the Web*, vol. 2, no. 4, pp. 22:1–22:34, 2008. (Cited on pages 44, 92, 97, 99, 116, 123, and 128.)
- [91] S. S. Tsai, D. M. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, and B. Girod, “Fast geometric re-ranking for image-based retrieval,” in *Proceedings of the IEEE International Conference on Image Processing*, pp. 1029–1032, Sep. 2010. (Cited on page 48.)
- [92] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. (Cited on pages 55, 56, 57, 58, 59, 61, 73, and 83.)
- [93] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “Global contrast based salient region detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 409–416, Jun. 2011. (Cited on page 55.)
- [94] J. Fan, D. K. Y. Yau, A. K. Elmagarmid, and W. G. Aref, “Automatic image segmentation by integrating color-edge extraction and seeded region growing,” *IEEE Transactions on Image Processing*, vol. 10, no. 10, pp. 1454–1466, 2001. (Cited on page 55.)

-
- [95] R. Desimone and J. Duncan, “Neural mechanisms of selective visual attention,” *Annual Review of Neuroscience*, vol. 18, no. 1, pp. 193–222, 1995. (Cited on page 55.)
- [96] T. Judd, F. Duranda, and A. Torralba, “Fixations on low-resolution images,” *Journal of Vision*, vol. 11, no. 4, pp. 1–20, 2011. (Cited on page 55.)
- [97] J.-S. Lee, F. D. Simone, and T. Ebrahimi, “Efficient video coding based on audio-visual focus of attention,” *Journal of Visual Communication and Image Representation*, vol. 22, no. 8, pp. 704–711, 2011. (Cited on page 55.)
- [98] U. Rutishauser, D. Walther, C. Koch, and P. Perona, “Is bottom-up attention useful for object recognition?,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. II–37–II–44, Jun. 2004. (Cited on page 55.)
- [99] C. Siagian and L. Itti, “Rapid biologically-inspired scene classification using features shared with visual attention,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 300–312, 2007. (Cited on page 56.)
- [100] H. Yu, J. Li, Y. Tian, and T. Huang, “Automatic interesting object extraction from images using complementary saliency maps,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 891–894, Oct. 2010. (Cited on page 56.)
- [101] J. M. Wolfe and T. S. Horowitz, “What attributes guide the deployment of visual attention and how do they do it?,” *Nature Reviews Neuroscience*, vol. 5, no. 6, pp. 495–501, 2004. (Cited on page 56.)
- [102] S. K. Mannan, C. Kennard, and M. Husain, “The role of visual salience in directing eye movements in visual object agnosia,” *Current Biology*, vol. 19, no. 6, pp. R247–R248, 2009. (Cited on page 56.)
- [103] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. PrePrints, 2012. (Cited on pages 57 and 59.)
- [104] R. Achanta, S. Hemami, F. Estrada, and S. Ssstrunk, “Frequency-tuned salient region detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1597–1604, Jun. 2009. (Cited on pages 57, 58, 60, 63, 71, 72, 73, and 83.)
- [105] R. Achanta, F. Estrada, P. Wils, and S. Ssstrunk, “Salient region detection and segmentation,” in *Proceedings of the International Conference on Computer Vision Systems*, pp. 66–75, May 2008. (Cited on page 57.)
- [106] S. Frintrop, M. Klodt, and E. Rome, “A real-time visual attention system using integral images,” in *Proceedings of the International Conference on Computer Vision Systems*, pp. 1–10, Mar. 2007. (Cited on page 57.)

Bibliography

- [107] Y.-F. Ma and H.-J. Zhang, “Contrast-based image attention analysis by using fuzzy growing,” in *Proceedings of the ACM International Conference on Multimedia*, pp. 374–381, Nov. 2003. (Cited on page 57.)
- [108] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Jun. 2007. (Cited on page 57.)
- [109] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems*, vol. 19, pp. 545–552, Dec. 2006. (Cited on pages 57, 58, 60, 61, 73, and 83.)
- [110] L. Itti and C. Koch, “Computational modelling of visual attention,” *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001. (Cited on page 59.)
- [111] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002. (Cited on page 63.)
- [112] C. M. Christoudias, B. Georgescu, and P. Meer, “Synergism in low level vision,” in *Proceedings of the IEEE International Conference on Pattern Recognition*, pp. IV–150–IV–155, Aug. 2002. (Cited on page 63.)
- [113] T. Ojala, M. Pietikäinen, and D. Harwood, “A comparative study of texture measures with classification based on featured distributions,” *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996. (Cited on page 64.)
- [114] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, “Evaluation of gist descriptors for web-scale image search,” in *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 19:1–19:8, Jul 2009. (Cited on page 64.)
- [115] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “BRIEF: Binary robust independent elementary features,” in *Proceedings of the European Conference on Computer Vision*, pp. 778–792, Sep. 2010. (Cited on page 64.)
- [116] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, “Fast keypoint recognition using random ferns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 448–461, 2010. (Cited on page 64.)
- [117] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, “Compressed histogram of gradients: A low-bitrate descriptor,” *International Journal of Computer Vision*, vol. 96, no. 3, pp. 384–399, 2012. (Cited on page 64.)
- [118] M. J. Swain and D. H. Ballard, “Color indexing,” *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991. (Cited on page 64.)

-
- [119] ISO/IEC JTC1/SC29/WG11, “ISO/IEC 15938-3: Information technology – Multimedia content description interface – Part 3: Visual,” tech. rep., International Organization for Standardization, 2002. (Cited on page 64.)
- [120] ISO/IEC JTC1/SC29/WG11, “ISO/IEC 15938-5: Information technology – Multimedia content description interface – Part 5: Multimedia description schemes,” tech. rep., International Organization for Standardization, 2002. (Cited on page 64.)
- [121] ISO/IEC JTC1/SC29/WG11, “ISO/IEC 15938-8: Information technology – Multimedia content description interface – Part 8: Extraction and use of MPEG-7 descriptions,” tech. rep., International Organization for Standardization, 2002. (Cited on page 64.)
- [122] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 703–715, 2001. (Cited on page 64.)
- [123] H. G. Barrow and J. Tenenbaum, “Interpreting line drawings as three-dimensional surfaces,” *Artificial Intelligence*, vol. 17, no. 1–3, pp. 75–116, 1981. (Cited on page 66.)
- [124] J. Canny, “A computational approach to edge detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, 1986. (Cited on page 66.)
- [125] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99–109, 1943. (Cited on page 66.)
- [126] R. C. Gonzalez and R. E. Woods, eds., *Digital Image Processing*. Upper Saddle River, New Jersey, USA: Pearson Prentice Hall, 3 ed., 2008. (Cited on page 66.)
- [127] A. K. Jain and F. Farrokhnia, “Unsupervised texture segmentation using Gabor filters,” *Pattern Recognition*, vol. 24, no. 12, pp. 1167–1186, 1991. (Cited on page 66.)
- [128] J. S. Beis and D. G. Lowe, “Shape indexing using approximate nearest-neighbour search in high-dimensional spaces,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognitions*, pp. 1000–1006, Jun. 1997. (Cited on page 67.)
- [129] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, “A comparison of affine region detectors,” *International Journal of Computer Vision*, vol. 65, no. 1–2, pp. 43–72, 2005. (Cited on page 67.)
- [130] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. I–886–I–893, Jun. 2005. (Cited on pages 68 and 152.)
- [131] C. Fellbaum, ed., *WordNet: An electronic lexical database*. Cambridge, Massachusetts, USA; London, UK: The MIT Press, May 1998. (Cited on pages 91 and 98.)

Bibliography

- [132] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, “To search or to label?: Predicting the performance of search-based automatic image classifiers,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 249–258, Oct. 2006. (Cited on page 91.)
- [133] P. Zurek, “GPS data problem on Flickr,” Feb. 2013. Available at: <http://www.flickr.com/photos/piotrzurek/1016089766/>. (Cited on page 93.)
- [134] Funsherpa.com, “The impact of the Internet and social media on travel,” Jul. 2012. Available at: <http://blog.funsherpa.com/2012/07/travel-infographic>. (Cited on page 92.)
- [135] Z. Xu, Y. Fu, J. Mao, and D. Su, “Towards the semantic Web: Collaborative tag suggestions,” in *Proceeding of the ACM International Conference on World Wide Web*, pp. 1—8, May 2006. (Cited on pages 92, 97, 101, 102, 114, and 116.)
- [136] R. Krestel and L. Chen, “Using co-occurrence of tags and resources to identify spammers,” in *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 38–46, Sep. 2008. (Cited on pages 92, 97, 101, 108, and 114.)
- [137] J. Luo, D. Joshi, J. Yu, and A. Gallagher, “Geotagging in multimedia and computer vision – A survey,” *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 187–211, 2011. (Cited on page 94.)
- [138] J. Hays and A. A. Efros, “IM2GPS: Estimating geographic information from a single image,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Jun. 2008. (Cited on page 94.)
- [139] L. S. Kennedy and M. Naaman, “Generating diverse and representative image search results for landmarks,” in *Proceedings of the ACM International Conference on World Wide Web*, pp. 297–306, Apr. 2008. (Cited on pages 94 and 95.)
- [140] Y. T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T. Chua, and H. Neven, “Tour the world: Building a web-scale landmark recognition engine,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1085–1092, Jun. 2009. (Cited on pages 94 and 95.)
- [141] Technical Standardization Committee on AV & IT Storage Systems and Equipment, “Exchangeable image file format for digital still cameras: Exif Version 2.2,” Tech. Rep. JEITA CP-3451, Japan Electronics and Information Technology Industries Association, Apr. 2002. (Cited on pages 95 and 195.)
- [142] IPTC Standards Committee, “IPTC photo metadata standard, IPTC Core 1.1 and IPTC Extension 1.1,” tech. rep., International Press Telecommunications Council, Jul. 2010. (Cited on page 95.)

-
- [143] L. Hollenstein and R. Purves, “Exploring place through user-generated content: Using Flickr to describe city cores,” *Journal of Spatial Information Science*, vol. 1, no. 1, pp. 21–48, 2010. (Cited on page 96.)
- [144] A. Budanitsky and G. Hirst, “Evaluating WordNet-based measures of lexical semantic relatedness,” *Comput. Linguist.*, vol. 32, pp. 13–47, Mar. 2006. (Cited on page 98.)
- [145] R. L. Cilibrasi and P. M. B. Vitanyi, “The Google similarity distance,” *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, pp. 370–383, Mar. 2007. (Cited on page 98.)
- [146] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999. (Cited on page 101.)
- [147] T. Bogers and A. Van den Bosch, “Using language models for spam detection in social bookmarking,” in *Proceedings of the ECML/PKDD Discovery Challenge Workshop*, pp. 1–12, Sep. 2008. (Cited on pages 123, 126, 130, and 135.)
- [148] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, “A Bayesian approach to filtering junk e-mail,” Tech. Rep. WS-98-05, AAAI-98 Workshop on Learning for Text Categorization, Jul. 1998. (Cited on page 124.)
- [149] D. Fetterly, M. Manasse, and M. Najork, “Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages,” in *Proceedings of the International Workshop on the Web and Databases*, pp. 1–6, Jun. 2004. (Cited on page 124.)
- [150] S. Marti and H. Garcia-Molina, “Taxonomy of trust: Categorizing P2P reputation systems,” *International Journal of Computer and Telecommunications Networking*, vol. 50, no. 4, pp. 472–484, 2006. (Cited on page 124.)
- [151] A. Thomason, “Blog spam: A review,” in *Proceedings of the Conference on Email and Anti-Spam*, pp. 1–4, Aug. 2007. (Cited on page 124.)
- [152] A. Jøsang, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” *Decision Support Systems*, vol. 43, no. 2, pp. 618–644, 2007. (Cited on page 125.)
- [153] A. Whitby, A. Jøsang, and J. Indulska, “Filtering out unfair ratings in bayesian reputation systems,” in *Proceedings of the International Workshop on Trust in Agent Societies*, pp. 106–117, Jul. 2004. (Cited on page 125.)
- [154] Y. Yang, Y. L. Sun, S. Kay, and Q. Yang, “Defending online reputation systems against collaborative unfair raters through signal modeling and trust,” in *Proceedings of the ACM Symposium on Applied Computing*, pp. 1308–1315, Mar. 2009. (Cited on page 125.)
- [155] P. Heymann, G. Koutrika, and H. Garcia-Molina, “Fighting spam on social web sites: A survey of approaches and future challenges,” *IEEE Internet Computing*, vol. 11, pp. 36–45, Nov. 2007. (Cited on page 125.)

Bibliography

- [156] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford, “CAPTCHA: Using hard AI problems for security,” in *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 294–311, May 2003. (Cited on page 125.)
- [157] L. von Ahn, B. Maurer, C. Mcmillen, D. Abraham, and M. Blum, “reCAPTCHA: Human-based character recognition via Web security measures,” *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008. (Cited on page 125.)
- [158] Yahoo! Inc., “Flickr – Tags,” Oct. 2012. Available at: <http://www.flickr.com/help/tags>. (Cited on page 125.)
- [159] G. Mori and J. Malik, “Recognizing objects in adversarial clutter: Breaking a visual CAPTCHA,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. I–134–I–141, Jun. 2003. (Cited on page 125.)
- [160] B. Markines, C. Cattuto, and F. Menczer, “Social spam detection,” in *Proceedings of the International Workshop on Adversarial Information Retrieval on the Web*, pp. 41–48, Apr. 2009. (Cited on pages 125, 126, and 135.)
- [161] B. Krause, C. Schmitz, A. Hotho, and S. G., “The anti-social tagger: Detecting spam in social bookmarking systems,” in *Proceedings of the International Workshop on Adversarial Information Retrieval on the Web*, pp. 61–68, Apr. 2008. (Cited on pages 125 and 126.)
- [162] B. Matthews, “Comparison of the predicted and observed secondary structure of T4 phage lysozyme,” *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975. (Cited on page 132.)
- [163] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, California, USA: Morgan Kaufmann Publishers Inc., 2 ed., 2005. (Cited on page 132.)
- [164] H. Liu and R. Setiono, “Chi2: Feature selection and discretization of numeric attributes,” in *Proceedings of the International Conference on Tools with Artificial Intelligence*, pp. 388–391, Nov. 1995. (Cited on page 133.)
- [165] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York, New York, USA: Wiley-Interscience, 2 ed., 2000. (Cited on pages 133 and 134.)
- [166] L. Rainie, J. Brenner, and K. Purcell, “Photos and videos as social currency online,” Sep. 2012. Available at: <http://pewinternet.org/Reports/2012/Online-Pictures.aspx>. (Cited on page 143.)
- [167] F. Frey, M. Rodriguez Adames, Y. fang Tsai, F. Cost, and S. Farnand, “Print versus screen – Presentation medium-dependent picture consumption,” tech. rep., Printing Industry Center, Rochester, NY, Sep. 2010. (Cited on page 143.)
- [168] Wikimedia Foundation Inc., “A picture is worth a thousand words,” Feb. 2013. Available at: http://en.wikipedia.org/wiki/A_picture_is_worth_a_thousand_words. (Cited on page 143.)

- [169] C. G. M. Snoek, “HP Challenge 2010: High Impact Visual Communication,” Feb. 2010. Available at: <http://comminfo.rutgers.edu/conferences/mmchallenge/2010/02/10/hp-challenge-2010>. (Cited on page 143.)
- [170] M. Snider, “Social games: news and survey findings,” Feb. 2010. Available at: <http://content.usatoday.com/communities/gamehunters/post/2010/02/social-games-news-and-survey-findings/1>. (Cited on pages 143 and 164.)
- [171] S. Harada, M. Naaman, Y. J. Song, Q. Wang, and A. Paepcke, “Lost in memories: Interacting with large photo collections on PDAs,” in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pp. 325–333, Jun. 2004. (Cited on page 145.)
- [172] M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina, “Automatic organization for digital photographs with geographic coordinates,” in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pp. 53–62, Apr. 2004. (Cited on page 145.)
- [173] B. C. Atkins, “Adaptive photo collection page layout,” in *Proceedings of the IEEE International Conference on Image Processing*, pp. 2897–2900, Oct. 2004. (Cited on page 145.)
- [174] J. Geigel and A. Loui, “Using genetic algorithms for album page layouts,” *IEEE Multimedia*, vol. 10, no. 4, pp. 16–27, 2003. (Cited on page 145.)
- [175] M. Rabbath, P. Sandhaus, and S. Boll, “Automatic creation of photo books from stories in social media,” *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 7S, no. 1, pp. 27:1–27:18, 2011. (Cited on page 145.)
- [176] L. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2:524–2:531, Jun. 2005. (Cited on pages 145 and 152.)
- [177] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid, “Local features and kernels for classification of texture and object categories: A comprehensive study,” *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007. (Cited on page 145.)
- [178] P. Felzenszwalb, D. McAllester, and D. Ramanan, “A discriminatively trained, multiscale, deformable part model,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2008)*, pp. 1–8, 2008. (Cited on page 145.)
- [179] A. Torralba, R. Fergus, and W. T. Freeman, “80 million tiny images: A large data set for nonparametric object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008. (Cited on page 145.)
- [180] E. Law and L. von Ahn, “Input-agreement: A new mechanism for collecting data using human computation games,” in *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, pp. 1197–1206, Apr. 2009. (Cited on page 146.)

Bibliography

- [181] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, M. Jaskolski, and D. Baker, “Crystal structure of a monomeric retroviral protease solved by protein folding game players,” *Nature Structural & Molecular Biology*, vol. 18, no. 10, pp. 1175–1177, 2011. (Cited on page 146.)
- [182] A. Waheed, “Top 10 most popular games to play on Facebook,” Jan. 2013. Available at: <http://www.top10facts.com/2013/01/top-10-most-popular-games-to-play-on-facebook>. (Cited on page 164.)
- [183] C. H. Lampert, M. B. Blaschko, and T. Hofmann, “Beyond sliding windows: Object localization by efficient subwindow search,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, Jun. 2008. (Cited on page 181.)
- [184] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, “Learning to detect a salient object,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011. (Cited on page 181.)
- [185] M. J. Huiskes, B. Thomee, and M. S. Lew, “New trends and ideas in visual concept detection: The MIR Flickr retrieval evaluation initiative,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 527–536, Mar. 2010. (Cited on pages 182 and 189.)
- [186] M. J. Huiskes and M. S. Lew, “The MIR Flickr retrieval evaluation,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, Oct. 2008. (Cited on page 182.)
- [187] A. Carranza, “Instagram: The popularity and influence of photos,” Jan. 2013. Available at: <http://www.examiner.com/article/the-impact-and-influence-of-instagram-photo-phenomenon>. (Cited on page 189.)
- [188] R. Tous, J. Nin, J. Delgado, and P. Toran, “Approaches and standards for metadata interoperability in distributed image search and retrieval,” in *Database and Expert Systems Applications* (A. Hameurlain, S. W. Liddle, K.-D. Schewe, and X. Zhou, eds.), Lecture Notes in Computer Science, pp. 234–248, Springer Berlin Heidelberg, 2011. (Cited on page 189.)
- [189] R. van Zwol, “Flickr: Who is looking?,” in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 184–190, Nov. 2007. (Cited on page 189.)
- [190] ISO/IEC JTC1/SC29/WG1, “ISO/IEC TR 24800-1: Information technology – JPSearch – Part 1: System framework and components,” tech. rep., International Organization for Standardization, 2007. (Cited on pages 190 and 193.)
- [191] ISO/IEC JTC1/SC29/WG1, “ISO/IEC 24800-4: Information technology – JPSearch – Part 4: File format for metadata embedded in image data (JPEG and JPEG 2000),” tech. rep., International Organization for Standardization, 2010. (Cited on pages 190, 192, 195, and 196.)

-
- [192] F. Temmermans, F. Dufaux, and P. Schelkens, "JPSearch: Metadata interoperability during image exchange," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 134–139, 2012. (Cited on pages 190, 191, 192, 193, and 197.)
- [193] J. M. Martínez, R. Koenen, and F. Pereira, "MPEG-7: The generic multimedia content description standard, part 1," *IEEE MultiMedia*, vol. 9, no. 2, pp. 78–87, 2002. (Cited on page 190.)
- [194] S.-F. Chang, T. Sikora, and A. Puri, "Overview of the MPEG-7 standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 688–695, 2001. (Cited on page 190.)
- [195] J. R. Smith, M. Döller, R. Tous, M. Gruhne, K. Yoon, M. Sano, and I. S. Burnett, "The MPEG Query Format: Unifying access to multimedia retrieval systems," *IEEE MultiMedia*, vol. 15, no. 4, pp. 82–95, 2008. (Cited on page 191.)
- [196] Dublin Core Metadata Initiative, "Dublin Core metadata element set, version 1.1," Dec. 2012. Available at: <http://dublincore.org/documents/dces/>. (Cited on page 191.)
- [197] World Wide Web Consortium, "Resource Description Framework (RDF)," Jan. 2013. Available at: <http://www.w3.org/RDF/>. (Cited on page 191.)
- [198] P. Toran and J. Delgado, "Image search based on a broker approach," in *Proceedings of the International Workshop on Multimedia Metadata Community*, pp. 166–172, May 2010. (Cited on page 191.)
- [199] R. Tous, J. Delgado, T. Zinkl, P. Toran, G. Alcalde, M. Goetz, and O. Ferrer Roca, "The anatomy of an optical biopsy semantic retrieval system," *IEEE MultiMedia*, vol. 19, no. 2, pp. 16–27, 2012. (Cited on page 191.)
- [200] F. Temmermans, B. Jansen, R. Deklerck, P. Schelkens, and J. Cornelis, "The mobile museum guide: Artwork recognition with eigenpaintings and SURF," in *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1–4, Apr. 2011. (Cited on page 191.)
- [201] F. Dufaux, M. Ansorge, and T. Ebrahimi, "Overview of JPSearch: A standard for image search and retrieval," in *Proceedings of the International Workshop on Content-based Multimedia Indexing*, pp. 138–143, Jul. 2007. (Cited on page 192.)
- [202] T. Ebrahimi, A. Skodras, and P. Schelkens, "Ongoing standardization efforts," in *The JPEG 2000 Suite* (P. Schelkens, A. Skodras, and T. Ebrahimi, eds.), pp. 481–489, Chichester, UK: John Wiley & Sons, Ltd., 2009. (Cited on page 192.)
- [203] K. Yoon, Y. Kim, J.-H. Park, J. Delgado, A. Yamada, F. Dufaux, and R. Tous, "JPSearch: New international standard providing interoperable framework for image search and sharing," *Signal Processing: Image Communication*, vol. 27, no. 7, pp. 709–721, 2012. (Cited on pages 192 and 193.)

Bibliography

- [204] ISO/IEC JTC1/SC29/WG1, “ISO/IEC 24800-3: Information technology – JPSearch – Part 3: Query format, TECHNICAL CORRIGENDUM 1 ,” tech. rep., International Organization for Standardization, 2010. (Cited on page 192.)
- [205] ISO/IEC JTC1/SC29/WG1, “ISO/IEC 24800-5: Information technology – JPSearch – Part 5: Data interchange format between image repositories,” tech. rep., International Organization for Standardization, 2011. (Cited on page 193.)
- [206] ISO/IEC JTC1/SC29/WG1, “ISO/IEC 24800-2: Information technology – JPSearch – Part 2: Registration, identification and management of schema and ontology,” tech. rep., International Organization for Standardization, 2011. (Cited on pages 193 and 196.)
- [207] ISO/IEC JTC1/SC29/WG1, “ISO/IEC 24800-6: Information technology – JPSearch – Part 6: Reference software,” tech. rep., International Organization for Standardization, 2012. (Cited on page 193.)
- [208] M. Doeller, R. Tous, F. Temmermans, K. Yoon, J. Park, Y. Kim, F. Stegmaier, and J. Delgado, “JPEG’s JPSearch standard: Harmonizing image management and search,” *IEEE MultiMedia*, vol. PP, no. 99, p. 1, 2012. (Cited on pages 193 and 199.)
- [209] Creative Commons, “Creative Commons license: Attribution-NonCommercial-NoDerivs 3.0 Unported – CC BY-NC-ND 3.0,” Jan. 2013. Available at: <http://creativecommons.org/licenses/by-nc-nd/3.0/>. (Cited on page 194.)
- [210] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*. Norwell, Massachusetts, USA: Kluwer Academic Publishers, 1 ed., 1992. (Cited on page 195.)
- [211] D. S. Taubman and M. W. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Norwell, Massachusetts, USA: Kluwer Academic Publishers, 2001. (Cited on page 195.)



Ivan Ivanov

CONTACT INFORMATION


Work address:

EPFL STI IEL GR-EB
ELE 231, Station 11
1015 Lausanne, Switzerland

Phone: +41 21 693 4714

E-mail: ivan.ivanov@epfl.ch

Web: <http://people.epfl.ch/ivan.ivanov>

 <http://www.linkedin.com/in/ivanivanov>

Home address:

Rue de Genève 79
1004 Lausanne, Switzerland
Mobile phone: +41 78 762 1464
E-mail: ivicaivanov@gmail.com

Born on October 2nd, 1982
in Bor, Serbia

KEY COMPETENCES

- Excellent **technical** and **analytical skills**, with strong interest in **solving technical problems**
- Experienced in **social media analysis** and **multimedia processing** – **PhD degree** from **EPFL**
- Comprehensive knowledge of IT (**hardware systems development** and **mobile network deployment**)
- Open-minded, reliable, persistent, organized, creative, ambitious, quick learner

EDUCATION

04/2008 – present **Doctor of Science in Electrical Engineering (PhD), major: Multimedia Signal Processing**

School of Engineering, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

09/2001 – 09/2006 **Diploma Engineer in Electrical Engineering (equiv. to Master of Science), major: Electronics**

School of Electrical Engineering, University of Belgrade, Belgrade, Serbia,
GPA: 9.69/10.00 (top 1 %)

PROFESSIONAL EXPERIENCE

- 04/2008 – present **Research Assistant**
Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
- Conducted research in various fields of social media analysis, multimedia processing (image and video), computer vision, and metadata management.
 - Designed and developed, in the framework of Swiss and EU funded projects:
 - An efficient photo-sharing web application “Cheese” (<http://cheese.epfl.ch>) for automatic image annotation in large-scale database of over 1 million images (**the world’s first** platform to be JPSearch - Part 4 compliant; **Best Poster Award** at the S3MR, Turkey, in 2011).
 - An innovative social game “Epitome” (<http://apps.facebook.com/epitome>) as Facebook and Android OS application for photo album summarization (**finalist** of the Multimedia Grand Challenge, Italy, in 2010). Carried out user-satisfaction survey, compiling results into detailed reports.
 - A tool for crawling Twitter social network. Collected and analyzed a large-scale set of 10 million users, their relationships and associated metadata.
 - Published 15 research papers in highly reputed international journals and conferences, including IEEE Signal Processing Magazine (**ranked first** among all Electrical & Electronic Engineering journals).
 - Acted as advisor and reviewer of papers for leading international journals in multimedia signal processing: Multimedia Tools and Applications, IEEE Transactions on Image Processing, IEEE Journal of Selected Topics in Signal Processing, IEEE Transactions on Multimedia.
 - Supervised and supported 6 students in their theses and semester projects at Master level.
 - Organized laboratory sessions for Master’s level courses: Image and Video Processing, and Multimedia Security. Conducted correction and grading of laboratory reports.
- 12/2007 – 03/2008 **Radio Access Network Conceptual Planning Expert**
Vip mobile d.o.o. (member of Telekom Austria Group), Belgrade, Serbia
- Directed definition and production of a conceptual plan of voice and data mobile network, which assured reliable and timely serving of new subscribers in the initial rollout of mobile services.
- 06/2006 – 04/2007 **Digital Design Engineer**
Texas Instruments, Nice, France
- Participated in a long-term and schedule-driven project on the development of OMAPV1035 single-chip for low-power multimedia devices (smartphones, tablets), the first fully-integrated digital baseband, RF and applications processor.
 - Managed integration of heterogeneous power domains at the chip’s top level, and dealt with large number of signals and Intellectual Property modules.
- 04/2006 – 10/2007 **Hardware Design Engineer**
ELSYS Eastern Europe (subsidiary of ELSYS Design France), Belgrade, Serbia
- Worked as a consultant within a hardware design team for Texas Instruments, France.

AWARDS AND HONORS

- 07/2011 **Best Poster Award** for the poster “Swiss Cheese for Visual Search” at the 2nd Summer School on Social Media Retrieval (S3MR), Turkey.
- 10/2010 **Finalist** of the Multimedia Grand Challenge with the presentation “Epitome – A Social Game for Photo Album Summarization” at the ACM International Conference on Multimedia, Italy.
- 12/2005 Awarded as **outstanding student** with one year scholarship by Fund for Young Talents, The Government of the Republic of Serbia.
- 12/2002 Ranked **1st in the first-year student class** in School of Electrical Engineering, University of Belgrade, Serbia.
- 05/2001 Won **4th place** at Serbia’s **national programming competition**.
- 03/2001 Won **2nd place** at Serbia’s **national mathematics competition**.

PUBLICATIONS

Journal articles

- [1] I. Ivanov, P. Vajda, P. Korshunov, and T. Ebrahimi, “Comparative study of trust modeling for automatic landmark tagging,” *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, 2013.
- [2] I. Ivanov, P. Vajda, J. S. Lee, and T. Ebrahimi, “In tags we trust: Trust modeling in social tagging of multimedia content,” *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 98–107, 2012.
- [3] I. Ivanov, P. Vajda, J.-S. Lee, L. Goldmann, and T. Ebrahimi, “Geotag propagation in social networks based on user trust model,” *Multimedia Tools and Applications*, vol. 56, no. 1, pp. 155–177, 2012.
- [4] P. Vajda, I. Ivanov, J.-S. Lee, and T. Ebrahimi, “Epitomize your photos,” *International Journal of Computer Games Technology*, vol. 2011, no. 706893, pp. 1–11, 2011.
- [5] P. Vajda, I. Ivanov, L. Goldmann, J.-S. Lee, and T. Ebrahimi, “Robust duplicate detection of 2D and 3D objects,” *International Journal of Multimedia Data Engineering and Management*, vol. 1, pp. 19–40, Jul.–Sep. 2010.

Book chapters

- [1] I. Ivanov, P. Vajda, J.-S. Lee, P. Korshunov, and T. Ebrahimi, “Geotag propagation with user trust modeling,” in *Social Media Retrieval* (N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, and X.-S. Hua, eds.), Computer Communications and Networks, pp. 283–304, Springer-Verlag London, Jan. 2013.

Conference papers

- [1] S. Yazdani, I. Ivanov, M. Analoui, R. Berangi, and T. Ebrahimi, “Spam fighting in social tagging systems,” in *Proceedings of the International Conference on Social Informatics*, pp. 448–461, Dec. 2012.
- [2] P. Vajda, I. Ivanov, L. Goldmann, and T. Ebrahimi, “Social game Epitome versus

- automatic visual analysis,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, Jul. 2011.
- [3] P. Vajda, I. Ivanov, L. Goldmann, and T. Ebrahimi, “Let Epitome summarize your photo collection!,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Jul. 2011. demo paper.
 - [4] P. Vajda, I. Ivanov, L. Goldmann, and T. Ebrahimi, “Omnidirectional object duplicate detection,” in *Proceedings of the IEEE Digital Signal Processing Workshop and the IEEE Signal Processing Education Workshop*, pp. 332–337, Jan. 2011.
 - [5] I. Ivanov, P. Vajda, J.-S. Lee, and T. Ebrahimi, “Epitome – A social game for photo album summarization,” in *Proceedings of the ACM International Conference on Multimedia, Workshop on Connected Multimedia*, pp. 33–38, Oct. 2010.
 - [6] P. Vajda, I. Ivanov, J.-S. Lee, L. Goldmann, and T. Ebrahimi, “Propagation of geotags based on object duplicate detection,” in *Applications of Digital Image Processing XXXIII*, vol. 7798 of *Proceedings of the SPIE International Conference on Optics and Photonics*, pp. 77980S–1–77980S–8, Aug. 2010.
 - [7] P. Vajda, I. Ivanov, L. Goldmann, J.-S. Lee, and T. Ebrahimi, “3D object duplicate detection for video retrieval,” in *Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*, pp. 1–4, Apr. 2010.
 - [8] I. Ivanov, P. Vajda, L. Goldmann, J.-S. Lee, and T. Ebrahimi, “Object-based tag propagation for semi-automatic annotation of images,” in *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 497–506, Mar. 2010.
 - [9] I. Ivanov, F. Dufaux, T. M. Ha, and T. Ebrahimi, “Towards generic detection of unusual events in video surveillance,” in *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 61–66, Sep. 2009.
 - [10] I. Ivanov, “Digital acquisition of sine wave signals using TI MSP430F449,” in *Proceedings of the Telecommunications Forum*, pp. 761–764, Nov. 2006.
 - [11] I. Ivanov and P. Cvijic, “Quadrature demodulation of amplitude modulated signal,” in *Proceedings of the Telecommunications Forum*, pp. 1–4, Nov. 2005.

Standard contributions

- [1] I. Ivanov, P. Vajda, L. Goldmann, and T. Ebrahimi, ““Cheese” – A JPSearch Part 4 compliant image database,” Tech. Rep. JPEG2011/N5826, ISO/IEC JTC1/SC29/WG1, Jul. 2011.
- [2] P. Vajda, I. Ivanov, and T. Ebrahimi, ““Cheese” – Towards JPSearch compliant image database,” Tech. Rep. JPEG2011/N5739, ISO/IEC JTC1/SC29/WG1, Jan. 2011.
- [3] I. Ivanov, P. Vajda, and T. Ebrahimi, “A user-friendly platform to collect test images for mobile visual search,” Tech. Rep. MPEG2011/M18952, ISO/IEC JTC1/SC29/WG11, Jan. 2011.
- [4] M. T. Andrade, V. Barbosa, A. Carreras, P. Carvalho, and *et al.*, “Revised Contribution to the Advanced Surveillance AF,” Tech. Rep. MPEG2009/M16623, ISO/IEC JTC1/SC29/WG11, Jun. 2009.

- [5] M. T. Andrade, V. Barbosa, A. Carreras, P. Carvalho, and *et al.*, “Revised Extended template for the Advanced Surveillance AF,” Tech. Rep. MPEG2009/M16622, ISO/IEC JTC1/SC29/WG11, Jun. 2009.
- [6] M. T. Andrade, V. Barbosa, A. Carreras, P. Carvalho, and *et al.*, “Contribution to the Advanced Surveillance AF,” Tech. Rep. MPEG2009/M16161, ISO/IEC JTC1/SC29/WG11, Feb. 2009.
- [7] M. T. Andrade, V. Barbosa, A. Carreras, P. Carvalho, and *et al.*, “Extended template for the Advanced Surveillance AF,” Tech. Rep. MPEG2009/M16159, ISO/IEC JTC1/SC29/WG11, Feb. 2009.

TECHNICAL SKILLS

Project and process flow	Systems design and development; Integration; Qualitative and quantitative research methodologies; Proposal writing; Technical documentation.
Analytical and problem solving	Analytical and logical thinking; Data gathering, analysis and organization; Data mining, machine learning, pattern recognition and modeling.
Information technology	Multimedia processing; Signal processing; Social media analysis; Visual search in images and videos; Software engineering; Hardware systems design; Mobile network deployment; Wireless radio access networks; GSM/GPRS/EDGE (certificate from Alcatel-Lucent University Timisoara, Romania, in 2008); 2.75G BSS conceptual planning and hardware dimensioning.
Operating sys.	Microsoft Windows (XP/7), Unix/Linux, SUN Solaris.
Software tools	Microsoft Office (including MS Project and MS Visio), IBM Rational ClearCase, Adobe Photoshop.
Program. and script. languages	C/C++, Perl, Java, MatLab, OpenCV, HTML, PHP, MySQL, VHDL, Assembly x86 and TI MSP430, Tcl.

LANGUAGES

Serbian	Native language
English	Fluent, CEF level C2
French	Intermediate, CEF level B1
German	Basic

SOCIAL SKILLS

- Basketball player (Bor Basketball Club, 3rd Serbian league 1997–2001, the point guard)

HOBBIES

- Photography
- Travelling

“Twenty years from now you will be more disappointed
by the things that you didn’t do than by the ones you did do.
So throw off the bowlines. Sail away from the safe harbor.
Catch the trade winds in your sails.
Explore. Dream. Discover.”

— Samuel Langhorne Clemens (Mark Twain),
*American humorist, novelist,
writer and lecturer (1835–1910)*