

# **Semantic Image Understanding using Image Analysis and Related Information**

**画像分析と関連情報を利用した画像意味理解  
に関する研究**

SUPHEAKMUNGKOL SARIN

Global Information and Telecommunication Studies  
Multimedia Representation Research II

February 2012

Waseda University



*Dedicated to*

My beloved Grand-mother, CHHEVE Un (Yey Touch),  
and my dearest Aunty, SAN Nimol (Ming Ny).

You are always in our mind even physically you are not with us anymore...



## **Abstract**

In recent years, digital cameras have become ubiquitous; storage is less expensive, Internet access is available nearly everywhere and digital social interaction is an increasingly popular trend. Due to these reasons, digital images have grown exponentially and have been making it beyond the abilities of people to easily manage these important contents. In an effort to solve this burden, the author investigates on image understanding in order to bridge the semantic gap between human and machine. Towards this goal, the author proposed image analysis methods and system designs that go beyond the superficial image content analysis. The proposed schemes (i) fully exploit the holistic content analysis by utilizing not only the whole original image, but also its salient regions and its background; (ii) leverage other related information about the image such as GPS, temporal, layout, optical, and contextual information; or (iii) combine these schemes to complete this difficult task. The author also examines user's behaviour, user's perception, aesthetic values and photography grammar. In the scope of this dissertation, the author focuses on automatic image annotation, result re-ranking, and categorization and quality assessment tasks. These tasks are among the most fundamental and essential ones for semantic understanding of image. The contents of the thesis can be summarized as the following.

Chapter 1 sets the stage by giving the background of the research problem as well as the scope of the thesis namely, automatic image annotation, result re-ranking, and categorization and aesthetics quality assessment.

Chapter 2 gives the state-of-the-art research work on the related techniques towards image understanding, and the positioning and contributions of this thesis in this regard.

Chapter 3 explores the problem of automatic image annotation in a general case. One of the main bottlenecks in this area is the lack of integrity and diversity of features. The author proposes to solve this problem by utilizing 43 image features that cover the holistic content of the image from global to subject, background and scene. In the approach, salient regions and the background are separated without prior knowledge. Each of them together with the whole image are treated independently for feature extraction. Extensive experiments were designed to show the efficiency and the effectiveness of the approach. Two publicly available datasets manually annotated with the diverse nature of images were chosen for the experiments, namely the Corel5K and ESP Game datasets. The results confirm the superior performance of the proposed approach over the use of a single whole image using sign test with  $p$ -value  $< 0.05$ . Furthermore, the proposed combined feature set gives satisfactory performance compared to recently proposed approaches especially in terms of generalization even with just a simple combination. The approach also achieves a better performance with the same feature set versus the grid-based approach. More importantly, when using the proposed set of features with the state-of-the-art technique, the results show higher performance in a variety of standard metrics.

Chapter 4 focuses on the problem of automatic annotation in the personal case. By analysing users' behaviour and technology trends, the author proposes a novel solution for this task. The method integrates all contextual information available *to* and *from* the users, such as their daily emails, schedules, chat archives, web browsing histories, documents, online news, Wikipedia data, and so forth. Subsequently, the integrated information is analysed and important semantic terms are extracted. The keywords are in the form of named entities, such as names of people, organizations,

locations, and date/time as well as high frequency terms. They serve as annotation candidates for the photograph. Users can choose to validate these candidates. Experiments conducted with 10 subjects and a total of 313 photos prove that the proposed approach can significantly help users with the annotation process. The approach achieves a 33% gain in annotation time as compared to manual annotation. The results also demonstrate encouraging accuracy rate of the suggested keywords.

Chapter 5 is dealing with results re-ranking in the image retrieval task. Image search systems have a very limited usefulness since it is still difficult to provide different users with what they are searching for. This is because most research efforts to date have only been concentrating on relevancy rather than diversity which is also a quite important factor, given that the search engine knows nothing about the user's context. In the chapter, the author describes the proposed approach for photographic retrieval task (within the scope of ImageCLEF 2008). The novelty of the approach is the use of *AnalogySpace*, the reasoning technique over commonsense knowledge for document and query expansion, which aims to increase the diversity of the results. The proposed technique combines *AnalogySpace* mapping with other two mappings namely, *location* and *full-text*. Re-ranking mechanism is employed to the resulting images from the mapping by trying to eliminate duplicate and near duplicate results in the top 20. The experiments and the results conducted using the IAPR TC-12 photographic collection, with 20,000 still natural photographs, are represented. The results show that the integrated method with *AnalogySpace* yields better performance in terms of cluster recall and the number of relevant photographs retrieved by maintaining precision. The author finally identifies the weakness in the approach and ways on how the system could be optimized and improved.

Chapter 6 is interested in the problem of high quality photo categorization and aesthetic quality assessment. The chapter outlines the proposed framework for the tasks. The author addresses these challenges by exploring the aesthetics from the

combined perspectives of the artists and photographers. The author proposes to use the aesthetic primitives of images for visualization as a guideline for high and low-level image feature extraction and to classify this high quality content into six creative exposure themes, which are commonly followed by the professional photographers. Furthermore, the proposed framework suggests evaluating the quality of the photograph accordingly to these themes. In the proposed approach, the tasks are solved using statistical modelling and learning schemes. A small experiment using only the camera setting features was conducted and the result was encouraging.

Chapter 7 concludes the findings. Then, the future perspectives in structuring the image collections and eventually in making sense out of them are presented.

These analysis and methodology designs presented in the thesis shall contribute to the better understanding of visual content beyond the conventional approaches. In addition, it is shown that they meet one or more of the user's requirement attributes. Therefore, many fully targeted visual related applications and services - not limited to the image related ones - could rise from these findings.

# Declaration

In this thesis, part of the work on personal digital image annotation is based on a joint-research with SEIKO EPSON Corporation carried out at the KAMEYAMA Laboratory of the GITS, Waseda University. In addition, part of the work on general image annotation was conducted when the author was with DOCOMO Europe Communications Laboratories Europe GmbH.

It is to be noted that the views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies or endorsement, wither expressed or implied, of GITS, SEIKO EPSON Corporation or DOCOMO Communications Laboratories Europe GmbH.

It is also to be noted that some parts of this dissertation are taken or adapted from the author's published papers.



# Acknowledgments

First and foremost, I would like to express my deep gratitude to my supervisor, Professor Wataru KAMEYAMA who has always encouraged me in my research. Throughout these last six years, he has shown me the research, and guided me along this very interesting journey. He always excites me with lots of interesting and insightful ideas that I have never thought of. I also appreciate his responsiveness as an adviser, despite serving as the Vice-Dean and later on as the Dean of this graduate school, while having many other tasks to handle regularly. I have learned a lot from him not only about the research itself, but also about working methodologies and life conduct as a whole. He has been my real mentor.

I also would like to thank Professor Hidenori NAKAZATO, Professor Shigekazu SAKAI and Professor Hiroshi WATANABE for agreeing to serve as my dissertation committee. Undoubtedly, their valuable inputs have contributed towards the successful completion of this thesis.

Second, I would like to thank SEIKO EPSON Corporation especially Dr. Tadashi MIYOSAWA and Mr. Toshinori NAGAHASHI for their kind cooperation since the beginning of this project. Without their kind help, their time, and their exchange of ideas during our monthly meeting, this work could never have been accomplished.

Third, I would like to thank Dr. Masami YABUSAKI for giving me the opportunity to do my research at DOCOMO Communications Laboratories Europe GmbH.

I also would like to thank my manager, Dr. Michael FAHRMAIR, my director, Dr. Matthias WAGNER, and my colleagues for their kindness during my stay there.

I also would like to thank a number of people who helped me along the projects in this thesis, especially those who accepted to be the subjects of the experiments. I really appreciate their time and their kind help.

I also would like to thank all my friends everywhere who always cheer me up each and every day. Millions of thanks also go to friends, colleagues and staffs in the KAMEYAMA laboratory, GITS and HOWARP who always provide me a friendly and helpful environment. I am sad to lose one of them, Mr. Kok-Meng ONG, who passed away earlier this year. He was a great friend and colleague.

I am totally indebted to the Government of Japan that kindly provides me the Monbugakakusho (MEXT) scholarship so that I could pursue my Master's and Doctor's degrees here at WASEDA University.

Last but not the least, this thesis is for my family. They are always with me. Their unconditional support has helped me greatly throughout my years. Thank you so much for everything! I love you!

# Contents

<b>Abstract</b>	<b>v</b>
<b>Declaration</b>	<b>ix</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Objective . . . . .	2
1.3 Dissertation Organization . . . . .	3
<b>2 Representative Methods &amp; Models in Semantic Image Understanding, and Thesis’s Positioning &amp; Contributions</b>	<b>7</b>
2.1 Introduction . . . . .	7
2.2 Towards Semantic Image Understanding . . . . .	8
2.2.1 Representative Methods . . . . .	8
2.2.2 Processing Stages . . . . .	10
2.2.3 Models . . . . .	10
2.3 Positioning and Contributions of this Thesis . . . . .	13
<b>3 On Automatic Image Annotation: The General Case</b>	<b>17</b>
3.1 Introduction . . . . .	17
3.1.1 Background and Motivation . . . . .	17

3.1.2	Problem Formulation . . . . .	18
3.1.3	General Concept . . . . .	18
3.1.4	Contributions . . . . .	21
3.2	Related Works . . . . .	22
3.2.1	Prior Art in Image Pre-processing Techniques . . . . .	22
3.2.2	Prior Art in Label Propagation Techniques . . . . .	23
3.3	The Proposed Approach . . . . .	25
3.3.1	Overview . . . . .	25
3.3.2	Salient Regions and Background Extraction for Holistic Image Representation . . . . .	26
3.3.3	Holistic Feature Extraction . . . . .	29
3.3.3.1	Color Features . . . . .	29
3.3.3.2	Texture Features . . . . .	30
3.3.3.3	Scene Feature . . . . .	30
3.3.3.4	Advanced Local Invariant Features . . . . .	30
3.3.4	Experiment Setting . . . . .	33
3.3.4.1	Datasets . . . . .	33
3.3.4.2	Performance Metrics . . . . .	34
3.3.4.3	Validation Procedure . . . . .	35
3.4	Results . . . . .	37
3.4.1	Joint Equal Combination Model . . . . .	38
3.4.1.1	Joint Equal Combination Annotation Scheme . . . . .	38
3.4.1.2	Results . . . . .	38
3.4.2	TagProp Model . . . . .	39
3.4.2.1	TagProp Annotation Scheme . . . . .	39
3.4.2.2	Performance as Image Retrieval from Single-keyword Queries Task . . . . .	41
3.4.2.3	Performance as Image Retrieval from Multi-keywords Queries Task . . . . .	42

3.4.2.4	Some Qualitative Results in the Retrieval Task . . .	46
3.4.2.5	Image Auto-annotating Performance . . . . .	50
3.4.2.6	Some Qualitative Results in the Annotation Task . .	54
3.4.2.7	Number of Worse, Draw and Better Results of Keyword- wise and Image-wise Precision . . . . .	54
3.4.3	Discussion . . . . .	58
3.5	Conclusion . . . . .	60
<b>4</b>	<b>On Automatic Image Annotation: The Personal Case</b>	<b>63</b>
4.1	Introduction . . . . .	63
4.1.1	Background and Motivation . . . . .	63
4.1.2	Problem Formulation and General Idea . . . . .	65
4.2	Related Works . . . . .	66
4.2.1	Manual Annotation (with UI enhancement) . . . . .	66
4.2.2	Semi-automatic Annotation (including collaborative annotation)	67
4.2.3	Automatic Annotation . . . . .	68
4.3	Proposed Approach: Leveraging Context to Bridge Semantic Gap . .	70
4.3.1	Nature of Personal Digital Photographs . . . . .	70
4.3.2	Gathering Contextual Information . . . . .	70
4.3.3	Using Time + Location as Photograph Filters . . . . .	72
4.3.4	Extracted Keywords . . . . .	73
4.4	System Design and Implementation . . . . .	74
4.4.1	Data Acquisition . . . . .	76
4.4.2	Relevant Files Generation . . . . .	79
4.4.3	Keywords Generation . . . . .	80
4.4.3.1	Named Entity Generation . . . . .	80
4.4.3.2	Statistical Keywords . . . . .	81
4.4.4	Annotation GUI and Metadata Coverage . . . . .	83
4.5	Empirical Evaluations . . . . .	85

4.5.1	Validation goals . . . . .	85
4.5.2	Participants and Data sets . . . . .	85
4.5.2.1	Subjects . . . . .	85
4.5.2.2	Personal Photographs . . . . .	85
4.5.3	Experiment Process . . . . .	86
4.5.3.1	Manual Annotation . . . . .	87
4.5.3.2	Annotation with Keyword Suggestion Features . . . . .	87
4.5.3.3	Keywords Judging . . . . .	88
4.5.4	Results and Discussion . . . . .	89
4.5.4.1	Experimental Results and Analysis . . . . .	89
4.5.4.2	Discussion . . . . .	94
4.6	Other Features . . . . .	96
4.6.1	Searching . . . . .	96
4.6.2	Browsing . . . . .	96
4.7	Conclusion . . . . .	98
<b>5</b>	<b>On Result Re-ranking in Image Retrieval Task</b>	<b>101</b>
5.1	Introduction . . . . .	101
5.2	The Proposed Approach . . . . .	102
5.3	Related Works . . . . .	104
5.4	Implementation . . . . .	104
5.4.1	Matching . . . . .	104
5.4.1.1	Location matching . . . . .	106
5.4.1.2	AnalogySpace matching . . . . .	106
5.4.1.3	Full-text matching . . . . .	107
5.4.2	Re-ranking . . . . .	108
5.4.2.1	Pair distance similarity . . . . .	108
5.4.2.2	Re-rank . . . . .	108
5.5	Evaluation . . . . .	109

5.5.1	Protocol . . . . .	109
5.5.2	Dataset . . . . .	109
5.5.3	Query . . . . .	110
5.5.4	Measurement techniques . . . . .	111
5.6	Results and Discussions . . . . .	112
5.7	Conclusion . . . . .	115
<b>6</b>	<b>On Categorization and Aesthetics Quality Assessment</b>	<b>117</b>
6.1	Introduction . . . . .	117
6.1.1	Background and Motivation . . . . .	117
6.1.2	Problem Formulation and General Idea . . . . .	118
6.2	Related Works . . . . .	119
6.2.1	Categorization and Annotation . . . . .	119
6.2.2	Aesthetic Quality Assessment . . . . .	119
6.2.2.1	Content-based approach . . . . .	119
6.2.2.2	Subjective approach . . . . .	120
6.3	Proposed Approach and Framework . . . . .	121
6.3.1	Conceptual Approach . . . . .	121
6.3.1.1	Aesthetics and Categorization . . . . .	121
6.3.1.2	Camera Setting Context . . . . .	124
6.3.2	Research Framework . . . . .	125
6.3.2.1	Framework of the Approach . . . . .	125
6.3.2.2	Feature Extraction . . . . .	125
6.3.2.3	Feature Selection and Classification . . . . .	128
6.4	Evaluation Protocol . . . . .	129
6.4.1	Dataset . . . . .	129
6.4.2	Analysis . . . . .	130
6.5	Challenges . . . . .	132
6.6	A Preliminary Experiment . . . . .	133

6.6.1	Dataset and Extracted Features . . . . .	133
6.6.2	Model Building, Evaluation and Results . . . . .	133
6.7	Conclusion . . . . .	135
<b>7</b>	<b>Conclusion and Future Perspectives</b>	<b>137</b>
7.1	Summary . . . . .	137
7.2	Future Perspectives . . . . .	139
7.2.1	Structuring . . . . .	139
7.2.2	Making sense . . . . .	140
	<b>Appendix</b>	<b>143</b>
<b>A</b>	<b>List of publications</b>	<b>143</b>
	<b>Bibliography</b>	<b>149</b>

# List of Figures

1.1	Research objective: Image analysis and its methodology designs for semantic image understanding using image analysis and related information . . . . .	3
1.2	The organizational structure of the main chapters of the dissertation .	6
2.1	Processing steps and their respective methods towards semantic image understanding . . . . .	11
2.2	Chapter’s positioning in the processing steps and summary of covered user’s requirements . . . . .	16
3.1	Example showing the importance of the separation between (a) original whole image, (b) the background, and (c) the salient regions. In many cases, using the background and salient regions in addition to the whole image can leverage the chance of getting all the related images and can subsequently lead to better recall of relevant keywords. This is the case particularly for an incomplete labeled training set where the image is not labeled with all relevant keywords. Moreover, weakly labeled training data are the usual case of data obtained from the Internet. . . . .	19
3.2	Example showing the importance of salient regions: from the color feature space, the relatively bigger proportion of the background with different colors can make the two images very different from each other.	20

3.3	Example showing different methods used prior to image feature extraction: (a) the image is segmented into different regions, (b) the image is decomposed into predefined and fixed blocks, (c) the image is dense sampled (left) or is sampled by points of interest (right) . . .	24
3.4	Overall architecture of our proposed approach . . . . .	27
3.5	Combined model for salient region and background extraction . . . . .	29
3.6	Processing steps in local invariant features ( SIFT and Color SIFT) extraction . . . . .	31
3.7	Grid-based salient regions and background extraction . . . . .	37
3.8	Corel5K dataset retrieval examples in comparison with the baseline approaches . . . . .	51
3.9	ESP game dataset retrieval examples in comparison with the baseline approaches . . . . .	52
3.10	ESP dataset annotation examples in comparison with the baseline approaches . . . . .	55
3.11	Corel5K dataset annotation examples in comparison with the baseline approaches . . . . .	56
3.12	Example showing some complex images that result in failure in salient regions and background extraction: (a) the original image, (b) the extracted salient regions and (c) the extracted background. . . . .	59
4.1	Contextual ambient information gathering . . . . .	71
4.2	Overall View of the Concept . . . . .	73
4.3	System architecture of the implemented prototype . . . . .	75
4.4	Data acquisition of personal and public information with Google Desktop Search . . . . .	78
4.5	Process of generating relevant files to the photo with location and time as event filter . . . . .	79
4.6	Named entity keyword extraction process . . . . .	82

4.7	Statistical keyword extraction process . . . . .	83
4.8	(A) Blank annotation interface; (B) Annotating interface with key- words suggestion feature . . . . .	88
4.9	(A) Acceptable hit rate of Who (People’s name) and Org. (Organi- zation names) keywords; (B) Coverage rate for 1 acceptable keyword of Who (People’s name) and Org. (Organization names) . . . . .	90
4.10	(A) Acceptable hit rate of statistical keywords; (B) Number of ac- ceptable keywords of each photo; (C) Coverage rate for at least 8 acceptable keywords of statistical keywords . . . . .	91
4.11	(A) Manual Annotation and Annotation with Keyword Suggestion Features of Each Subject; (B) Average Annotation Without and With Keyword Suggestion Features . . . . .	92
4.12	Proposed conceptual annotation inferface layout for future implemen- tation . . . . .	95
4.13	Search engine . . . . .	97
4.14	Browse engine . . . . .	98
4.15	The future goal . . . . .	100
5.1	The proposed approach . . . . .	103
5.2	Flow diagram of the system architecture . . . . .	105
5.3	AnalogySpace matching . . . . .	106
5.4	Full-text matching . . . . .	107
5.5	Re-ranking process . . . . .	109
5.6	Example of a photograph of the collection and its attached metadata	110
6.1	Example images of the six creative exposure themes . . . . .	123
6.2	Conceptual approach of <i>exposure theme classification</i> and <i>photo qual- ity assessment</i> . . . . .	126
6.3	Generated decision tree model . . . . .	134

7.1	Extracting knowledge from images . . . . .	142
-----	--	-----

# List of Tables

2.1	Image related tasks: Human Vs. Computer . . . . .	8
2.2	Some taxonomy in image understanding (adapted from ToC of [123])	9
2.3	The four different models towards semantic image understanding and some example methods . . . . .	12
2.4	Positioning of the thesis and its chapters in each image understanding model . . . . .	15
3.1	Statistics of the two datasets: Corel5K and ESP Game. . . . .	34
3.2	Summary of performance comparison when using our features with the JEC approach. Note that JEC-15 is the result reported in [52] of the JEC method using their 15 features. . . . .	39
3.3	Performance comparison when using only whole image versus whole+roi+bg in terms of MAP (A) . . . . .	39
3.4	Performance comparison between our proposed approach and the grid-based one in terms of MAP (A) . . . . .	40
3.5	Performance comparison between our work and the state-of-the-art methods for the Corel5K dataset. Note that TagProp is the original results claimed in 7). TagProp* is our implementation of the results using the same features, the portion of the code provided by the authors in their website and the same number of neighbors ( $k = 200$ )	42
3.6	Performance comparison between our work and the state-of-the-art methods for the ESP Game dataset. . . . .	43

3.7	Performance comparison between our work and the state-of-the-art methods in terms of multi-keyword queries in the Corel5K dataset. . . . .	44
3.8	Performance comparison between our work and the state-of-the-art methods in terms of multi-keyword queries in the ESP Game dataset. . . . .	45
3.9	Performance comparison when using only whole image versus whole+roi+bg in terms of multi-keyword queries of the Corel5K dataset. . . . .	47
3.10	Performance comparison between our proposed approach and the grid-based one in terms of multi-keyword queries of the Corel5K dataset. . . . .	48
3.11	Performance comparison when using only whole image versus whole+roi+bg in terms of multi-keyword queries of the ESP Game dataset. . . . .	49
3.12	Summary of performance of our auto-annotating performance . . . . .	53
3.13	Performance comparison when using only whole image versus whole+roi+bg in terms of our auto-annotating performance . . . . .	53
3.14	Performance comparison between our approach and the grid-based one in terms of auto-annotating performance . . . . .	53
3.15	Number of worse, draw and better results in keyword-wise MAP of our whole+roi+bg versus other approaches in the Corel5K datasets . . . . .	57
3.16	Number of worse, draw and better results in keyword-wise MAP of our whole+roi+bg versus other approaches in the ESP Game datasets . . . . .	57
3.17	Number of worse, draw and better results in image-wise MAP of our whole+roi+bg versus other approaches in the Corel5K datasets . . . . .	57
3.18	Number of worse, draw and better results in image-wise MAP of our whole+roi+bg versus other approaches in the ESP Game datasets . . . . .	58
5.1	Example of a query topic . . . . .	111
5.2	Precision (P) at the top $n$ results . . . . .	113
5.3	Cluster Recall (CR) at the top $n$ results . . . . .	114

5.4	Other metrics: Number of Relevant Retrieved images (NumRelRet), Number of Relevant images (NumRel), Mean Average Precision (MAP), Geometric Mean Average Precision (GMAP), Blind RElevance Feed- back (BREF) . . . . .	115
6.1	ImageCLEF VCDT Concepts . . . . .	130
6.2	Correspondence between: (A) Creative exposure themes and Anno- tation concepts, (B) Quality and Annotation concepts . . . . .	131
6.3	Confusion matrix . . . . .	134
6.4	Precision, Recall/Sensitivity, Specificity and Accuracy rates (Let $TP$ : <i>TruePositive</i> ; $TN$ : <i>TrueNegative</i> ; $FP$ : <i>FalsePositive</i> ; $FN$ : <i>FalseNeg-</i> <i>ative</i> ) . . . . .	135



# Chapter 1

## Introduction

### 1.1 Background

Today's low cost of digital cameras and digital storage devices, combined with the rapid adoption of broadband Internet connectivity and the increasingly popular social websites, have enabled us to generate and consume a tremendous number of images. In parallel, as the number of images is rapidly expanding, we have also encountered grave difficulties with image-related works even the fundamental ones such as organizing, searching and browsing. The current methods of organizing, browsing, searching and sharing as well as the results that we obtain from those tasks are *very limited* and *unnatural* [121]. Thus, we cannot fully enjoy and make use of our image contents. This is a crucial problem because the real value of the content depends on how we can easily manage, access, and infer useful information from them, and yet until today, there is no complete real-world solution towards this matter.

These above mentioned problems are due to the *lack of semantic understanding of image*. As goes the saying "*image is worth a thousand words*", we need ways to enable the computer to understand the image beyond just the *pixel values*. This should

incorporate *different interpretations* about the image or set of images from different perspectives depending on the context, environment or situation. Researchers have paid attention in this field especially in the recent years. There have been research efforts in different spectrums from lower level in image processing such as edge detection, feature extraction to a higher level in computer vision such as object/scene recognition, classification and retrieval.

There are many challenges in Image Understanding (IU). They include view point variation, illumination, occlusion, scale, deformation, background clutter, object intra-class variation, local ambiguity and more importantly individual user's perception. This is because IU is a decision task situated at the last stage of computer vision. Usually, it involves the user's interpretation. Towards this goal, it is thus vital to look beyond the conventional image content by also leveraging related information about the image such as GPS information, temporal information, layout information, optical information, user's behavior, user's contextual information and user's perception. The author is doing as such in this thesis.

## 1.2 Objective

The objective of this thesis is to explore and derive image analysis methods and designs towards the semantic understanding of image by using image content analysis as well as other related information about the image. Figure 1.1 shows this objective. The methods and the designs shall contribute to the reduction of the semantic gap. In the scope of this dissertation, the author focuses on *automatic image annotation, result re-ranking, categorization and quality assessment* tasks. These tasks are among the most fundamental and essential ones for semantic understanding of image:

1. Automatic Image Annotation: Many image-related applications would become efficient and effective once every image is meaningfully described. Therefore,

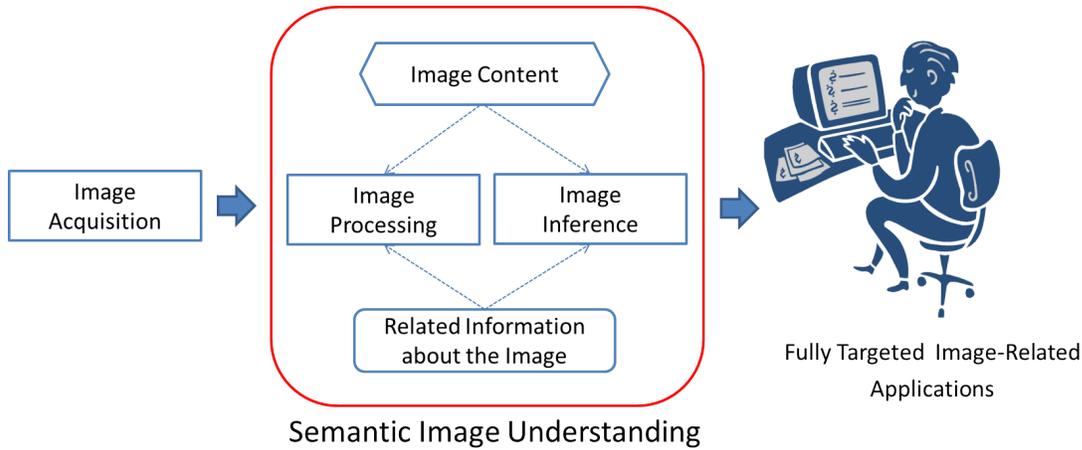


Figure 1.1: Research objective: Image analysis and its methodology designs for semantic image understanding using image analysis and related information

this thesis looks into the problem of automatic labeling. Both the general purpose image and personal image scenarios are explored.

2. Result Re-ranking: It is obvious that without much information about the users, one cannot give a general retrieval result set that would please every user. In this case, re-ranking mechanism of the result set is very essential. This thesis explores a practical and natural technique for doing as such.
3. Categorization and Quality Assessment: This sub topic is becoming increasingly important with the exponential growth of images. The target of this sub topic is to give a framework on how such highly subjective tasks could be realized. The case of high quality photograph is studied.

### 1.3 Dissertation Organization

This thesis consists of seven main chapters and the organization is as follows. This Chapter introduces the background of the research problem and its objectives. The following is the roadmap to subsequent chapters.

Chapter 2 begins with the introduction of the related techniques towards image understanding and follow by the positions and the contribution of this thesis.

Chapter 3, 4, 5 and 6 are the main chapters. They explore *automatic image annotation, results-ranking, and categorization and quality assessment* respectively.

- Chapter 3 presents our investigations on Automatic Image Annotation (AIA) in a general context. The focus is on image content based feature extraction. It presents our combined model in image saliency and background extraction as well as the scheme for holistic feature extraction for an AIA task. Extended results and comparison with the state-of-the-art techniques are presented.
- Chapter 4 discusses AIA in the personal context. We present our novel method in exploiting users' personal and public information for a semi-automatic image annotation.
- Chapter 5 introduces the result re-ranking problem in image retrieval task. In the developed method, commonsense knowledge is used as key to promote diversity in the result sets and yet maintaining the precision.
- Chapter 6 presents the study on categorization and quality inference. The chapter introduces a framework for the tasks by considering the perspectives of the professional photographers and artists. Aesthetic primitives are investigated.

Chapter 7 summarizes the key findings of the thesis then follows by the insightful perspectives for the future works.

The dissertation ends with a bibliography reference, and a list of papers published within the scope of this thesis. The flow of the structure of this dissertation is also illustrated in Figure 1.2.

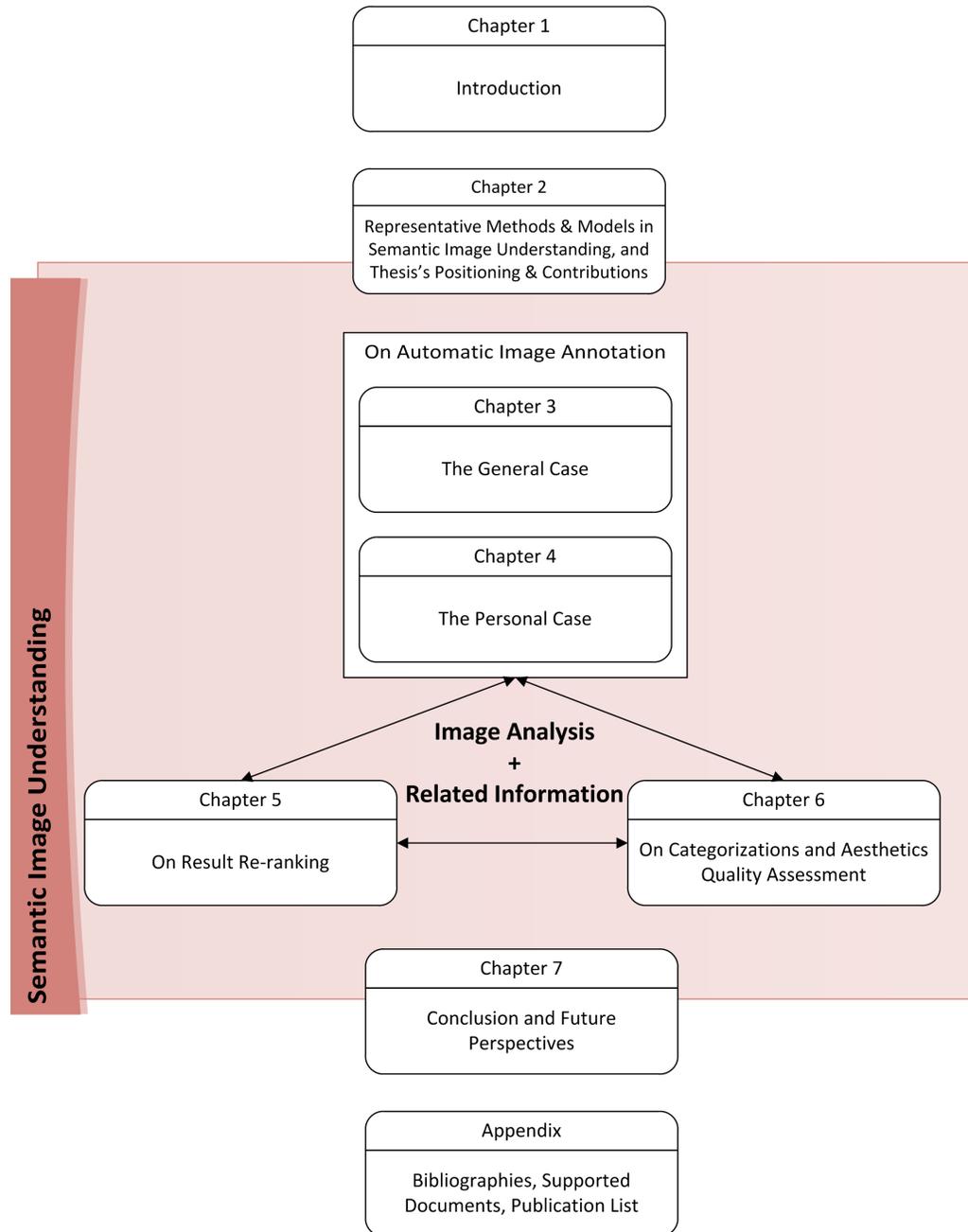


Figure 1.2: The organizational structure of the main chapters of the dissertation

# Chapter 2

## Representative Methods & Models in Semantic Image Understanding, and Thesis's Positioning & Contributions

### 2.1 Introduction

In Semantic Image Understanding, we would like to teach machine to see the image like human does (i.e. beyond the pixel values) so that it can render fully adaptive services back to human. There have been important advancements in image processing and computer vision in the last 50 years. Nowadays, computer can handle some vision tasks accurately and efficiently. For example, machine is better than human being in the tasks such as aligning images, doing face morphing, etc. However, these tasks tend to be very specific and context independent. That means if we can provide computer with a set of instructions to solve a problem, it will excel in the task. When it comes to image understanding tasks such as segmentation, contextual tracking, object recognition, etc., which cannot be easily formulated, human is still far better than machine. Table 2.1 shows the examples. That is because those tasks are rather complex and involve perceptual and cognition understanding. Neverthe-

	Computer	Human
Complexity ↑		<ul style="list-style-type: none"> <li>• Quality inference</li> <li>• Photo summarization</li> <li>• Clustering</li> <li>• Recognition/Labeling</li> <li>• Contextual tracking</li> <li>• Segmentation</li> <li style="text-align: center;">⋮</li> </ul>
	<ul style="list-style-type: none"> <li>• Aligning images</li> <li>• Face morphing</li> <li style="text-align: center;">⋮</li> </ul>	

Table 2.1: Image related tasks: Human Vs. Computer

less, we have to bridge this gap between human and computer because only when machine could understand the image better that they can provide better targeted image related services or applications to the consumers.

## 2.2 Towards Semantic Image Understanding

### 2.2.1 Representative Methods

Towards semantic image understanding, there have been a lot of research efforts from image acquisition, basic image processing tasks to very advanced inference tasks [123]. These include image formation, low-level feature detection and representation, mid/high-level feature detection and representation, segmentation, salient region extraction, segmentation, salient region extraction, feature-based alignment, structure from motion, dense motion estimation, image stitching, computational photography, image recognition, etc. The Table 2.2 shows some of the taxonomy.

<b>Methods / Domains</b>	<b>Sub-domains or Tasks</b>
Image formation	Light, Camera, Geo-metric transformation, Image formation, Magnetic Resonance, etc.
Low-level Feature detection and representation	Color, Texture, Points and patches, Edge, Lines, etc.
Mid/High-level Feature detection and representation	Bag-of-feature model, contextual/multi-modal features, etc.
Segmentation	Active contours, Normalized cut, Graph cut, etc.
Salient region extraction	Spectral residual, Frequency tuned, etc.
Feature-based alignment	2D and 3D features based alinement, Pose estimation, etc.
Structure from motion	Triangulation, Bundle adjustment, Factorization, etc.
Dense motion estimation	Translational alignment, Parametric motion, etc.
Image stitching	Motion models, etc.
Computational photography	Image matting, Image composition, Image/camera calibration, HDR, etc.
Stereo correspondence	Epipolar geometry, Sparse correspondence, Dense correspondence, Local methods, Global optimization Multi-view stereo, etc.
3D construction	Shapes, Surface representation, Active range finding, Model-based reconstruction, etc.
Image-based rendering	View interpolation, video-based rendering, etc.
Recognition	Object detection, Face recognition, Category recognition, Automatic annotation, Emotion, Quality, Context and scene understanding, etc.
Result set re-ranking	Clustering, Diversity promotion, etc.

Table 2.2: Some taxonomy in image understanding (adapted from ToC of [123])

## 2.2.2 Processing Stages

We can group the sub-domains into three different stages namely, image acquisition (hardware processing), image feature extraction and representation (low level and mid level processing), and image inference tasks (high/decision level processing). Figure 2.1 gives categorization of the tasks into the three main processing stages.

## 2.2.3 Models

The approaches of the methods that have been proposed from low-level image processing to inference tasks can further be divided into four categories:

- Brute-force: In this category, manual work is conducted. There have been works trying to design the interface in a convenient way so that users can easily perform the manual tasks such as annotation [11] [103] [9]. There have also been efforts trying to build a game with purpose to leverage user's joy of playing a game to describe the images such as [44] [133].
- Image Analysis: This is a category of the conventional research. Researchers have been trying to make sense of the image from its content through color, texture, edge, patches, as well as others that can be derived. The short coming is the gap between context and content. Though, some work could be performed using the image content alone, there is an obvious limitation when it comes to be used towards user's fully targeted applications or services. In addition, most attempts in this category make use only the original whole image. Datta et al. have the survey of all the related works [26].
- Context / Related Information: This approach tries to leverage other information related to the image beside its visual content. Usually, this is achieved by trying to associate some context information. It can range from time, location, sound, video, activities, etc. [69] [68]. Recently, leveraging contextual

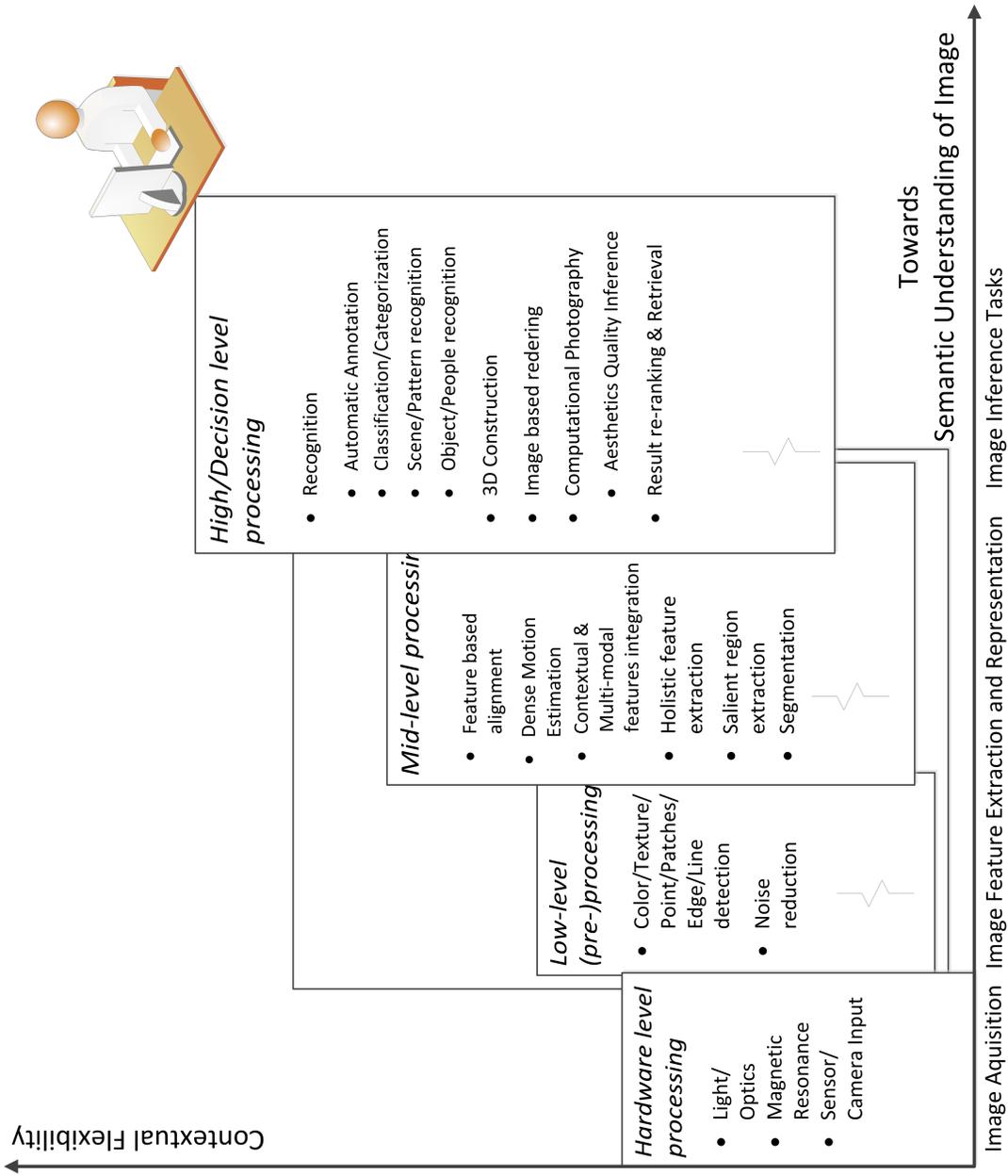


Figure 2.1: Processing steps and their respective methods towards semantic image understanding

		<b>Model</b>				
		Brute-force	Image Analysis	Context/Related Information	Multi-modal	
	[11, 103, 9, 44]	<b>X</b>				
Representative Methods in each Model	[26]		<b>X</b>			
	[68, 69, 37, 112, 146]			<b>X</b>		
	[66, 25, 27]					<b>X</b>

Table 2.3: The four different models towards semantic image understanding and some example methods

information from the user’s social circle is getting increasingly important due to its popularity [37] [112] [146].

- Multimodal: This a hybrid solution combining image analysis model and context/related information model. Recently, many works have been proposed in this direction due to the increasingly available sensory data. Katti et al., for instance, tried to categorize interestingness using not only image content, but also some optical features [66]. For quality assessment, we see the work of Datta et al. in photo quality assessment and emotion inference [25] [27]. However, often times, only one or a few aspects of image semantics are covered. Thus, an integrated and holistic solution is still needed.

## 2.3 Positioning and Contributions of this Thesis

From the literature, it is clear that there are two of the fundamental problems in computer vision and image understanding. They are the superficial usage of image content information and the lack of information about the image. Often times, only the content (i.e. pixel values) of the image is known. Moreover, many researchers only make use of this original whole image. *If we would like the computer to imitate how human sees the images, it is important (i) to imitate the way human recognizes an image; and (ii) to provide the computer with the related information about the image in a similar manner. This is the position of this thesis.* We would like (i) to fully exploit the image content beyond just the original whole image; and (ii) to adopt the multi-modal model by trying to leverage not only the image content, but also all the related information about the image. There have been works on the same direction as shown earlier. However, they are still superficial. Our approaches presented in this thesis maximize holistic content analysis and rich contextual related information, diversify the result sets, and aim at user's perception and requirements. For the latter, we try to target one or more of the user's requirement attributes. We believe that the final integration that leverage the synergy of these proposed approaches will be one of the ultimate solutions. The following describes the focus of each chapter:

Chapter 3 investigates on image annotation in the general case. The proposed approach responses to the relevancy and diversity requirements by leveraging salient regions and background in addition to the whole image for *holistic feature extraction and representation*.

In Chapter 4, the author presents the study of automatic image annotation in the case of personal usage. By studying the technology trends and user's information consumption behavior, a novel mechanism incorporating *user's personal informa-*

*tion and public information* is derived. The method is thus adaptive, contextual and user centered. This responses to the following requirements: relevancy, familiarity, trustworthy, interactiveness, freshness and enjoyment.

In Chapter 5, the author presents the discussion on promoting diversity through leveraging commonsense knowledge base and other resources. This is in response to a problem in relevancy, diversity and familiarity.

In Chapter 6, the study on categorization and quality inference is presented. In this chapter, the author propose to categorize the image based on the perspective of the professional photographers. In addition, the framework of the proposed approach follows the guideline of aesthetics primitives for visualization. Thus, it responses well to the issues raised in the quality, relevancy and diversity.

Figure 2.4 illustrates the positioning of the thesis in image understanding models. The position of each chapter as well as the position of the thesis are shown. Furthermore, our combined research tasks in this thesis have responded to nine important criteria in user's requirements. Figure 2.4 gives the contributions of our methods to the criteria. This Chapter ends with a summary of the contributions of each chapter in different processing stages as well as all the covered image requirement attributes, as given in Figure 2.2.

		<b>Model</b>			
		Brute-force	Image Analysis	Context/Related Information	Multi-modal
	[11, 103, 9, 44]	x			
Representative Methods in each Model	[26]		x		
	[68, 69, 37, 112, 146]			x	
	[66, 25, 27]				x
	Chpt. 3		x		
Positioning of this Thesis	Chpt. 4			x	x
	Chpt. 5			x	x
	Chpt. 6		x	x	x
	All		x	x	x

Table 2.4: Positioning of the thesis and its chapters in each image understanding model

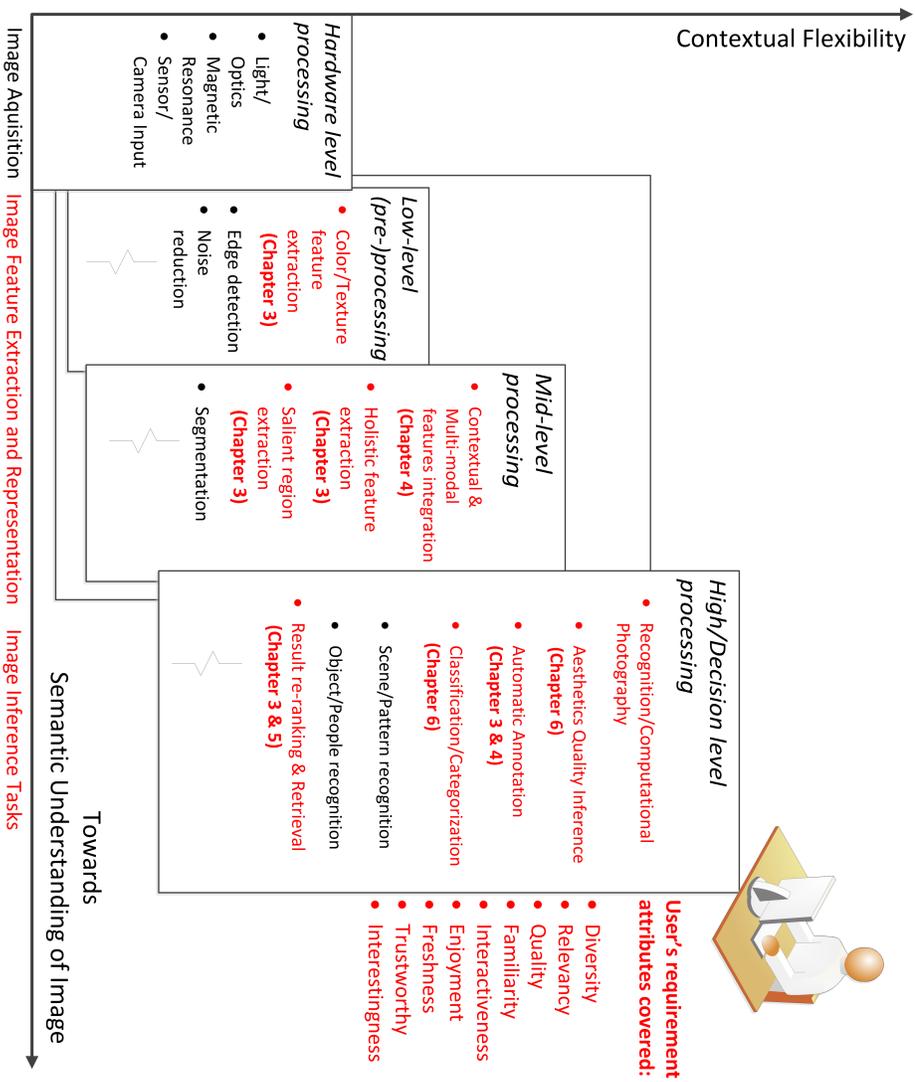


Figure 2.2: Chapter's positioning in the processing steps and summary of covered user's requirements

# Chapter 3

## On Automatic Image Annotation: The General Case

### 3.1 Introduction

#### 3.1.1 Background and Motivation

We are now living in the world with billions of images [38]. As for illustrative examples, Flickr reported that it reached 5 billion photos back in September of 2010 [5] and Facebook has announced 2.5 billion as the number of photos uploaded to its social sharing website per month [4]. Given the fact that the number will only keep increasing at an exponential rate, there is a critical demand for an efficient and effective tool that can help the users manage their large volume of content. The positive side is that we also have a huge amount of images that are partially labeled by the owner or the crowds through these popular digital social networking websites. Automatic Image Annotation (AIA) is a very important research field because it addresses the issue by supporting a keyword-based search and organization system. AIA has been an ongoing research for more than a decade and has been very active in the recent years. Researchers have been trying to exploit different kinds of resources and learning mechanisms from visual, textual, ontology to social

labeling over the Internet [26]. Though it is a highly challenging task, progress has been made throughout the years. However, there is one main problem that we could observe. It is the integrity and the diversity of the features. We tackle this issue in this chapter.

### 3.1.2 Problem Formulation

We formulate the annotation problem as a sample based one in which keywords for unknown images are inferred from a labeled training dataset.

Let  $TD = \{(I_1, W_{I_1}), (I_2, W_{I_2}), \dots, (I_p, W_{I_p})\}$  be the annotated training dataset which contains  $p$  pairs of  $(I_n, W_{I_n})$ , where  $I_n$  represents the image  $n$  and  $W_{I_n}$  is its description;  $W = \{w_1, w_2, \dots, w_m\}$  is a set of  $m$  words and  $F = \{f_1, f_2, \dots, f_k\}$  is a set of  $k$  visual features. The automatic image annotation aims to select a subset of top ranked words from the dictionary  $W$  and can be formally defined as follows:

$$AIA(J, TD, W, F) = \langle P_{J,w_1}, P_{J,w_2}, \dots, P_{J,w_m} \rangle \quad (3.1)$$

where  $J$  is a previously unknown image to be annotated and  $P_{J,w_r}$  is the probability generated by the annotator  $AIA$  of the word  $w_r$  for image  $J$ . Finding a good set of keywords involves (i) having a good machine learning algorithm, and (ii) defining and selecting important features. This chapter focuses on the latter.

### 3.1.3 General Concept

Fig. 3.1 illustrates the general idea of our approach. For an unknown image, it is obvious that the concurrent use of its salient regions, its background and its original whole image will enable a better chance of finding all relevant keywords for the image from the training set. This is intuitive and also corresponds to human's perception response when trying to search, recognize or describe a new image. Despite the fact, to the best of our knowledge, none of the previous works has made use of the

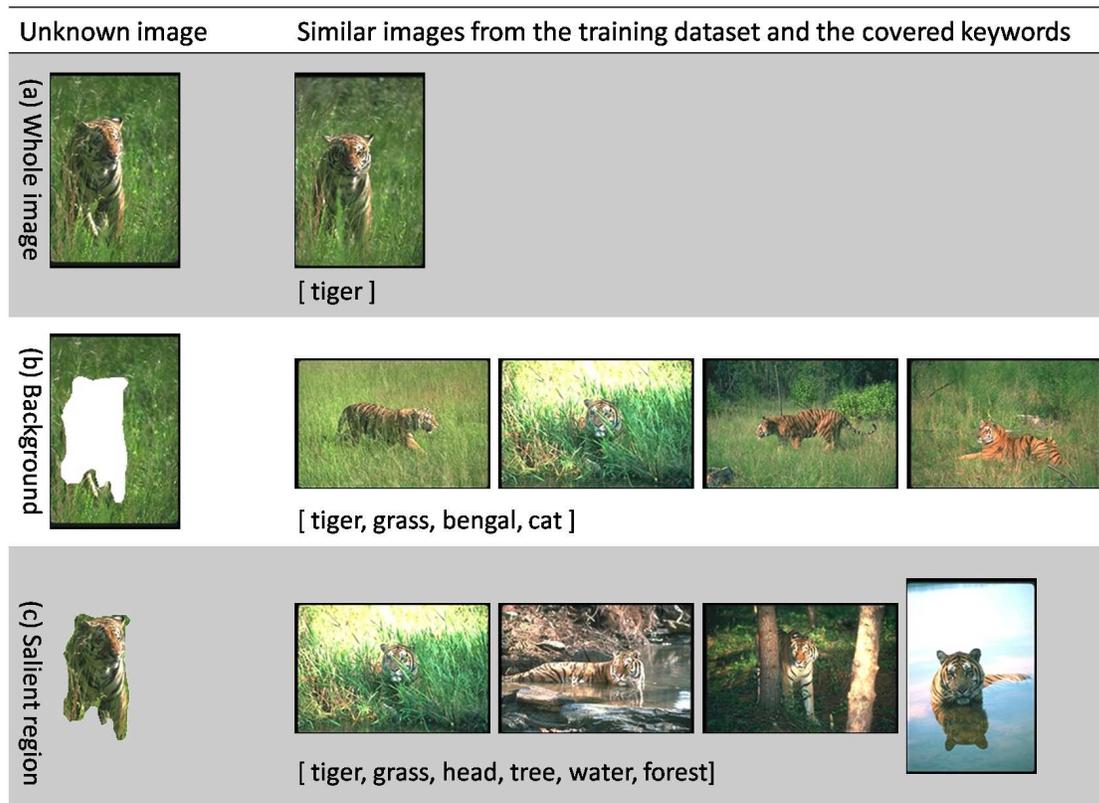


Figure 3.1: Example showing the importance of the separation between (a) original whole image, (b) the background, and (c) the salient regions. In many cases, using the background and salient regions in addition to the whole image can leverage the chance of getting all the related images and can subsequently lead to better recall of relevant keywords. This is the case particularly for an incomplete labeled training set where the image is not labeled with all relevant keywords. Moreover, weakly labeled training data are the usual case of data obtained from the Internet.

*The Figure is taken from Figure 1 of the author's paper [J1]*



(a)



(b)

Figure 3.2: Example showing the importance of salient regions: from the color feature space, the relatively bigger proportion of the background with different colors can make the two images very different from each other.

*The Figure is taken from Figure 2 of the author's paper [J1]*

*background image* and used it in synergy with *salient regions* and the *whole image*. With the recent progress in salient region extraction methods, we believe that there can be an improvement in the image annotation technique when processing the three images altogether. This is because there can be many variations (e.g. level of illumination, view points or occlusion) of an object or a scene depending on how the image is taken. To be able to get the maximum number of keywords from the training dataset, we have to be able to find all the related images. In Fig. 3.2, we show another difficult problem of judging the similarity between images when treating them as a whole one. In this case, using the color space, we are unable to confirm the similarity of the two images. Yet, using the salient region (bird in these images) as an addition, we can better represent both images. Therefore, we propose methods to extract features from the three images (i.e. whole, salient regions, and background images) for the AIA task.

### 3.1.4 Contributions

Our main contributions are as follows.

1. We propose to use the *background area* and *salient regions* in conjunction with the *whole image* for AIA. We present a method combining two recently published models to automatically extract salient regions and the background without prior knowledge about the image.
2. We show that we can effectively employ the bag-of-features model on the whole, salient regions and background image. 43 features that cover the holistic content of the image are extracted and used in this chapter ranging from the color, the texture, the scene to local invariant descriptors. With the integrity and diversity of our features, yet the number of the total dimension of our feature is also nearly three times less than that of the ones that have been used in the state-of-the-art approach in [52].
3. We show the strength of our combined features in three settings:
  - over the use of same features extracted from a single whole image,
  - over the use of the same feature set with a grid-based method,
  - over the state-of-the-art results [87] [52] when integrating with their proposed models. It is shown that by using an adhoc combination method [87], we have received a very good performance compared to the same approach. More importantly, by using the more advanced model in [52] which better exploits different features, our feature set beats its performance in many performance metrics.

The structure of this chapter is organized as follows. This section gives the background of the research, formally outlines the problem, the general idea of the chapter and the main contributions. Section 3.2 summarizes the related works. Section 3.3

presents the proposed approach. Section 3.4 gives the experiment settings for evaluation. The detailed results and discussion are presented in section 3.5. Section 3.6 wraps up the finding and provides the future perspectives. It is also noted that all the images illustrated in this chapter are taken from the Corel5K and the ESP Game datasets [31] [133].

## 3.2 Related Works

This section provides the prior works of the research described in this chapter and the context within which the work is situated. Here, we only present the closely related works. We divide the works into two categories, namely, image pre-processing techniques for feature extraction and label propagation techniques.

### 3.2.1 Prior Art in Image Pre-processing Techniques

To increase the efficacy in image representation, researchers have been trying to extract features from local parts of the image in addition to the global image because features that consider the image as a whole cannot describe the local regions effectively. To achieve this, popular approaches are achieved either by first performing image segmentation and then by a feature extraction mechanism, by the use of bag-of-feature model or by the combination of them.

1. In automatic image annotation, two approaches have been employed for the segmentation task: region based and block (also known as tile) based segmentation.
  - (a) The region based approach represents the ideal idea of defining the region for each object in the image. Some popular approaches include color image segmentation [29], normalized cut [113], random walker [46], minimum spanning tree-based segmentation [141] and isoperimetric par-

titioning [47]. However, in many cases, it is a complex algorithm that involves machine learning or uses some prior knowledge.

- (b) In the block based approach, the image is simply split into different blocks of predefined shapes designed to capture some important regions [73] [90] [109] [71] [114] [143] [128] [93]. It is shown in the literature that such decomposition can yield better results than using only one whole image in the image annotation. However, each block does not represent any semantic object unless we know the kind of images that we are dealing with and design the region template accordingly. Usually, it is not possible to create a one-size-fit-all template for every image.
2. In the bag-of-features model [49] [134] [139], often the image or the region of image is first sampled. It can be dense sampled or sampled by points of interest. Additionally, there is another sampling way called spatial pyramid [74] which builds on the top of the two approaches mentioned earlier. In the spatial pyramid sampling, the whole image is divided into blocks or at different resolutions, and the sampling points are selected from each block and aggregated together in order to give significance to sub regions. Then, a vector quantization is performed on the extracted local features from the sampling points, usually by using clustering algorithms. The resulting feature descriptor is a fix-length histogram of the visual occurrence.

Fig. 3.3 summarizes these related techniques in image pre-processing prior to image feature extraction.

### 3.2.2 Prior Art in Label Propagation Techniques

As for keyword propagation, a number of models have been proposed ranging from discriminative [48] [55], generative [16] [94] [20], to nearest neighbor ones (also known as K Nearest Neighbor or KNN). The KNN approach is the special case of the

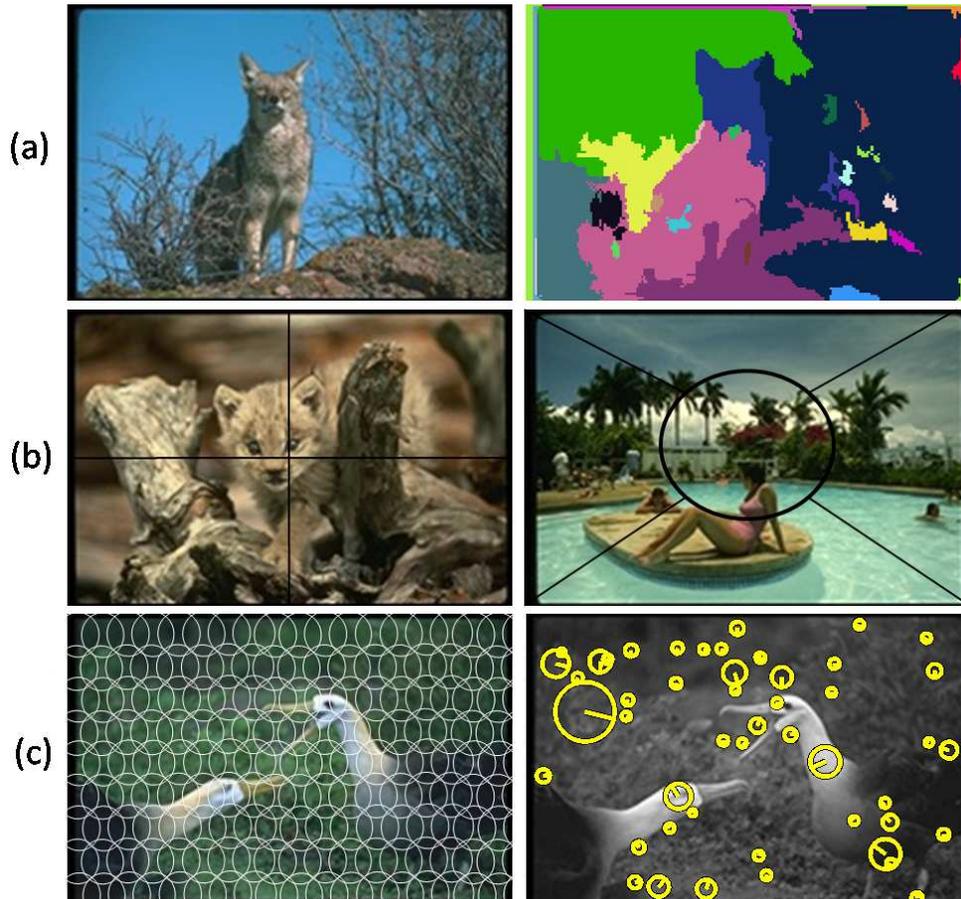


Figure 3.3: Example showing different methods used prior to image feature extraction: (a) the image is segmented into different regions, (b) the image is decomposed into predefined and fixed blocks, (c) the image is dense sampled (left) or is sampled by points of interest (right)

*The Figure is taken from Figure 3 of the author's paper [J1]*

equation 3.1 in which we aim to select a subset of top ranked words of the dictionary  $W$  from the top  $k$  nearest neighbors. The pioneer systems include the Continuous Relevant Model (CRM) [62] and Multiple Bernoulli Relevance Models (MBRM) [35]. Nearest neighbor approaches have gained popularity in recent years due to the availability of larger datasets and the increased computational power. It has been shown that this approach is best suited for the image annotation task particularly for weakly labeled dataset. For instance, Torralba et al. in [126], show that despite the noise when using 80 million images, the accuracy improves consistently with the larger training set. In the recent years, the KNN approaches in [87] and [52] achieved the state-of-the-art performances. Therefore, we use the KNN model for keyword propagation.

## 3.3 The Proposed Approach

### 3.3.1 Overview

It is ideal if we could have a perfect segmentation method where we can separate all the objects inside the image. However, in practice, it is a chicken-and-egg problem because we need to know some information about the image before we can solve this problem. The state-of-the-art approaches are still computationally expensive and introduce an unreliable segmentation. To identify an image, not all the detailed information is needed. Usually, a human observer would focus on some objects of interest or on the background scene. This should also be the case for an AIA system. To suggest relevant keywords for an unknown image, such a system should just need to find all the related images with the same or similar high interest objects and/or background in order to learn the keywords while the role of the whole image is to put constraints on the images found. This simplifies the task because identifying some salient regions is relatively easier compared to the detailed segmentation. Moreover, we do not need a perfect segmentation of the objects of interest. Some rough regions

that show these objects would just be fine. Fig. 3.4 shows the overall architecture of our proposed scheme for *holistic features extraction* in the AIA task. The following sub-sections describe the feature extraction processes of our approach. For keyword propagation, we employ the state-of-the-art techniques described in [87] and [52].

### 3.3.2 Salient Regions and Background Extraction for Holistic Image Representation

A recent progress in salient region detection algorithms convinces us that we could explore its usage for the salient region and the background extraction which serves for the holistic feature representation and thus can give an effective AIA. There has been a large body of works on salient regions extraction using different methods ranging from biologically inspired approaches to methods using real human eye tracking data [61] [64] [10] [57]. Here, we are interested in the model presented in [57] and [10] because of their simplicity and efficiency in terms of accuracy and computational cost.

Hou et al. in [57] proposed a bottom up approach where they make use of the scale invariance of natural image statistics. They calculate a spectral residual as the difference between the original log spectrum and its mean-filtered version. The saliency map is obtained by applying an inverse Fourier Transform to the spectral residual. Given an image  $I$  and its Fourier Spectrum  $f$ , the saliency map of the model can be defined as:

$$S_{\text{spectral residual}}(x, y) = g(x, y) \star F^{-1} [\exp(R(f) + P(f))]^2, \quad (3.2)$$

where  $g(x, y)$  is a Gaussian filter;  $F^{-1}$  is the inverse Fourier Transform;  $R(f) = L(f) - A(f)$  represents the spectral residual ( $L(f)$  is the log spectrum and  $A(f)$  is the general shape of the log spectrum);  $P(f)$  denotes the phase spectrum of the image.

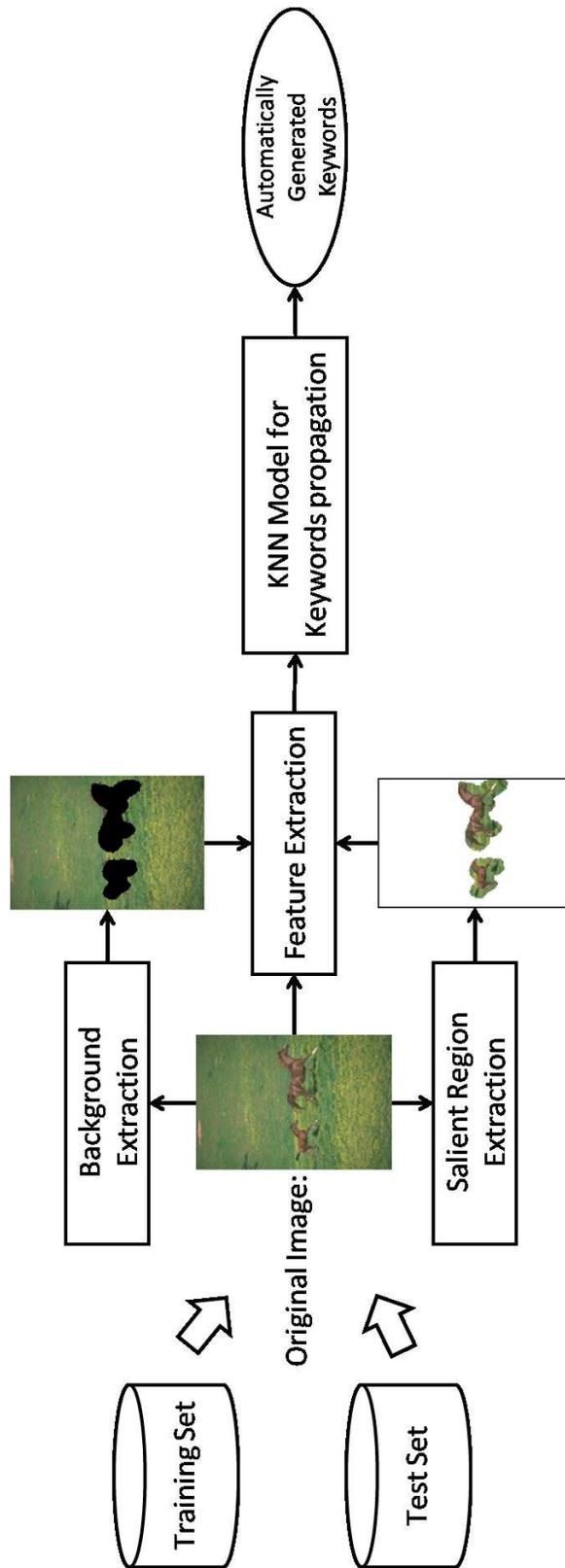


Figure 3.4: Overall architecture of our proposed approach

The Figure is taken from Figure 4 of the author's paper [J1]

Achanta et al. in [10] utilize features of color and luminance for saliency map calculation. Given an image  $I$  in the  $L^*a^*b^*$  color space, the saliency map of the model can be formulated as:

$$S_{frequency\ tuned}(x, y) = \|I_\mu - I_{\omega_{hc}}(x, y)\|, \quad (3.3)$$

where  $I_\mu$  is the mean image feature vector;  $I_{\omega_{hc}}(x, y)$  is the corresponding image pixel  $(x, y)$  vector value in the Gaussian blurred version and  $\| \cdot \|$  is the  $L_2$  norm.

For each model, let  $S_{map}(I)$  be the saliency map of the image  $I$ . We define a threshold for the final saliency cut as  $TH = mean(S_{map}(I)) + std(S_{map}(I))$ .  $TH$  is configured for a better compensation after verifying with a number of empirical tests. Eventually, we compute the final saliency map  $S_{final\ map}(I)$  by rejecting the salient points  $S(x, y)$  that are less than the threshold as:

$$S_{final\ map}(x, y) = \begin{cases} 1 & \text{if } S(x, y) > TH, \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

We take the advantages of both models by performing the union of the saliency maps extracted from each model. Let  $S_{SR}(I)$  and  $S_{FT}(I)$  be the final saliency maps of the image  $I$  from the spectral residual and frequency tuned models respectively, the combined saliency map  $S_{combined}(I)$  is formulated as the following:

$$S_{combined}(I) = S_{SR}(I) \cup S_{FT}(I) \quad (3.5)$$

Then, the background image is calculated accordingly by subtracting the salient regions from the whole image. Fig. 3.5 illustrates the processing steps.

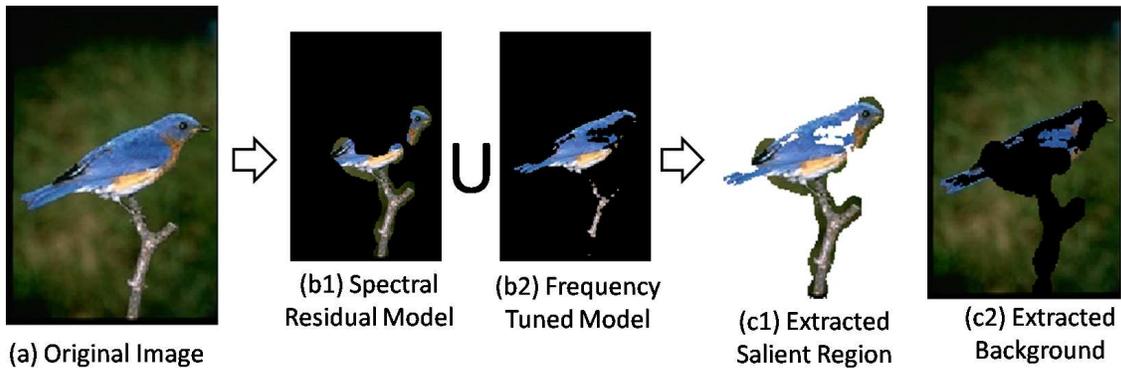


Figure 3.5: Combined model for salient region and background extraction

*The Figure is taken from Figure 5 of the author's paper [J1]*

### 3.3.3 Holistic Feature Extraction

We have studied features that have been proven to be effective in previous works on image annotation and classification using the whole image [111] [131] [99]. As a result, 43 image features  $F = \{f_{colors}, f_{textures}, f_{scenes}, f_{sift\&colorsifts(bag-of-features)}\}$  have been implemented and are described in the following sub-subsections. The Appendix summarizes all the 43 features.

#### 3.3.3.1 Color Features

Color features have been widely used. Though they are among the simplest features, they are important. We have extracted features from 5 color spaces.

- $RGB, L * a * b^*, HSV$ : are simple color histograms in the respective color spaces and computed in 3 channels each with 16 bins.
- *Opponent*: the histogram is calculated as a combination of three 1-D histograms based on the channels of the opponent color space [131].
- $rg$ : since the  $b$  component is redundant in the RGB normalized color space ( $r + g + b = 1$ ),  $r$  and  $g$  are recalculated by eliminating  $b$ . Afterward, the histogram is calculated [131].

### 3.3.3.2 Texture Features

Texture features are important features specifically for distinguishing the region, the surface or detecting objects. Two types of texture features are implemented.

- *Gabor*: a three scales and four orientations filter is used. Then, each of the response images are split into non-overlapping rectangular blocks. We calculate the mean filter response magnitudes from each block over all the twelve response images [87].
- *Haar*: a two by two edge filter is used. The wavelet responses are generated by a block-convolution of an image with Haar filters at three different orientations (vertical, horizontal and diagonal). Convolution with a sub-sampled image is conducted at different scales. Afterward, the image is rescaled to the size 64 x 64 pixels, then a Haar feature is generated by concatenating the Haar response magnitudes [87].

### 3.3.3.3 Scene Feature

Usually, a human observer of an image at a fraction of second can summarize the essential information (gist) about the image such as indoor/outdoor, street, beach, landscape, etc. [36] [104]. The gist descriptors [99] attempts to represent this exquisite ability of humans by describing the spatial layout of an image using global features derived from the spatial envelope. It is shown to be very good in scene categorization. We use the original implementation in [99].

### 3.3.3.4 Advanced Local Invariant Features

SIFT is a powerful local feature and have been confirmed in many publications because of its invariant to scale and orientation [80]. Recently, Color SIFT features have been proposed as extension to SIFT feature which provide additional flexibilities [132] [18] [131] [8].

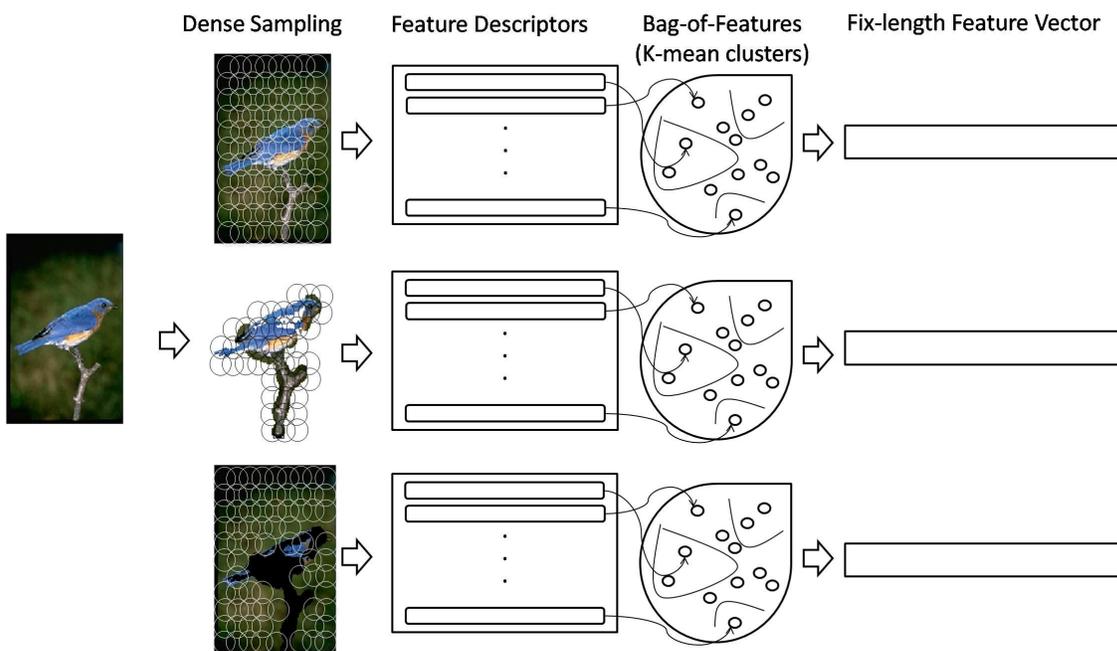


Figure 3.6: Processing steps in local invariant features ( SIFT and Color SIFT) extraction

*The Figure is taken from Figure 6 of the author's paper [J1]*

## SIFT and Color SIFT Descriptor Extraction

We extracted all the 7 SIFT and Color SIFT features.

- *SIFT*: As originally proposed by [80], first, locations of important interest points in the image are detected by a set of Difference of Gaussian filters applied at different scales of the image. Next, these locations are refined by removing points of low contrast. Each key point is then assigned with an orientation. Afterward, at each key point, the local feature descriptor is computed. This descriptor is based on the local image gradient and is transformed following the orientation of the key point in order to provide orientation invariance.
- *HueSIFT*: It is computed by a concatenation of the hue histogram with the SIFT descriptor.
- *HsvSIFT*: The descriptor is extracted by computing SIFT over all the three channels of HSV.
- *OpponentSIFT*: The descriptor describes all the channels in the *Opponent* color space using SIFT descriptors.
- *rgSIFT*: Descriptors are added for the *r* and *g* components of the normalized RGB color model. Then, for every normalized channel, the SIFT descriptor is computed.
- *C – SIFT*: Utilizes the *C* or the normalized opponent color space. SIFT is computed accordingly.
- *RGBSIFT*: SIFT descriptors are computed for every RGB channel independently.

## Point Sampling Strategy

In our setting, we employ dense sampling with an interval of 6 pixels for all the three images. A honey-rate structure is used by applying a sample spacing of 3 pixels.

## Bag-of-Features Model

For each feature, descriptors are calculated from each sampling point. We randomly use 125,000 of them. Next, they are clustered to form codebooks of size 512 using the K-mean algorithm. The total number of descriptors used for clustering and the number of clusters are rather small. Usually, the number of descriptors for clustering can be up to millions and the codebook size can be as many as 4096 or more. We purposefully chose this configuration for less computational cost. Finally, a fix-length feature vector of size 512 for each image is constructed for each feature. Fig. 3.6 shows the processing steps in features extraction for these advanced local invariant features. We made use of the software described in [80], by adapting it to our case.

### 3.3.4 Experiment Setting

In this section, we describe the datasets and the metrics used to assess the performance of our system as well as the validation procedure.

#### 3.3.4.1 Datasets

We have considered two publicly available datasets mainly because of the different nature of the images as well as the capability to compare with the state-of-the-art methods [88] [52] [35].

##### (i) Corel5K

The Corel5K dataset [31] originates from the Corel stock photo collection. It is a collection of 5,000 images including 4,500 images as the training set. Many kinds of

	Corel5K	ESP Game
Image size	128 x 192	variable
Vocabulary size	260	268
Number of training image	4500	18689
Number of test image	500	2081
Average number of words per image	3.4	4.7
Maximum number of words per image	5	15

Table 3.1: Statistics of the two datasets: Corel5K and ESP Game.

images are presented in the dataset from sunset to sport and portrait. Each image is labeled to describe the main objects. The annotation is assigned to have from one to five keywords. There are 371 keywords but only 260 appear in both train and test sets. It is arguably the most used collection in image annotation and retrieval research.

### (ii) ESP Game

The ESP game [133] is a recent dataset collected over the Internet through means of social labeling game. It has diverse contents of web images from personal photos to drawings and logos. Only a subset of the collection (20,770 images) is used in this chapter for fair comparison with other published methods [88] [52]. A total of 268 keywords can be found in both training and test sets.

Table 3.1 summarizes the properties of the two datasets.

#### 3.3.4.2 Performance Metrics

We perform our evaluation based on a number of different metrics as described in the following.

##### (i) Fix-length Precision, Recall, and Recalled keywords

We compute precision, recall and the coverage rate of keywords. For a given keyword, let  $N_H$  be the number of images labeled with the keyword in the ground-truth;  $N_{App}$

be the number of images that are assigned with the keyword by the system; and  $N_C$  be the number of images that are correctly assigned. The precision ( $P$ ) is defined as  $\frac{N_C}{N_{App}}$ ; recall ( $R$ ) is formulated as  $\frac{N_C}{N_H}$ ; and the coverage rate of keywords ( $N+$ ) is the number of keywords with a positive recall. We report the average of each measure. It is noted that each image is assigned with 5 keywords in this experiment setting, although some may have more or less than this number in the ground-truth.

### (ii) Precision at Different Levels of Recall (PDLR)

For PDLR, we calculate the Mean Average Precision ( $MAP$ ) and Break-Even Point ( $BEP$ ) (also known as R-Precision) following [48] and [52].  $MAP$  is the average of the precision at each position where a relevant image is retrieved, defined as  $\frac{1}{|R(w)|} \sum_{I \in R(w)} Pr(rk(w, I))$  where  $rk(w, I)$  is the rank of an image  $I$  for a query  $w$ .  $BEP$  gives the percentage  $Pr(|R(w)|)$  in the top  $|R(w)|$  ranking position. To measure the auto-annotating performance, we calculate iMAP and iBEP by changing the role of the keyword and the image as proposed in [51]. iMAP measures the average precision over the images while iBEP is the break-even point accordingly.

### (iii) Success, Draw and Worse Results in MAP Distribution

We compute and compare the performance of our best features with those of other features as well as state-of-the-art results in terms of the number of worse, draw and better results of the MAP distribution of both the keywords and the images.

#### 3.3.4.3 Validation Procedure

The objective of this experiment is threefold. The first two goals are to show the superiority of our approach versus the use of a single whole image, and the grid-based approach with the same feature set. The third goal is to show that we can effectively employ our feature set with the state-of-the-art methods to beat their performances. For each metric, we present 7 results using different combinations of

features:

1. *whole*: only features from the whole image are used. The total number of features used is 15.
2. *roi*: only features from salient regions (also known as region of interests or roi) are used. The total number of features used is 14.
3. *bg*: only features from background are used. Total number of features used is 14.
4. *whole + roi*: features from the whole image and salient regions are used. The total number of features used is 29.
5. *whole + bg*: features from the whole image and the background are used. The total number of features used is 29.
6. *roi + bg*: features from salient regions and the background are used. The total number of features used is 28.
7. *whole + roi + bg*: features from the whole images, salient regions and the background are used. The total number of features used is 43.

In addition to proving that our best feature set (*whole+roi+bg*) gives a better performance than that of the state-of-the-art, we also give evidences that our proposed method is better than the conventional approach that uses only the *whole* image. To further prove the effectiveness of our approach, we also compare it with a grid-based approach with the same feature set. In the grid-based approach, we assume that salient regions are always at the center of the image. For a fair comparison, we consider the square-size region at the middle part of the image as the salient region and the rest as its background. Fig. 3.7 shows two example images and their respective salient region and background images. We extract the same set of features from the background and the salient region as in our approach. It is noted that for this case,



Figure 3.7: Grid-based salient regions and background extraction

*The Figure is taken from Figure 7 of the author's paper [J1]*

the experiment is only conducted on the Corel5K dataset because the ESP Game one includes some square-size images. We refer to this method as *Grid* for the rest of this chapter.

For statistical proof, we calculate the *sign test* of different metric distributions to reject the null hypothesis. The sign test is chosen because we do not want to assume the type of distribution of our results. In all cases, a *P - value*  $< 0.05$  is demanded in order to be statistically significant.

### 3.4 Results

Since the first two goals mentioned earlier can be encapsulated in the third one, we divide the results by the state-of-the-art label propagation techniques, namely, the

joint equal contribution and tagprop models.

### 3.4.1 Joint Equal Combination Model

#### 3.4.1.1 Joint Equal Combination Annotation Scheme

Makadia et al. in [88] introduced a simple yet efficient approach. The method called Joint Equal Contribution (JEC) simply combines all the features equally and the propagation is done by transferring the keywords from the nearest neighbors via the KNN scheme. Let  $d(i, j)$  be the combined distance of image  $i$  and  $j$ . If  $\tilde{d}_{(i,j)}^k$  is the scaled distance of feature  $k$ , then

$$d(i, j) = \frac{1}{N} \sum_{K=1}^N \tilde{d}_{(i,j)}^k \quad (3.6)$$

We present the results using our implemented approach with our proposed features and compare with the recently proposed works. Table 3.2 gives the summary of the comparison.

#### 3.4.1.2 Results

From the results, we can infer that our features (total combination: *whole+roi+bg*) give a better performance than other methods in most of the metrics. We received a superior performance except for recall (R) in the ESP Game dataset than those of [88] which in turn beats all the results before 2008. We especially maximize the number of keywords which means it is very good in terms of generalization. Our features also give better results than those used in the state-of-the-art results [52] in this combination scheme. Here, we only report the basic fix-length performance because we do not have the other metric results of other papers for this JEC scheme. Tables 3.3 and 3.4 present the comparison between *whole* and *whole+roi+bg*, and between *whole+roi+bg* of our approach and the grid-based one. For a detailed comparison, we calculate the MAP of all possible combinations of queries (maximum

	Corel5K			ESP Game		
	P	R	N+	P	R	N+
MBRM	24	25	122	18	19	209
JEC	27	32	139	22	25	224
JEC-15	28	33	140	24	19	212
Our work (JEC): whole	26.9	35.5	144	23.9	23.6	240
Our work (JEC): roi	11.7	9.3	59	35.9	14.3	223
Our work (JEC): bg	23	31.3	140	23.1	21.7	232
Our work (JEC): whole + roi	29.1	34.7	151	24.6	21.8	241
Our work (JEC): whole + bg	27.3	35.4	151	23.7	22.9	235
Our work (JEC): roi +bg	22.2	26.6	129	26.1	20.1	236
Our work (JEC): whole + roi + bg	28.8	36.2	156	24.1	22.5	241
Our work (JEC): whole + roi + bg	27.2	34.2	150	N/A	N/A	N/A

Table 3.2: Summary of performance comparison when using our features with the JEC approach. Note that JEC-15 is the result reported in [52] of the JEC method using their 15 features.

	Corel5K	ESP Game
	MAP (A)	MAP (A)
Our work (JEC): whole	21.0	9.1
Our work (JEC): whole + roi + bg	<b>21.1</b>	<b>9.2</b>
P-value (Sign Test)	$8.34 \times 10^{-34}$	$1.45 \times 10^{-161}$

Table 3.3: Performance comparison when using only whole image versus whole+roi+bg in terms of MAP (A)

size of 5). It is shown that *whole+roi+bg* gives a higher performance than a single *whole* for both datasets. It is also confirmed that our approach is better than the grid-based one. The results are statistically significant with p-value of sign test  $p \ll 0.05$ . In short, the results confirm the strength of our integrated features as well as our approach. We provide further analysis in the next section.

## 3.4.2 TagProp Model

### 3.4.2.1 TagProp Annotation Scheme

TagProp [52] generalizes the approach in [88] by introducing the weight of each feature and has become the current state-of-the-art. Since we implement the model,

	Corel5K
	MAP (A)
Grid (JEC): whole + roi + bg	21.0
Our work (JEC): whole + roi + bg	<b>21.1</b>
P-value (Sign Test)	$8.34 \times 10^{-34}$

Table 3.4: Performance comparison between our proposed approach and the grid-based one in terms of MAP (A)

we briefly describe the method and the features used for a quick overview.

### (i) Model

TagProp makes use of the Bernoulli model for keyword representation because keywords are either present or absent. Let  $y_{iw} \in \{+1, -1\}$  denotes the absence or presence of a keyword, the keyword presence prediction  $p(y_{iw} = +1)$  for an image  $i$  is defined as a weighted sum over the training images, indexed by  $j$ :

$$p(y_{iw} = +1) = \sum \pi_{ij} p(y_{iw} = +1|j), \quad (3.7)$$

while  $\pi_{ij}$  is the weight of image  $j$  for predicting the keywords of image  $i$ . In other words, it is the probability to use the image  $j$  as a neighbor for the image  $i$ . It can be defined using the image rank or the image distance. We are interested in the image distance based variant which is more suitable to represent different distances according to the feature:

$$\pi_{ij} = \frac{\exp(-\rho^T d(i, j))}{\sum_{j'} \exp(-\rho^T d(i, j'))}, \quad (3.8)$$

while  $j' \in J$  is the subset of the  $k$  most similar images to  $i$ . The weights of the rest of images can be set to 0.  $d(i, j')$  is the vector of each base distance between image  $i$  and  $j$ . They maximize the log-likelihood of the prediction of the training set to estimate the parameter  $\rho$  that control  $\pi_{ij}$  as  $L = \sum_{i,w} c_{iw} \ln p(y_{iw})$ , where  $c_{iw}$  is the cost of the imbalance between keyword presence and absence.  $c_{iw} = \frac{1}{n^+}$  if  $y_{iw} = +1$  and  $c_{iw} = \frac{1}{n^-}$  if  $y_{iw} = -1$ . The model is extended to incorporate the

word-specific logistic discriminant to boost the recall among the rare annotation.

## (ii) Features

15 distinct features are used in TagProp: 1 gist descriptor, 6 color histograms including RGB, L\*a\*b\*, HSV, and 8 local bag-of-features (2 features types x 2 descriptors x 2 layouts) including SIFT and HUE resulted in 32752 dimensions.

We have implemented the model using the information in the paper and their code available on the website<sup>1</sup>. We also used their published features. We got a similar performance but did not get the claimed results. This might be due to some small parameters or feature normalization that are different since only the code of the model is provided. We use the default setting parameters. We list down both results: the original ones noted as TagProp and our implementation noted as TagProp\* for fair comparison. It is generally noted that TagProp\* has better precision rates than the original ones but suffers in recall rates and the number of keywords as shown in Tables 3.5 and 3.6.

### 3.4.2.2 Performance as Image Retrieval from Single-keyword Queries Task

In this setting, we divide the results into two categories, namely, fix-length and precision at different recall levels. Tables 3.5 and 3.6 summarize the results of Corel5K and ESP Game, respectively. In the fix-length mode, we achieve better results than the implemented state-of-the-art performance (TagProp\*) in all the 3 metrics ( $P$ ,  $R$  and  $N+$ ) and on both datasets. In the other mode, we obtain less MAP and BEP in the Corel5K dataset but beat the state-of-the-art results in the ESP Game dataset. We believe that this is because our feature set tends to produce the holistic description about the content of the images, while Corel5K images are

---

<sup>1</sup><http://lear.inrialpes.fr/people/guillaumin/code/>

Approach	Corel5K				
	Fixed-length			PDLR	
	P	R	N+	MAP	BEP
TagProp	32.7	42.3	160	41.8	36.3
TagProp*	33.5	37.5	153	42.4	37.3
Our work (TagProp): whole	31.7	37.3	147	38.1	34.5
Our work (TagProp): roi	22.6	29.2	127	30	26
Our work (TagProp): bg	26.5	33.1	137	35.2	31.3
Our work (TagProp): whole + roi	32.9	39.8	154	39.4	36.5
Our work (TagProp): whole + bg	31.3	37.6	147	38.7	35.3
Our work (TagProp): roi + bg	28.7	36.8	141	37.2	32.3
Our work (TagProp): whole + roi +bg	<b>34.8</b>	<b>40.6</b>	<b>160</b>	<b>39.9</b>	<b>36.5</b>
Grid (TagProp): whole + roi + bg	31.1	36.7	147	38.6	35.0

Table 3.5: Performance comparison between our work and the state-of-the-art methods for the Corel5K dataset. Note that TagProp is the original results claimed in 7). TagProp\* is our implementation of the results using the same features, the portion of the code provided by the authors in their website and the same number of neighbors ( $k = 200$ )

not labeled with all the possible keywords. This problem has been addressed in the literature. We will discuss the problem again in the next subsection when we perform detailed analysis. Beside this, our approach beats all other approaches including the use of a single whole image and the grid-based approach in both datasets.

It is noted that we have reached our results presented in Tables 3.5 and 3.6 with only 100 and 170 as the number of nearest neighbor  $k$  for Corel5K and ESP Game datasets, respectively. Though we do not get better results using a larger  $k$ , this shows the importance of having diverse features because we can accumulate more related images with less  $k$ .

### 3.4.2.3 Performance as Image Retrieval from Multi-keywords Queries Task

In order to give a better insight on the effectiveness of our system, we measure the performance in multi-keywords queries. To allow for direct comparison, as in [52] [48], we use a subset of 179 of the 260 keywords of the Corel5K dataset that

Approach	ESP Game				
	Fixed-length			PDLR	
	P	R	N+	MAP	BEP
TagProp	39.2	27.4	239	28.1	31.3
TagProp*	41.3	20.7	226	23.8	26.4
Our work (TagProp): whole	42.2	22.8	231	26.2	29.2
Our work (TagProp): roi	41.1	20.2	226	22.7	25.6
Our work (TagProp): bg	40.2	21.5	225	24.3	26.8
Our work (TagProp): whole + roi	42.5	23	232	26.4	29.2
Our work (TagProp): whole + bg	42.2	22.8	231	26.2	29.2
Our work (TagProp): roi + bg	41.7	22.7	230	25.4	28.4
Our work (TagProp): whole + roi +bg	<b>43.1</b>	<b>23.2</b>	<b>233</b>	<b>26.4</b>	<b>29.4</b>

Table 3.6: Performance comparison between our work and the state-of-the-art methods for the ESP Game dataset.

appear at least twice in the dataset. The keywords queries are divided into easy, hard, single and multiple. Easy queries are those that have more than 3 relevant images while hard queries have at most 2 relevant images. Images are considered relevant when they are annotated by all the query keywords. We follow the same setting for the ESP Game dataset. We use all the 268 keywords because they appear in both testing and training sets and more than once. The maximum number of multiple keywords is set to 5 in both datasets.

We arrive at the results presented in Tables 3.7 and 3.8.  $MAP(S)$ ,  $MAP(M)$ ,  $MAP(E)$ ,  $MAP(H)$ , and  $MAP(A)$  are  $MAP$  results for *single*, *multiple*, *easy* and *hard* queries, respectively. In the Corel5K dataset, we obtain a better performance when comparing to *whole-only* and grid-based approaches in all the metrics. As expected, we achieve good performance in easy queries. First, it is because of the diverse range of our features from salient regions and the background that help finding more related images. Second, the easy queries usually target specific objects such as *sun*, *flower*, *person*, *building*, etc. Although we obtain less point in  $MAP(S)$  comparing to TagProp\*, we obtain the same performance in other  $MAP$  metrics and we still receive the same overall performance of  $MAP$  and  $BEP$  in this dataset.

Corel5K							
Approach	MAP (S)	MAP (M)	MAP (E)	MAP (H)	MAP (A)	BEP (A)	
PAMIR	34	26	43	22	26	17	
TagProp	46	35	55	32	36	27	
TagProp*	45	35	54	31	36	27	
Our work (TagProp): whole	42	34	54	30	34	26	
Our work (TagProp): roi	35	26	45	23	27	19	
Our work (TagProp): bg	40	31	51	27	32	23	
Our work (TagProp): whole + roi	43	36	55	32	36	27	
Our work (TagProp): whole + bg	43	34	54	30	34	26	
Our work (TagProp): roi + bg	41	33	54	29	34	25	
Our work (TagProp): whole + roi + bg	<b>44</b>	<b>35</b>	<b>56</b>	<b>31</b>	<b>36</b>	<b>27</b>	
Grid (TagProp): whole + roi + bg	42	33	54	29	34	26	

Table 3.7: Performance comparison between our work and the state-of-the-art methods in terms of multi-keyword queries in the Corel5K dataset.

Approach	ESP Game						
	MAP (S)	MAP (M)	MAP (E)	MAP (H)	MAP (A)	BEP (A)	
PAMIR	-	-	-	-	-	-	-
TagProp	-	-	-	-	-	-	-
TagProp*	24	15	18	15	15	10	10
Our work (TagProp): whole	26	16	19	16	16	10	10
Our work (TagProp): roi	23	14	17	14	14	9	9
Our work (TagProp): bg	24	15	17	15	15	10	10
Our work (TagProp): whole + roi	26	16	19	16	16	10	10
Our work (TagProp): whole + bg	26	16	19	16	16	10	10
Our work (TagProp): roi + bg	25	15	18	15	15	10	10
Our work (TagProp): whole + roi + bg	26	16	19	16	16	10	10

Table 3.8: Performance comparison between our work and the state-of-the-art methods in terms of multi-keyword queries in the ESP Game dataset.

In the ESP Game dataset, we attain better performance in every scale except for BEP(A). The good performance comes from the fact that the images from this dataset usually have one clear concept. The dataset also contains diverse ranges of web images and has a relatively large number of training set. Moreover, the test set is also relatively large compared to the Corel5K one and includes a variety of images. The bad performance in BEP is due to the large gap between the minimum and maximum number of keywords in the ground truth.

To further prove that the combination of *whole+roi+bg* is more effective than the use of a single *whole* image, and that our approach is better than the grid-based one, we compare the MAP results between the approaches. We compute the *p* – *value* of the sign test. Tables 3.9 and 3.10 summarize the results of the Corel5K dataset. It is shown that in all the metrics the higher performance of our approach and the combined feature set is statistically significant by the low value of  $p \ll 0.05$ . Table 3.11 shows that the better performance of our method in the ESP Game dataset is statistically significant for the easy, multiple, hard and all queries. Although the *p* – *values* of MAP(S) and BEP(A) are superior to 0.05, we can still observe the improvement in the result sets. The next subsection shows some examples of the retrieval task.

In overall, our approach and feature set give better performance in most of these keyword retrieval metrics for both datasets.

#### 3.4.2.4 Some Qualitative Results in the Retrieval Task

Here, we present two retrieval examples for each dataset to illustrate and compare the performance of our method to the ones from the baselines. The first is a single query retrieval task and the second one is a multiple query one. Fig. 3.8 and 3.9 show the tasks in the Corel5K dataset and the ESP Game dataset respectively. The resulting images are sorted by the level of relevancy. Seven images are shown for

	Corel5K					
	MAP (S)	MAP (M)	MAP (E)	MAP (H)	MAP (A)	BEP (A)
Our work (TagProp): whole	42.40	33.72	54.26	29.82	34.41	26.03
Our work (TagProp): whole + roi + bg	43.75	34.99	55.68	31.07	35.69	27.07
P-value (Sign Test)	$9 \times 10^{-5}$	0.0003	$9 \times 10^{-8}$	0.0156	$4 \times 10^{-6}$	0.0001

Table 3.9: Performance comparison when using only whole image versus whole+roi+bg in terms of multi-keyword queries of the Corel5K dataset.

		Corel5K						
		MAP (S)	MAP (M)	MAP (E)	MAP (H)	MAP (A)	BEP (A)	
Grid (TagProp): whole + roi + bg		42.37	33.33	54.41	29.34	34.05	25.75	
Our work (TagProp): whole + roi + bg		43.75	34.99	55.68	31.07	35.69	27.07	
P-value (Sign Test)		0.0372	0.0003	0.0001	0.0096	$4.38 \times 10^{-5}$	0.0137	

Table 3.10: Performance comparison between our proposed approach and the grid-based one in terms of multi-keyword queries of the Corel5K dataset.

	ESP Game					
	MAP (S)	MAP (M)	MAP (E)	MAP (H)	MAP (A)	BEP (A)
Our work (TagProp): whole	26.19	15.96	18.66	15.85	15.99	10.50
Our work (TagProp): whole + roi + bg	26.36	16.08	18.88	15.96	16.11	10.57
P-value (Sign Test)	0.1995	$9 \times 10^{-65}$	0.0001	$8 \times 10^{-62}$	$4 \times 10^{-65}$	0.0671

Table 3.11: Performance comparison when using only whole image versus whole+roi+bg in terms of multi-keyword queries of the ESP Game dataset.

each query in each method.

These result sets show that our approach give the most relevant outputs when comparing with the same top  $n$  images, thanks to the features extracted from the salient regions and the background. It is also noted that the grid-based approach performs quite well. This is because many of the images in the Corel5K dataset have the salient objects placed in the middle of the image and thus our setup to extract the squared center of the image is quite generous. Even though, our approach still performs better.

#### 3.4.2.5 Image Auto-annotating Performance

So far, we measure the performance of the annotation as a search task. It is also very important to measure how relevant our suggested keywords are. This is particularly essential for the interactive recommendation task as well as auto-annotating. Table 3.12 reports the performance results for this case.

It is noted that there is no report on iBEP and iMAP in the original paper of TagProp in [52]. It is shown that we receive very good results comparing to the state-of-the-art ones. In the Corel5K dataset, we gain about 8 and 10 points in iMAP and iBEP, respectively. We also get 2 points higher of both measures in the ESP Game dataset. With these results, we can be sure that more than half of the suggested keywords are relevant in the case of the Corel5K dataset and about 40% of relevancy rate can be achieved in the case of the ESP Game dataset.

Table 3.13 reports the results of the comparison between our proposed integrated feature versus the use of only *whole* image. It is shown that our approach leads to better performance for both metrics (iMAP and iBEP) and for both datasets. In Table 3.14, the improvement over the grid-based approach could not lead us to reject the null hypothesis by the calculated p-value. As discussed earlier, we believe this is

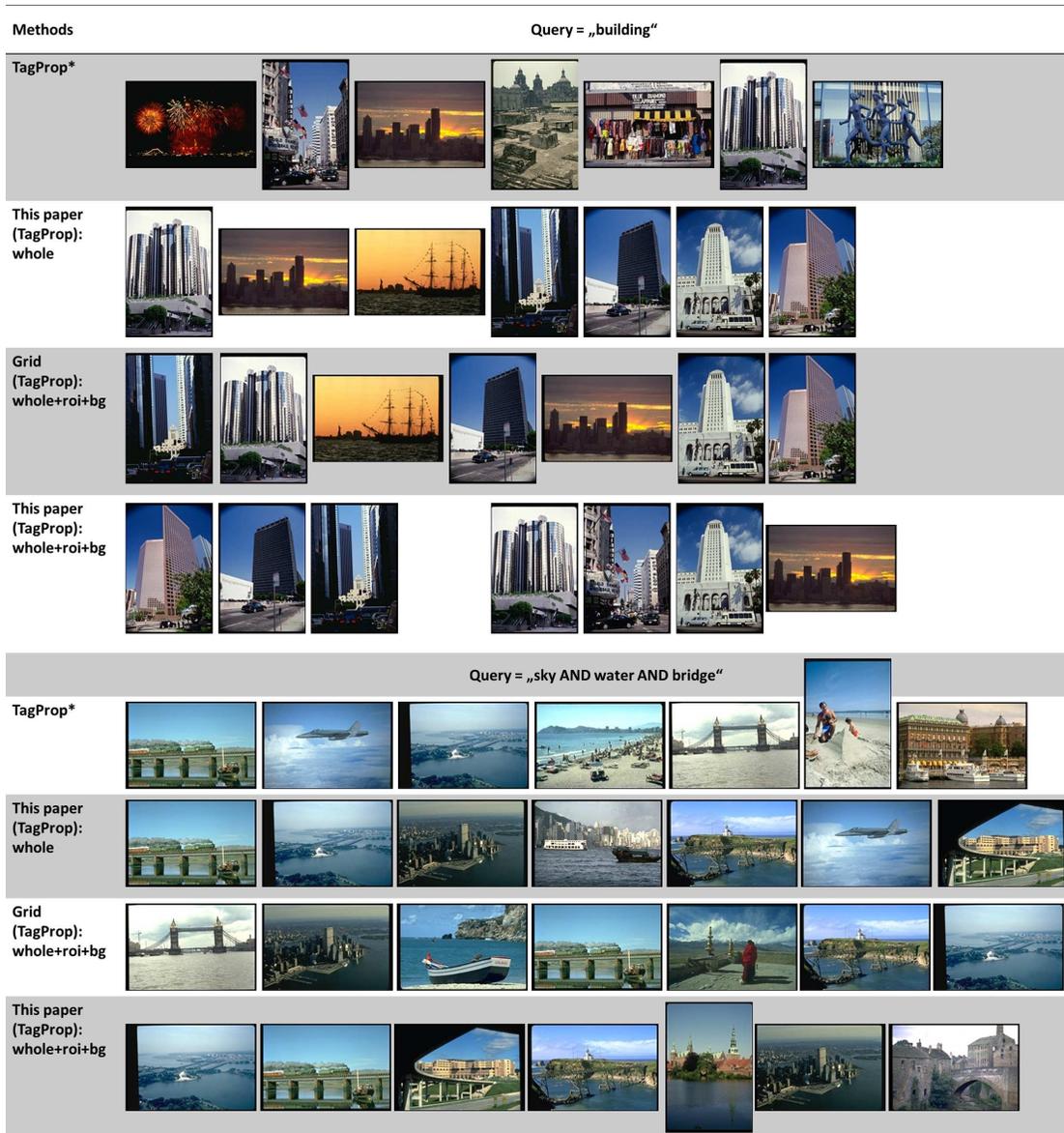


Figure 3.8: Corel5K dataset retrieval examples in comparison with the baseline approaches

*The Figure is taken from Figure 8 of the author’s paper [J1]*

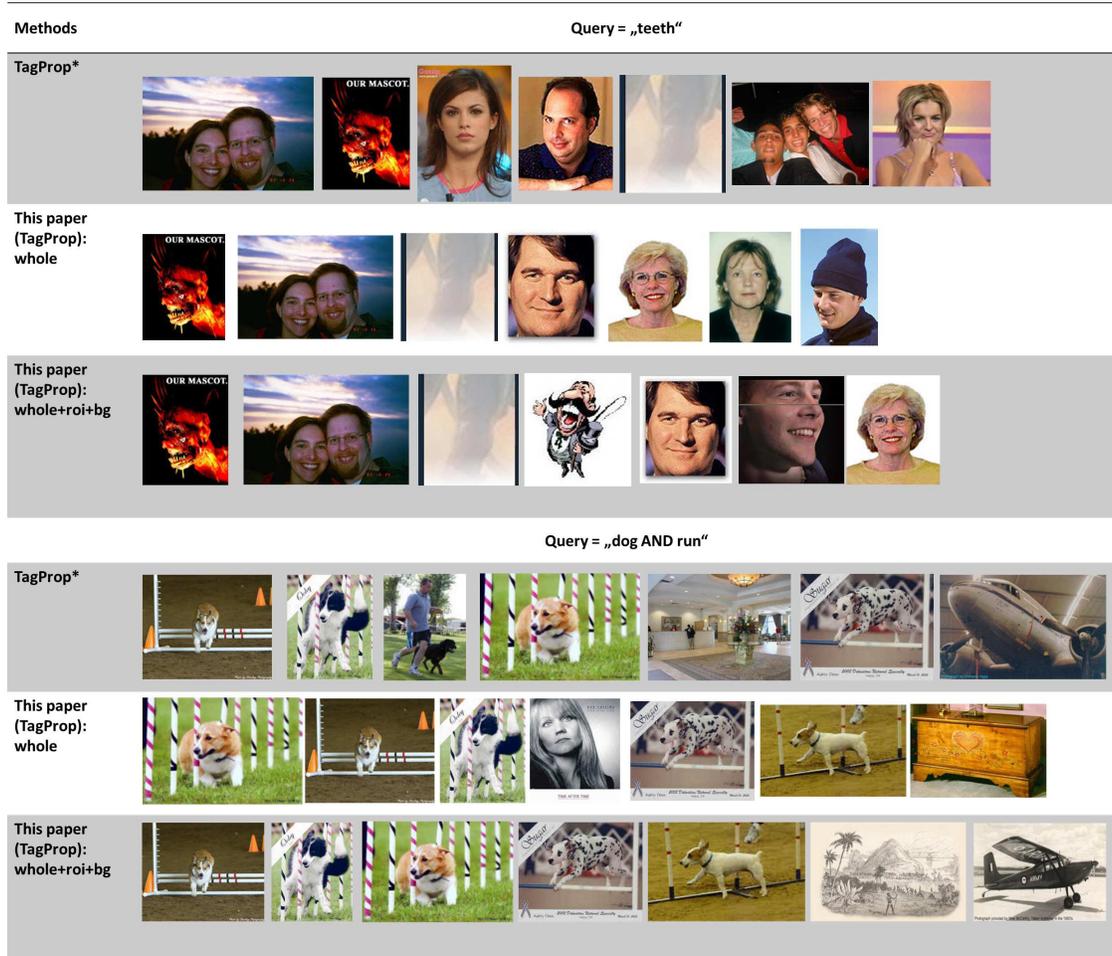


Figure 3.9: ESP game dataset retrieval examples in comparison with the baseline approaches

*The Figure is taken from Figure 9 of the author’s paper [J1]*

	Corel5K		ESP Game	
	iMAP	iBEP	iMAP	iBEP
TagProp	-	-	-	-
TagProp*	49.7	42.1	40.7	36.5
Our work (TagProp): whole	56.6	50.7	42.3	38.1
Our work (TagProp): roi	48.7	43.4	39.6	35.8
Our work (TagProp): bg	53.2	48.6	40.1	36.4
Our work (TagProp): whole + roi	57.7	52.5	42.7	38.6
Our work (TagProp): whole + bg	57	51.6	42.3	38.1
Our work (TagProp): roi + bg	56	50.9	41.9	37.9
Our work (TagProp): whole + roi + bg	<b>57.9</b>	<b>52.7</b>	<b>42.8</b>	<b>39</b>
Grid (TagProp): whole + roi + bg	57.5	51.5	N/A	N/A

Table 3.12: Summary of performance of our auto-annotating performance

	Corel5K		ESP Game	
	iMAP	iBEP	iMAP	iBEP
Our work (TagProp): whole	56.58	50.74	42.37	38.14
Our work (TagProp): whole + roi + bg	<b>57.93</b>	<b>52.71</b>	<b>42.80</b>	<b>39.07</b>
P-value (Sign Test)	0.0283	0.0065	0.0335	0.0292

Table 3.13: Performance comparison when using only whole image versus whole+roi+bg in terms of our auto-annotating performance

because of the favor of the Corel5K dataset for our salient region extraction setting of the grid-based approach. However, we will show in the examples that follow that this improvement can be observed and it is important. Furthermore, we will show the performance in terms of the number of worse, draw and better results in subsection 5.2.7.

	Corel5K	
	iMAP	iBEP
Grid (TagProp): whole+roi+bg	57.55	51.53
Our work (TagProp): whole+roi+bg	57.93	52.71
P-value (Sign Test)	0.6567	0.3559

Table 3.14: Performance comparison between our approach and the grid-based one in terms of auto-annotating performance

### 3.4.2.6 Some Qualitative Results in the Annotation Task

This subsection shows some qualitative annotation results of the two datasets. Figures 3.10 and 3.11 show the result sets in the ESP Game and Corel5K datasets, respectively. For each feature and method, we show a generated five-keyword annotation. It is once again observed that our approach gives the best annotations when comparing with the ones from the baselines. When the salient regions or the background are distinctive, our approach gets a very good recall in terms of keyword. It still gets similar performance with the others for rather complex images.

### 3.4.2.7 Number of Worse, Draw and Better Results of Keyword-wise and Image-wise Precision

We compute the results from all the 260 and 268 keywords and from 500 and 2081 test images in Corel5K and ESP Game, respectively. Tables 3.15 and 3.16 give the results in keyword-wise for Corel5K and ESP Game datasets. Tables 3.17 and 3.18 show the results in image-wise for the Corel5K and the ESP Game respectively. In general, the results follow the trend of results we showed earlier in retrieval performance (keyword) and auto-annotation (image). However, they present additional information. For instance, Table 3.17 shows that we get a better image-wise precision in 281 of the total 500 images versus TagProp\*. For the ESP Game dataset, we obtain 189/268 (see Table 3.16) and 1152/2081 (see Table 3.18) as the numbers of *better results* in keyword-wise and image-wise performance versus TagProp\*. As for the comparison between *whole+roi+bg* and *whole*, the Tables 3.15, 3.16, 3.17 and 3.18 show that our approach leads to a larger number of *better results* than *worse ones* in all conditions. In the case of our approach versus the grid-based one (see Tables 3.15 and 3.18), it is shown that for keyword-wise, we lose to the grid-based by about 38% (99/260) but we are better in 49% (129/260) of the 260 keywords. We believe that these results are significant. In image-wise, we also gain a higher number of better results than the worse ones.

Images					
TagProp*	hat,man,people,white,woman	car,green,rock,tree,water	blue,chart,graph,line,white	blue,man,sky,water,white	blue,man,red,water,white
This paper (TagProp): whole	black,eat,man,people,woman	grass,green,river,tree,water	blue,chart,circle,red,white	blue,car,man,sky,white	boat,car,man,water,white
This paper (TagProp): whole+roi+bg	<u>black,man,people,white,woman</u>	<u>green,man,road,tree,water</u>	<u>circle,logo,red,square,white</u>	<u>blue,man,snow,water,white</u>	<u>boat,man,sky,water,white</u>
Ground Truth	anime,beard,black,cartoon,cloud,man,sky,yellow	man,photo,tree	circle,flag,red,sun	man,mountain,snow,tree,white	boat,man,red,water
Images					
TagProp*	colors,country,flag,red,square	blue,man,people,sky,water	building,painting,people,sky,yellow	blue,hat,man,red,white	blue,ocean,sea,sky,water
This paper (TagProp): whole	colors,country,flag,red,square	blue,man,sky,statue,tower	green,man,people,table,tree	grass,green,man,people,tree	cloud,ocean,sea,sky,water
This paper (TagProp): whole+roi+bg	<u>country,flag,rectangle,red,square</u>	<u>blue,building,sky,statue,tower</u>	<u>building,green,people,tower,tree</u>	<u>dog,grass,man,people,woman</u>	<u>cloud,mountain,sea,sky,water</u>
Ground Truth	flag,rectangle,red,shadow	blue,building,sky,statue,tower	cloud,crowd,people,sky,tower,white	black,dog,ear,fur,grass,nose,pole,red,run	cloud,mountain,sea,sky,sun,water

Figure 3.10: ESP dataset annotation examples in comparison with the baseline approaches

*The Figure is taken from Figure 11 of the author's paper [J1]*

Images					
TagProp*	sky,jet,plane,prop,smoke	water,beach,cars,tracks,turn	tree,forest,cat,tiger,bengal	grass,field,cat,tiger,bengal	tree,coyote,cars,tracks,turn
This paper (TagProp): whole	sky,jet,plane,prop,smoke	sky,water,tree,people,train	tree,forest,cat,tiger,bengal	grass,field,cat,tiger,bengal	water,tree,people,grass,elk
Grid (TagProp): whole+roi+bg	sky,jet,plane,prop,smoke	sky,water,tree,people,train	tree,forest,cat,tiger,bengal	grass,field,cat,tiger,bengal	water,tree,people,grass,forest
This paper (TagProp): whole+roi+bg	<u>sky,jet,plane,flight,smoke</u>	<u>sky,water,tree,beach,island</u>	<u>head,forest,cat,tiger,bengal</u>	<u>grass,field,cat,tiger,bengal</u>	<u>water,tree,grass,rocks,coyote</u>
Ground Truth	sky,jet,plane	sky,water,pool	forest,cat,tiger,bengal	grass,cat,tiger,bengal	tree,snow,forest,coyote
Images					
TagProp*	tree,grass,ground,rocks,tiger	water,boats,cars,tracks,turn	light,cars,cathedral,tracks,turn	leaf,flowers,plants,needles,cactus	tree,people,buildings,light,shops
This paper (TagProp): whole	tree,flowers,cat,tiger,den	mountain,water,boats,valley,desert	sky,people,buildings,light,night	tree,leaf,grass,flowers,plants	people,buildings,flowers,light,shops
Grid (TagProp): whole+roi+bg	tree,grass,flowers,close-up,fox	mountain,sky,water,boats,people	sky,people,buildings,light,night	tree,leaf,flowers,plants,stems	tree,people,flowers,window,shops
This paper (TagProp): whole+roi+bg	<u>tree,rocks,fox,den,moose</u>	<u>sky,water,boats,people,valley</u>	<u>people,buildings,light,restaurant,night</u>	<u>leaf,grass,flowers,close-up,plants</u>	<u>people,buildings,flowers,light,shops</u>
Ground Truth	grass,fox,den,arctic	water,boats,waves	people,restaurant	leaf,close-up,plants	light,shops

Figure 3.11: Corel5K dataset annotation examples in comparison with the baseline approaches

*The Figure is taken from Figure 10 of the author's paper [J1]*

Our work (TagProp): whole + roi + bg Vs.	Corel5K				
	2x <	Worse	Draw	Better	> 2x
TagProp*	135	135	23	102	26
Grid (TagProp): whole+roi+bg	19	99	32	129	29
Our work (TagProp): whole	9	84	36	140	22
Our work (TagProp): roi	22	67	11	182	64
Our work (TagProp): bg	20	79	27	154	41
Our work (TagProp): whole+roi	7	98	42	120	10
Our work (TagProp): whole+bg	13	91	32	137	15
Our work (TagProp): roi+bg	11	81	32	147	22

Table 3.15: Number of worse, draw and better results in keyword-wise MAP of our whole+roi+bg versus other approaches in the Corel5K datasets

Our work (TagProp): whole + roi + bg Vs.	ESP Game				
	2x <	Worse	Draw	Better	> 2x
TagProp*	79	79	0	189	16
Our work (TagProp): whole	2	123	0	145	2
Our work (TagProp): roi	3	51	0	217	16
Our work (TagProp): bg	3	74	0	194	9
Our work (TagProp): whole+roi	1	124	0	144	3
Our work (TagProp): whole+bg	2	125	0	143	3
Our work (TagProp): roi+bg	1	103	0	165	3

Table 3.16: Number of worse, draw and better results in keyword-wise MAP of our whole+roi+bg versus other approaches in the ESP Game datasets

Our work (TagProp): whole + roi + bg Vs.	Corel5K				
	2x <	Worse	Draw	Better	> 2x
TagProp*	157	157	62	281	98
Grid (TagProp): whole+roi+bg	10	200	90	210	10
Our work (TagProp): whole	7	179	97	224	9
Our work (TagProp): roi	14	138	55	307	90
Our work (TagProp): bg	11	153	78	154	41
Our work (TagProp): whole+roi	3	194	116	190	5
Our work (TagProp): whole+bg	9	184	109	207	8
Our work (TagProp): roi+bg	4	164	101	235	16

Table 3.17: Number of worse, draw and better results in image-wise MAP of our whole+roi+bg versus other approaches in the Corel5K datasets

Our work (TagProp): whole + roi + bg Vs.	ESP Game				
	2x <	Worse	Draw	Better	> 2x
TagProp*	850	850	79	1152	159
Our work (TagProp): whole	29	930	126	1025	48
Our work (TagProp): roi	62	792	90	1199	170
Our work (TagProp): bg	26	811	90	1180	112
Our work (TagProp): whole+roi	13	932	173	976	15
Our work (TagProp): whole+bg	21	924	143	1014	31
Our work (TagProp): roi+bg	19	863	149	1069	37

Table 3.18: Number of worse, draw and better results in image-wise MAP of our whole+roi+bg versus other approaches in the ESP Game datasets

### 3.4.3 Discussion

We have shown that our features give a higher performance in all of the metrics except the recall rate of the ESP Game dataset with the JEC method. The reason could be because JEC does not exploit all the different feature distances, but rather uses them as one feature distance by combining them all. Furthermore, for most cases, we could statistically prove the significance of our results over those of the baseline approaches with a sign-test by requiring  $p - value < 0.05$ . We have also given examples of our approach in action in terms of retrieval and annotation tasks. In all these examples and obtained results, our approach helps not only to obtain the most relevant images and annotations, but it also helps to promote diversity among result sets in both settings. This is important because diversity is one of the most important factors in image search and has become even more important in this era of image explosion. This outcome is due to the use of both salient and the background regions in addition to the whole image which maximizes the recall. It is also noted that features from salient regions and background contribute to the performance when using them with features from the whole image. However, the combination of all these features gives the best performance.

Two main problems that we could observe which reduce the performance of our features and method: (i) the complexity of the image and (ii) the poorly labeled

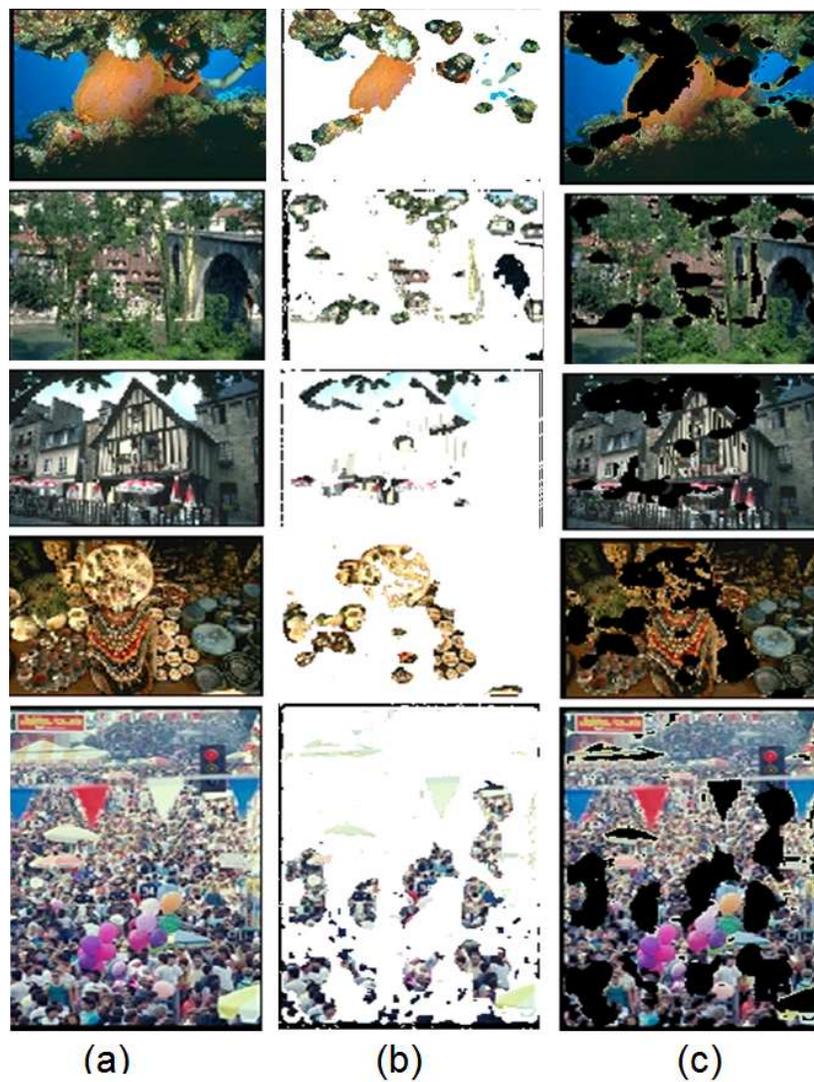


Figure 3.12: Example showing some complex images that result in failure in salient regions and background extraction: (a) the original image, (b) the extracted salient regions and (c) the extracted background.

*The Figure is taken from Figure 12 of the author's paper [J1]*

dataset. There are cases where the visual content of the image is rather complex which makes the resulting salient regions less accurate. In turn, this influences our extracted features. Fig. 3.12 shows some unsuccessful cases with complex images of the Corel5K dataset. We are considering extending the mechanism to effectively adapt the size of our saliency map. The drawback of the methods that we used is that they are completely based on the bottom up approach, i.e. no human data is used. We would like to further explore the complementary usage of the method in [64] where the authors extract salient regions using data learnt from human observers. For the second problem, we believe that having a rather good training dataset would lead to even better results with our feature set and approach. It could be observed that many times the approach gives the good result sets in terms of nearest neighbors but they are not annotated or poorly annotated with noise in the ground truth. One solution would be to do some pre-processing in the training dataset to reduce noise and include more annotation.

### 3.5 Conclusion

As the number of images keeps growing at an exponential rate, image annotation is a very important problem to solve. With the recent advancement of research in salient region extraction, we propose to extract features from the whole image as well as the regions of interest and the background. Methods designed to automatically extract the salient regions and the background and afterward the features from the respective areas are presented. A diverse range of features from the color, the texture, the scene to advanced local invariant features have been extracted. We report extensive experiments to confirm our approach as well as to show the strength of our features. It is shown that this new paradigm is very promising especially for the web image contents with weakly labeled training data.

## **Applications**

Our method can be used in many visual related applications. One immediate application is video annotation where we can use our approach for the key-frame images of each video. Other potential applications include surveillance systems, robot vision and medical image analysis. It can also be applied in the image aesthetics and image emotion inference fields through image feature analysis. However, it is not limited to these applications. Others that would make use of feature extraction, feature analysis, specific region detection or recognition, foreground and background detection can employ the method presented in this chapter.

## **Future Work**

We plan to further study on the selection of other advanced features to complement our existing ones. The self-similarity descriptor [111] can be one of them. Distance metrics are also very important in order to fully exploit the strength of each feature. Thus, we would like to investigate on other feature distance metrics. Moreover, we also intend to explore feature adaptation mechanism, as well as to enhance the salient region extraction method in order to be able to deal with complex images.



# Chapter 4

## On Automatic Image Annotation: The Personal Case

### 4.1 Introduction

#### 4.1.1 Background and Motivation

Nowaday, consumers capture and store thousands of their digital photographs on their personal computers. They can also speedily share them with their friends over the Internet. However, with the rapid growth of personal digital photos, the complexity and difficulty in archiving, searching, browsing and sharing photographs have also proportionately increased. The current photo management systems are still quite limited and unnatural. Hence, users cannot fully enjoy their photos because the real value of the photos depends largely on how they can effectively and efficiently access, manage, and share them.

These above mentioned problems are due to the lack of rich metadata associated with photos. Annotation is one of the key solutions to enable better access to digital photographs. In other words, users need to provide contextual metadata (meaningful descriptions) to each of their photograph files. This would allow them

to find their photos by searching using more abstract information instead of the file or directory names. However, this annotation process is tedious and time-consuming for users. Factor in the need to annotate hundreds or thousands of photos, and the task quickly becomes unrealistic for the average user to conduct or keep up with. Research shows that although people would like their photo albums to be organized, many do not label more than only a few, or they do not invest the effort to label their photos at all [97]. Therefore, most photos are poorly annotated or just retain the numerical file names that the camera defaults to.

Various research efforts on how to annotate images have been going on actively in the last decade. On one hand, there are techniques to extract relevant metadata directly from image content which include color/texture extraction, object identification, face detection/recognition, content-based categorization, etc. In 2000, Smeulders et al. published a comprehensive survey of these techniques [116]. However, these content-based technologies hold limited value as they are often inaccurate and too vague to accurately represent the interpretation of each individual. Other methods involve designing a better graphical annotation interface in order to allow users to easily input contextual metadata manually. In addition to this, there are approaches that depend on users' collaboration. One of them is an ESP game-like approach that is gaining popularity by using the power of anonymous volunteers to help manually label the photos over the web [133]. This concept is also adopted by Google Image Labeler [45]. However, this kind of approach has two drawbacks. First, it requires consistent participation from users, consuming both their time and energy. Second, it will never work for annotating personal photos, which often require private knowledge and contextual information of the owner's ambient environment and application of his or her personal interpretation of the environment and moment. Other methods try to use both content and context information such as that of Tuffield et al. [129]. However, the work is still very primitive and the authors only limit to a few kinds of contextual information. Datta et al., recently,

produced a detailed survey paper of the progress report in the field from the year 2000 [26]. We will also elaborate more about the closely-related techniques to ours in the Related Work section.

### 4.1.2 Problem Formulation and General Idea

In our study, we look at the problem by asking the following question: *how can we generate semantic metadata for photos without requiring the owner to manually input the data?*

We answer this question by proposing to *use the maximum amount of contextual information about the photos that are available from and to the users*. Information from the photo owners, such as their emails, schedules, web browsing histories, files, etc., and information available to the owners, such as news, encyclopedia, etc., is the focus of this study. We introduce a practical implementation paradigm to leverage the above mentioned information which serves as personalized and contextual metadata to suggest back as the semantic metadata for the photos. We do this by assuming that the exact location information is available for every captured photo based on the current trend in geo-photography. We use this location data in addition to timestamps data of the captured photo as “information filters” for relevant contextual information of that photo. By applying information extraction and retrieval techniques to the filtered contextual information, our system can suggest accurate semantic keywords to each photograph. Moreover, we propose to use named entities, such as the names of people and organizations, to represent the exact semantic meaning of the photos in addition to the high frequency terms.

We have designed and implemented a prototype of our proposed system. We have also performed the experiments to verify the effectiveness and accuracy of the system. Results show that users are able to annotate their photos significantly faster

using our proposed system. We have also obtained an encouraging rate of accuracy.

## 4.2 Related Works

This section provides the background for the research described in this chapter and the context within which the work is situated. The image annotation techniques that have been investigated thus far can be categorized into three major types: manual, semi-automatic and automatic.

### 4.2.1 Manual Annotation (with UI enhancement)

There are many image management tools (both commercial and research prototypes) that offer the manual annotation capability. What follows are descriptions of several selective systems that represent the essential functionalities of the currently available tools.

Adobe Photoshop Album [11] allows users to define customized keyword tags for people, places or events and drag them onto photos so that they can be searched later using these tags. Tags can be separated into categories and sub-categories for convenient annotation and dynamic organization of photos. Although the annotation system is limited, it is still more effective than the folder-based approach. On the other hand, the annotation process in Google's Picasa [103] and ACDSsee [9] is still very time-consuming. Users are required to input keywords manually. They only improve the look-and-feel of the GUI of their tools.

One research prototype, PhotoFinder [65] features a drag-and-drop technique that enables users to drag terms (such as person's name) and place them on an image. PhotoFinder associates annotation with coordinates in each photo that later allows for search queries such as "Nick next to Tommy". On the other end of the spectrum is Caliph, which is part of the Caliph & Emir project [84]. Caliph is a

semantic annotation tool designed to help users define semantic objects to be associated with their photos that can later be reused. Caliph can also perform efficient retrieval via the Emir tool.

Collectively, the two obvious burdens of these techniques are that they are time intensive and tedious. In addition, users need to pay great attention during the annotation process in order for it to be effective.

#### **4.2.2 Semi-automatic Annotation (including collaborative annotation)**

Semi-automatic techniques suggest some pieces of information to users in regards to arranging and clustering photos rather than having the users input everything themselves.

Wenyin et al. proposed the MiAlbum [137] system, which uses feedback to progressively improve annotation in the search process. When a user submits a keyword query, three kinds of results will be generated on the screen: images relevant to the keyword, images that are visually similar to the relevant images and randomly selected images. A user judges the resulting images using a thumb-up icon. If the user is satisfied, the search keyword will be attributed to that image. The overall quality of the annotations is improved with the extended use of such a system.

The MMM framework [107] allows camera phone users to annotate their photo immediately at the location where they captured the image. This system first displays time and location information and then generates other information from pre-populated lists that others have previously populated with their data through collaborative sharing of tags. A similar strategy is also employed in online photo management systems such as Yahoo! ZoneTag [145].

Naaman et al. [96] has presented a system that suggests identities inside a photo using the co-occurrence and re-occurrence patterns. The work assumes that accurate location information is available to the photo in addition to date/time information. The method relies on the identities that have previously been associated to the other photos in the collection.

Photocopain, created by Tuffield et al. [129], aims to take advantage of available information such as EXIF metadata, calendar data, community tags and GPS. However, there is more focus on content analysis than context, and only a few kinds of contextual information are taken into consideration. The work is still in an early stage.

There are many other interesting approaches in this category, but we focus here on those that are closely-related to our work. Other methods, such as the SmartAlbum system, assume that each photo comes with voice annotation, and the work analyzes speech signal using speech recognition methods [124]. Girgensohn et al. [43] use face recognition techniques to facilitate the annotation of people appearing inside the photos. The major concerns with these types of systems stem from the fact that most of them only target one aspect of the semantic information, thus creating a lack of scalability for practical implementation.

### **4.2.3 Automatic Annotation**

Many of today's image search engines, such as Google Image Search [44], use surrounding text as a way to generate metadata for the vast number of images on the web. In the web image domain there are an increasing number of investigative systems. One such recent system, AnnoSearch [136], does the annotation first by using an accurate initiative keyword obtained from file names or surrounding text in order to search for other web images. Then, the resultant images are compared and

clustered visually and semantically. Li and Wang have proposed an Automatic Linguistic Indexing of Pictures or Real-time (ALIPR) [77]. This system is an automatic image annotation system that learns from the training dataset and users and is able to achieve significant results in both time and accuracy. Zhou et al. have created an interactive approach for image annotation by incorporating keyword correlations and region matching [144]. However, the results could still be improved upon as well.

Aria [78] enables users to annotate their photos while composing emails. It automatically adds annotation to relevant photos in a collection as the email is being written. This is done using the information from a common sense database [115].

In conclusion, the systems currently in use are a part of a positive trend, and tools of this kind which do not require user intervention are very much needed. However, these systems are still in need of work, as the annotations are most often vague and inaccurate.

## Summary

Despite the diversity of efforts made in the previously mentioned work, the main challenge in generating annotation that represents an individual's interpretation of their photos remains unsolved. So goes the saying, "A picture is worth a thousand words". In an ideal world where a perfect object/face recognition algorithm exists, a computer would still not be able to mimic an individual's perception about a photo without considering its context. The Photocopain system nearly succeeds in integrating contextual information with annotation. However, a perfect system will need to go one level deeper and pay close attention to integrating all available information *to* and *from* users in their ambient environment. The systems presented here are trying to achieve this goal.

## 4.3 Proposed Approach: Leveraging Context to Bridge Semantic Gap

### 4.3.1 Nature of Personal Digital Photographs

An image or photograph can mean different things to different people. An image itself has no intrinsic meaning. Instead, meaning is bestowed upon the image by the viewer. Personal digital photographs have very different characteristics when compared with other types of images, such as those found in museums or web image collections. Usually a user's personal digital photos reflect their daily activities. The information from one's daily life is the ideal resource to be used to extract the semantic information needed to describe photos taken on a particular day or within a short interval.

### 4.3.2 Gathering Contextual Information

Many of us use computers both at home and at work. We use them to prepare or consult our schedules; read or write emails; surf the Internet; and get or share information with family, friends and colleagues via various Internet services such as chats, forums and blogs.

In a typical scenario, suppose that we are going for a trip, we might have planned this ahead in our schedules. Before leaving, we book a hotel room online, find the nearest public transportation and look for general information about the place we are to visit, such as weather, culture, main attractions and related news. We might use encyclopedia and tourism websites, online news and other sources. We might also email or chat with our friends and family about our upcoming trip. On the spot, we take lots of photos while we enjoy the trip. Upon returning, we share the photos as well as our impressions about the places with our friends and family via the Internet services mentioned earlier. This is often very useful information, as it

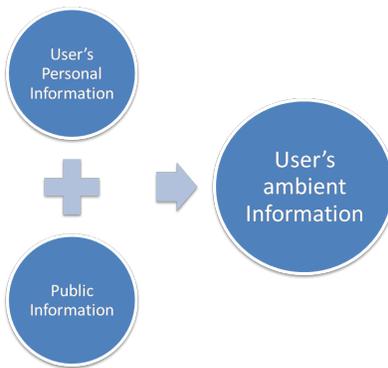


Figure 4.1: Contextual ambient information gathering

comes from a user’s direct personal interpretation of the photos (via their schedules, emails, chats, etc.) as well as from the other information they are processing from their environment (such as Wikipedia, tourism websites and online news websites, to name a few). These sources of information are important because what occurs in the ambient environment will add both direct and indirect effect to a user’s episodic memory. When looking for photos later; users are very likely to use the same keywords that they use in personal documents and in describing experiences in their ambient environment. We categorize these sources of information into two types:

1. **Personal information** refers to available contextual information from users such as schedules, notes, emails, chats, web browsing histories and all other documents residing in their computer or computers. These types of information link to users directly and personally.
2. **Public information** refers to contextual information that users consume freely or very cheaply such as local news, world news, encyclopedia information, tourism information, and other information from public repositories that are available online. These types of information link to users directly or indirectly.

Figure 4.1 depicts our concept.

### 4.3.3 Using Time + Location as Photograph Filters

As mentioned earlier, the *personal* and *public information* is readily or cheaply available, which provides for some huge advantages. However, a method is needed that allows us to distinguish which subset of the acquired information best represents the context of a captured photo. To do this, we consider the time and location of each photo as the key filters, because this information serves as the basic contextual metadata of the photo.

All digital cameras now provide time information. A timestamps indicating exactly when the photo was captured is embedded in each photo file itself. In addition, most camera phones can infer a rough location from GPS or Cell ID information. It is likely that all new cameras will eventually be equipped with location capturing systems. Additionally, most digital photographs support location data in addition to time information. This data can be stored in the form of a coordinate set (longitude and latitude) in the EXIF header [32] of every photograph<sup>1</sup>. There are documented trends as far as providing free location information database to the general public. For instance, Geonames [41] provides free geo-data such as geographical names and postal codes to the public, and its database contains over 8 million entries of geographical names within 2.2 million are cities and villages. Geonames's website boasts many features, including conversion from GPS coordinate set to nearby location. Consequently, there is no problem as far as translating a GPS coordinate set into an exact location name. As a result of services such as these, we will be able to obtain two key filters, namely timestamps and location, without much effort in the near future.

Based on the above facts and hypothesis, knowing the exact time and location where a photo was taken can be used to extract the subset of personal and public

---

<sup>1</sup>It is noted that EXIF is supported by only JPEG and TIFF.

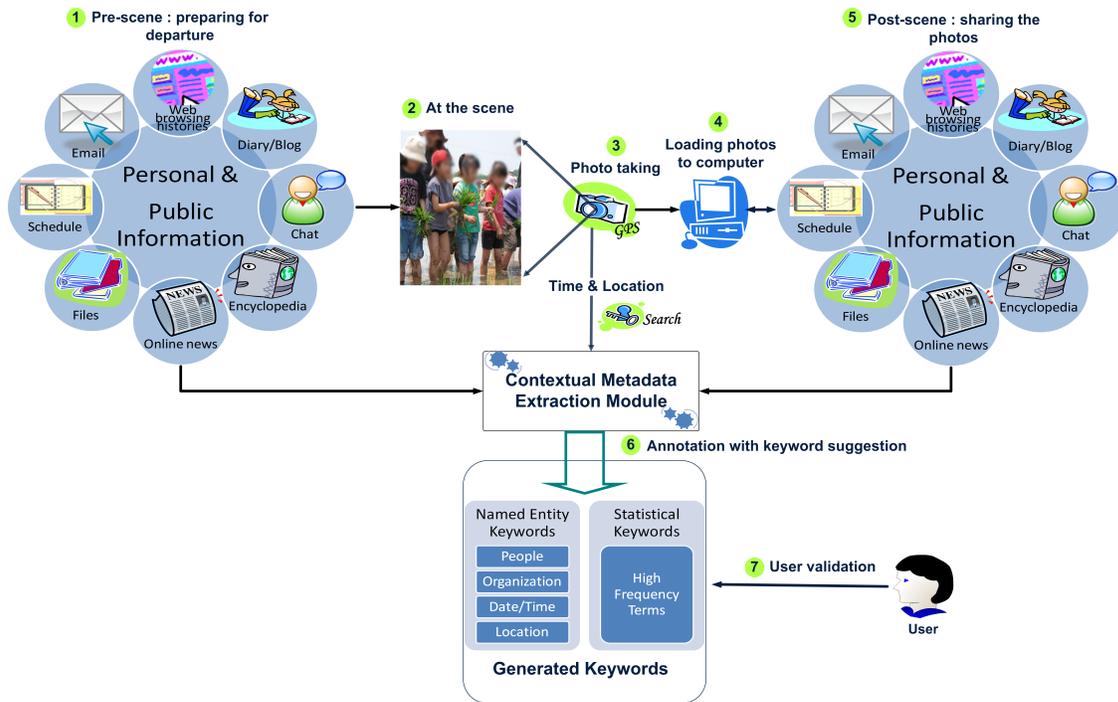


Figure 4.2: Overall View of the Concept

The Figure is taken from Figure 1 of the author's paper [J2]

information from a user's pre-scene (before going) and post-scene (after going) that strongly relates to a photo or group of photographs. By applying some Natural Language Processing techniques to this obtained information, we will be able to extract important representative keywords and suggest them to users for their validation.

#### 4.3.4 Extracted Keywords

We identify two classes of keywords to be extracted:

1. **Named Entity Keywords** refer to strong and exact proper noun identifications found in the relevant files. To generate this type of keywords, we employ computational linguistic techniques to intelligently parse documents and discover Named Entity (NE) information. In our case, we would like to get the important episodic memory information such as *dates*, *names of people*,

*location names and organization names.*

2. **Statistical Keywords** refer to terms that appear very frequently in the relevant files and that can be used to represent these files.

Fig. 4.2 illustrates our concept.

## 4.4 System Design and Implementation

We have designed and implemented a prototype of our system. The overall architecture of our system is depicted in Fig. 4.3. The following is the step-by-step explanation of the annotation process with our semi-automatic annotation system:

1. Users begin by choosing the photo that they would like to annotate. It is assumed that these photographs are embedded with Date/Time and Location information. In our case, the file name of each photo contains location name.
2. The extracted Date/Time and Location are used as key filters to search for related sources from user's computers including their personal and public information. Google Desktop Search (GDS) returns to us the relevant files from its index.
3. Relevant files to the photo with respect to Time and Location are sent to the Named Entity Extraction Module. In return, NEs from the relevant files with respect to their categories namely, Date, Location, People's name, Organization will be output. In addition, those output NEs are ranked by their frequencies of occurrence.
4. In the same manner as the previous step, all the relevant files related to the photo are sent to Statistical Keyword Extraction Module. This module processes the term ranking and outputs the top keywords ranked by their frequency of occurrence in the document sources.

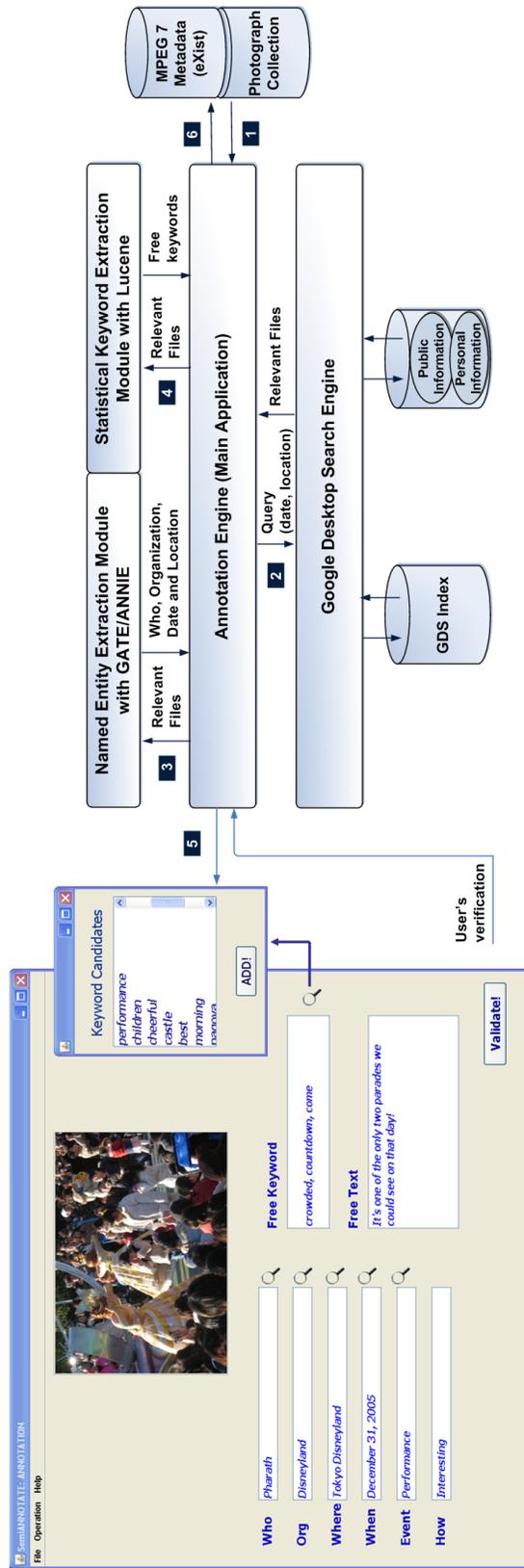


Figure 4.3: System architecture of the implemented prototype

*The Figure is taken from Figure 2 of the author's paper [J2]*

5. In this step, metadata (NEs + Statistical Keywords) found in step 3 and 4 are presented to the users. Top suggested keywords of each category are shown in their respective fields of the interface. Users may consult more keywords by clicking on the *magnifying icon* of each field. Finally, users validate the metadata candidates (They can always edit or augment the metadata if necessary).
6. All the metadata validated by users are converted to MPEG 7 MDS format and are sent to our eXist XML database.
7. All detailed processes are described as the following.

#### 4.4.1 Data Acquisition

Personal information of a user resides in their computers. Currently, there is a tremendous interest in Desktop search. Desktop search engine software can index and search files on a single computer or across multiple networked computers. The world's top software companies such as Google, Yahoo!, and Microsoft offer their proprietary versions of the Desktop Search Application. Lu et al. have a comprehensive analysis about the various kinds of desktop search software currently available and their performance metrics [81].

Google Desktop Search (GDS) [39] is among the most popular desktop search applications. GDS manages and indexes files found on personal computers. These files include email, schedule, web browsing history from Internet Explorer and Mozilla Firefox, office documents in the Open Document and Microsoft Office formats, memo, PDF, instant messenger transcripts from AOL, Google, MSN, Skype, and several multimedia file types. GDS includes plug-ins for different file formats that allow one to index and search through the contents of those local files. Google Desktop's email indexing feature is also integrated with Google's web-based email service called Gmail. GDS performs all tracking, cataloging and indexing entirely independently of the Windows caching of Internet pages. Therefore, should a user delete

their temporary Internet files, cache, and cookies, a record of the data is maintained by the GDS program. This means GDS caches all HTML Internet pages visited. Additionally, should a single web page have been visited repeatedly, the Google Desktop Search will store cached copies of all of these pages, giving exact information on what was presented to the browser on each visit. In addition, GDS is designed to index and retrieve user-created data only. Consequently, it does not index system-related files such as Microsoft Windows system files. Files stored within the default Windows directory, within the Recycle bin, or those that are invisible are not be indexed. They are excluded from indexing, increasing the efficiency of the program [130]. Another feature of GDS is called Search Across Computers. This feature enables us to search our files and viewed Web pages across all of our computers. For example, one can find files that he or she edited on the desktop from their laptop. To activate this feature, a Google Account is needed and the GDS program must be installed on each computer [110].

With these above mentioned qualifications, we decide to choose GDS as our data acquisition tool. This enables us to access all of the personal information residing on the user's computer. In our case, to make it simple, we also make public information available to GDS so that it can index this together with personal data. To do so, we download news and encyclopedia data from the Internet, and maintain them in the local directories on the user's personal computer. We consider the following online public repositories as the *public information* to be integrated:

1. News : MDN Mainichi Daily News [86], The Asahi Shimbun [14] (in English and in duration of two-year time)
2. Encyclopedia: English Wikipedia [138]

The news pages are downloaded via a tool called HTTrack [59]. The tool is configured to download only printer-friendly version of its HTML pages to minimize the

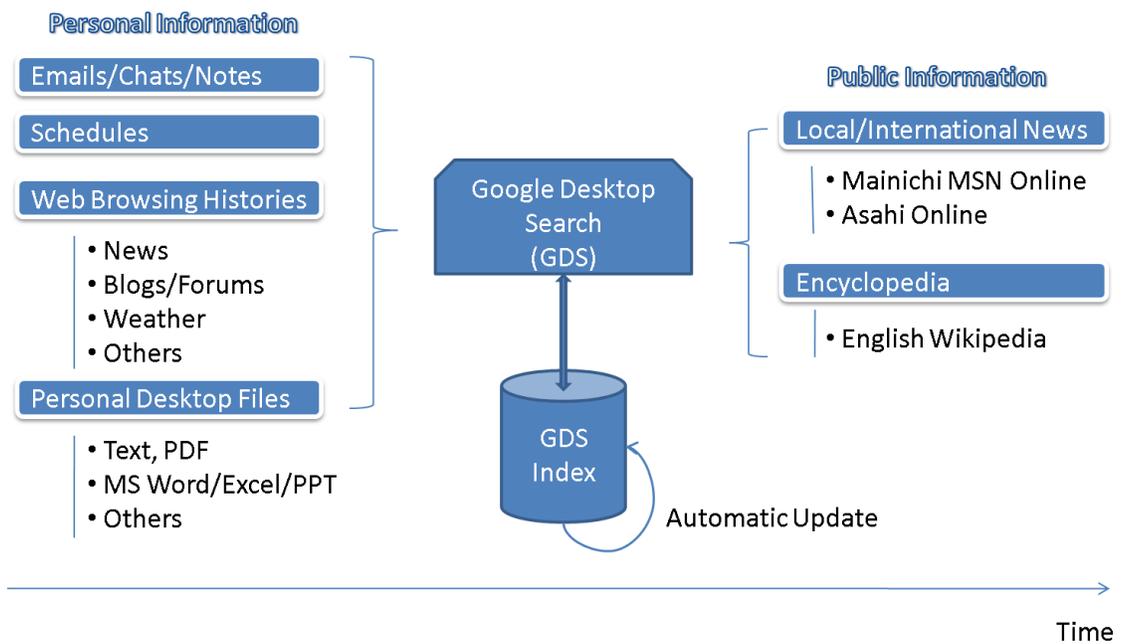


Figure 4.4: Data acquisition of personal and public information with Google Desktop Search

*The Figure is taken from Figure 3 of the author's paper [J2]*

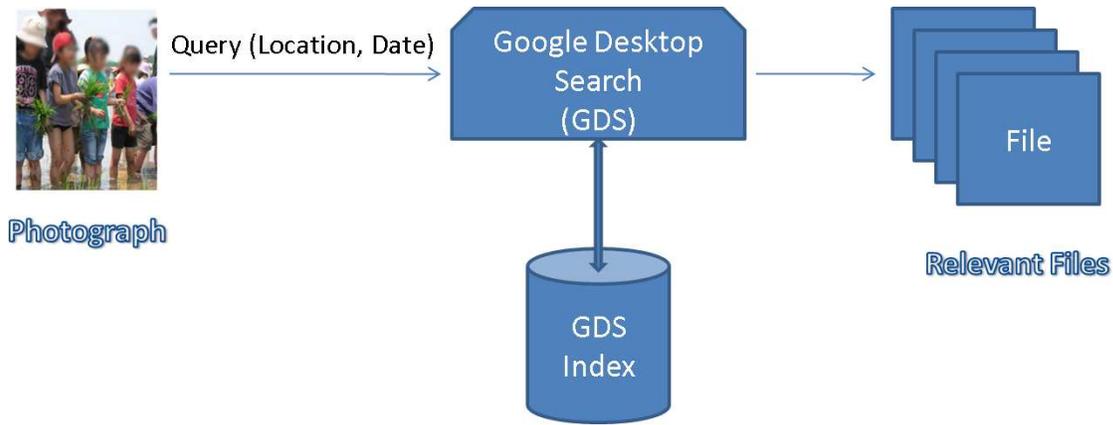


Figure 4.5: Process of generating relevant files to the photo with location and time as event filter

*The Figure is taken from Figure 4 of the author's paper [J2]*

tasks needed to clean up the unnecessary information in the page such as advertisements, pictures, flash media, etc. GDS is integrated into our system via its Java API, which is available from the SourceForge website [40]. Fig. 4.4 summarizes the process.

#### 4.4.2 Relevant Files Generation

Google Desktop Search also serves as our search tool for relevant indexed sources to date and location. This allows us to leverage Google's search technology. GDS is designed to narrow search space to areas that are more likely to contain documents stored by the user rather than files used to operate and maintain the computer. We define three patterns of queries to GDS to enable both exact and loose query in case the number of exact relevant sources are limited. We limit the maximum size of the result set to 100 in order to assure the quality of our metadata and the efficiency of the approach by maintaining both relevancy and computing performance. Fig. 4.5 shows the process in generating relevant files. Algorithm 4.1 is used to retrieve relevant contextual information for the photos from public and personal information resources.

---

**Algorithm 4.1** Generate relevant files

---

```
REQUIRE gds_index, date, location
ENSURE relevantFiles = generateRelevantFiles(gds_index, date, location)
1: resultSet1  $\leftarrow$  gds_index.query(date.getMonthYearDay(), location)
2: resultFiles  $\leftarrow$  resultSet1
3: IF relevantFiles.getSize() < 100
4: resultSet2  $\leftarrow$  gds_index.query(date.getMonthYear(), location)
5: relevantFiles  $\leftarrow$  relevantFiles.add(resultSet2)
6: IF relevantFiles.getSize() < 100
7: resultSet3  $\leftarrow$  gds_index.query(date.getYear(), location)
8: relevantFiles  $\leftarrow$  relevantFiles.add(resultSet3)
9: ENDIF
10: ENDIF
```

---

### 4.4.3 Keywords Generation

#### 4.4.3.1 Named Entity Generation

To get this type of keywords from relevant sources, information extraction techniques are needed. For this purpose, we integrate the General Architecture for Text Engineering (GATE) [22], a mature open source text engineering platform, into our system. GATE comes with A Nearly New Information Extraction (ANNIE) engine, a robust information extraction engine based on finite state algorithms. ANNIE depends on a number of language processing tools to do named entity extraction range from Unicode Tokenizer, Sentence Splitter, Part-of-Speech Tagger, Gazetteers, Semantic Tagger to Name Matcher and Pronominal Coreferencer. We introduce some linguistic resources specific to our situation such as company names, city names, people's names, etc. We also developed a NE sorting and ranking module associated with the GATE/ANNIE module. Top 20 NE keywords are generated for each category of keywords. Fig. 4.6 depicts the process of named entity keywords extraction. We describe each element as follows:

- The tokenizer splits the text into very simple tokens such as numbers, punctuation and words of different types.
- The gazetteer lists used are plain text files, with one entry per line. Each list

represents a set of names, such as names of cities, organizations, days of the week, etc.

- The sentence splitter is a cascade of finite-state transducers which segments the text into sentences. This module is required for the tagger. The splitter uses a gazetteer list of abbreviations to help distinguish sentence-marking full stops from other kinds.
- ANNIE's semantic tagger is based on the JAPE language. It contains rules which act on annotations assigned in earlier phases, in order to produce outputs of annotated entities.
- The name matcher module adds identity relations between named entities found by the semantic tagger, in order to perform coreference. It does not find new named entities as such, but it may assign a type to an unclassified proper name, using the type of a matching name.
- The pronominal coreference module performs anaphora resolution using the JAPE grammar formalism.
- Named Entity Sorter ranks and sorts the found NE according to their frequencies of appearance and their category.

#### 4.4.3.2 Statistical Keywords

Google Desktop Search is a closed technology of Google. We cannot fully configure and program it to analyze its index. Therefore, we also need a tool to index those related documents in order to perform other kinds of keyword extractions. Lucene [82] is a good tool to use to accomplish this. Lucene is the most famous open source information retrieval library. At the core of Lucene's logical architecture is the idea of a document containing fields of text. This flexibility allows Lucene's API to be independent of file formats. Text from PDFs, HTML, Microsoft Word documents

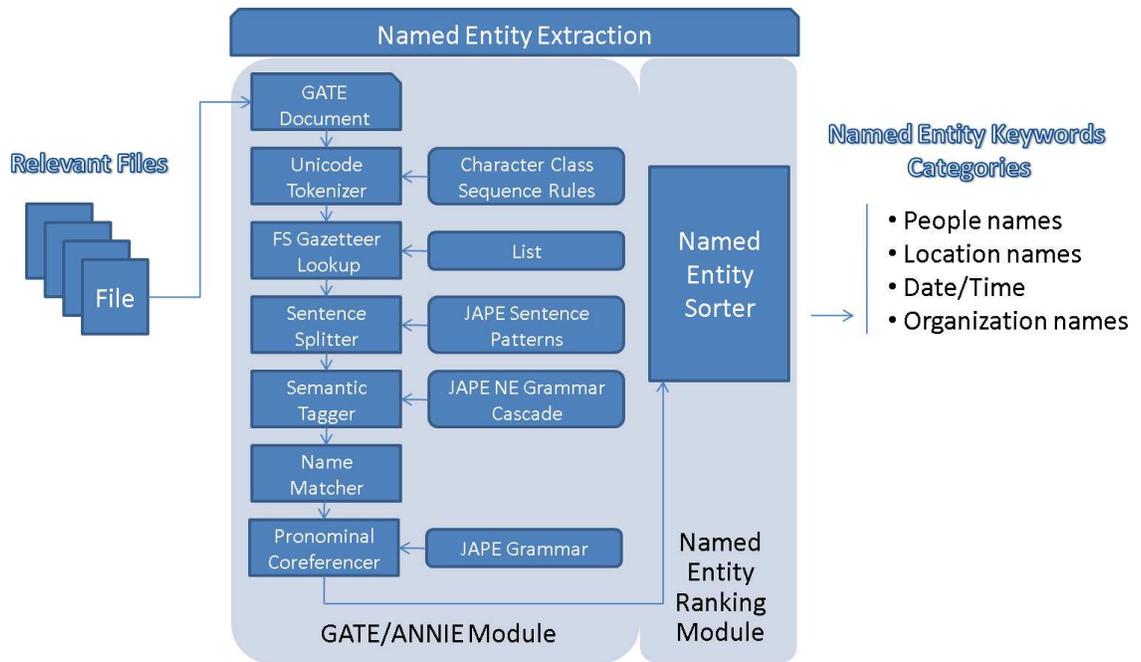


Figure 4.6: Named entity keyword extraction process

*The Figure is taken from Figure 5 of the author's paper [J2]*

and many others can all be indexed as long as their textual information can be extracted. In our case, we index all the relevant files in the different formats by the Lucene module that we developed using the Lucene's Java API. With this index, we calculate the statistics of each term to find the most frequent terms in the document collection that can be used as representative terms. Top 30 keywords are then generated for each photo. The following shows how we calculate the frequency of each term.

Let

- $TF(i, j)$  : the number of occurrences of term  $t(i)$  in document  $d(j)$
- $DL(j)$  : document length or the total of term occurrences in document  $d(j)$
- $n$  : the number of relevant sources

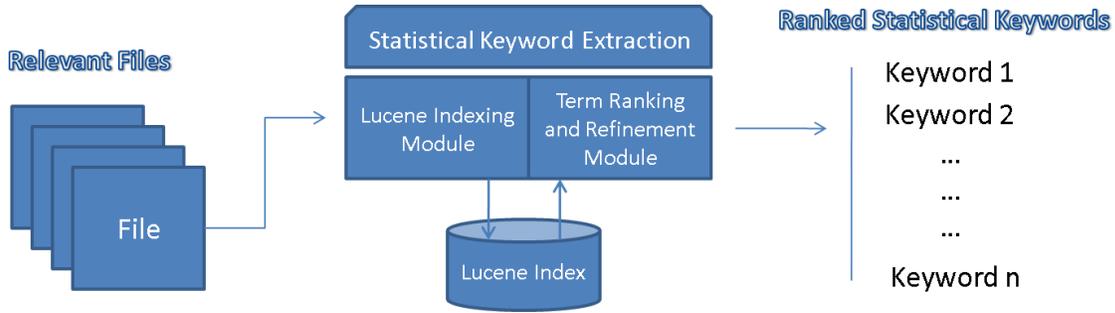


Figure 4.7: Statistical keyword extraction process

*The Figure is taken from Figure 6 of the author's paper [J2]*

A simple count is too crude because a term that occurs the same number of times in a short document is likely to be more valuable than in a long one. Therefore, we employ a simple adjustment based on the length of document. Hence, the term frequency is computed as the following:

$$TF_n(i) = \sum_{j=1}^n \frac{TF(i,j)}{DL(j)} \quad (4.1)$$

Fig. 4.7 illustrates the process in Statistical Keyword Extraction.

#### 4.4.4 Annotation GUI and Metadata Coverage

In our annotation GUI, we have correspondent text field for each of the categories of keywords. Below is the description of each one of them:

- *Who* refers to people's name
- *Org* refers to organization name
- *Where* refers to location name
- *When* refers to Date/Time
- *Free Keywords* refers to statistical keywords

Among generated NEs and statistical keywords, by default, the first top NE is inserted in the *Who* and *Org* fields while 3 statistical keywords are inserted in the *Free Keywords* field of the annotation interface. *When* and *Where* fields are also filled respectively with time and location of the photo. Users can always edit those default keywords if necessary. In addition to these automatically generated keywords, we also have other categories of keywords in our interface. They include:

- *Event* refers to reasons about the photos. We prepare some pre-set values for it with a list of events such as Birthday, Wedding, Meeting, Graduation, Festival, New Year, etc. that users can select from or add their own keywords.
- *How* refers to actions or emotions about the photos.
- *Free Text* refers to free text description about the photos.

We introduced these additional categories to improve the semantic integrity of our metadata for the retrieval task. Even though *Event* and *How* are not suggested by the current system, we believe that these keywords can be covered by the statistical keywords that we generate. Therefore, we can cover all of the related questions about photos including the *W5H1* (Who, What, Where, When, Why, and How) questions (*What* could also found in the statistical keywords). Please refer to Fig. 4.8 (B) for our annotation interface.

### **Metadata Format and Storage Database**

Contrary to Dublin core [30] which aims at simplicity, MPEG-7 [95] provides ways to give rich description for audio-visual media. Since our work focuses on semantic metadata about the photo, MPEG-7 element set is the best choice. In our case, we extended the *StructuredAnnotation* Basic Tool of MPEG-7 Multimedia Description Schemes (MDS) [106] to adapt and include all the categories of metadata extracted.

Since our MPEG-7 metadata is XML based, we also choose an XML native database to store the photo metadata in order to enhance the retrieval capabilities (search and browse). We choose eXist database for this purpose. eXist is an Open Source native XML database featuring efficient, index-based XQuery processing, automatic indexing, extensions for full-text search, XUpdate support and tight integration with existing XML development tools [33].

## 4.5 Empirical Evaluations

### 4.5.1 Validation goals

We investigate the performance of our system on two grounds:

1. The time difference between manual annotation and annotation by our proposed system using the built-in keyword suggestion features.
2. The accuracy of our proposed named entity keywords and statistical keywords by calculating their acceptable hit rates.

### 4.5.2 Participants and Data sets

#### 4.5.2.1 Subjects

We were able to recruit ten subjects for the experiments of our system. All subjects were computer science students at the graduate school of Global Information and Telecommunication Studies of Waseda University. They are all familiar with computers; they use and work with computers in their daily lives. Three of the subjects were women and seven were men.

#### 4.5.2.2 Personal Photographs

Each subject was asked to provide more than 30 personal photographs which had been taken over a period of six months. Photos are taken from events such as sight-

seeing, friend-gatherings, dinner parties, picnics, etc. Each subject provided photos for an average of 5 events. Each event had about 5 photos. We gathered 313 photographs in all.

Subjects were asked to install Google Desktop Search (GDS) and activate it each time they used their computers. Though GDS has its own cache index file system as described in section 3.1, the subjects were requested not to delete any of the files on their computers. This was required so that we can generate links to original files during the relevant files generation process. Subjects were also required to install our prototype system on their computers.

As mentioned in section 3.1, we manually downloaded the news from online repositories and Wikipedia. We then bundled this data into one single folder named *public\_information* and asked the subjects to save it on their computers. Google Desktop Search was then configured to include this folder into its index.

### 4.5.3 Experiment Process

First, in order to enable location information for each photo, we asked the subjects to label their own photos with the exact location name as the file name of the photo. To do this, we provide a drag-and-drop interface where subjects can easily input the location name on their photo(s).

The experiment is three part process. The first two parts are for time evaluation and the third one is to measure the accuracy. First, subjects are expected to annotate their own photos manually. Second, subjects were asked to annotate their photos using our proposed prototype system with keyword suggestion features. Between the two parts of the experiment, we leave a gap of 2 to 3 days so that subjects have time to forget their previously input keywords. This is done to avoid the

influence of a subject’s memory about the keywords of the photos that they have input into the system during the first step. Users were asked to input at least one keyword to the *Who* and *Org* fields. They have to input at least three keywords in the *Free Keyword* field. Lastly, subjects were requested to judge the accuracy of the automatically generated keywords for each photos that we saved into files before we performed the second step.

Please also note that we performed the experiment on the subject’s own computer, using their own contributed photos. Therefore, the timing varies depending on the configuration of their PCs. More details about the three parts of the experiment follow.

#### **4.5.3.1 Manual Annotation**

Users begin the experiment by manually annotating their photos with a blank interface. A blank interface is similar to the interface of our proposed system. It has all the fields for every category of keywords. However, the only difference is that there is no suggestion feature on this interface. Each text field represents a category of keywords accordingly. Thus, subjects have to manually input the annotation keywords to each text field. Annotation time is recorded for each photo. Fig. 4.8(A) shows our blank annotation interface.

#### **4.5.3.2 Annotation with Keyword Suggestion Features**

In this step, subjects annotate their photos with the help of our system. Top keywords of each field are shown in the respective text field. Subjects can consult other less ranked keywords by clicking on the magnifying icon and selecting from a drop-down list of suggested terms. At any time, subjects can modify the suggested keywords or add their own keywords if they find it necessary. Fig. 4.8(B) shows our annotation interface with the keyword suggestion features.

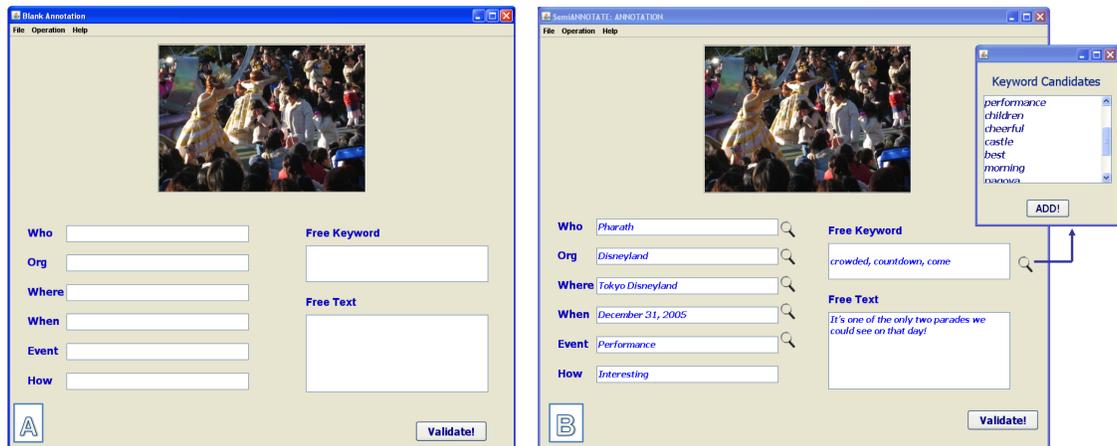


Figure 4.8: (A) Blank annotation interface; (B) Annotating interface with keywords suggestion feature

*The Figure is adapted from Figure 7 of the author's paper [J2]*

It is noted that in case no relevant file is found, the top NE keywords and statistical keywords found in the total index will be suggested. In the same way as in the previous task, we record the annotation time of each photo. It is also noted that at the beginning of this step, for each photo, we automatically generate the following: 30 free keywords, 5 person names, and 5 organization names. We then save these to a file for the last step of the experiment (keyword judging).

#### 4.5.3.3 Keywords Judging

In this step, we asked the subjects to work on the automatic keyword candidates of each field that we have generated. Subjects had to identify all the *acceptable keywords* of each field manually. Acceptable keywords refer to all the keywords that relate to the photo and are appropriate as keywords to describe or recall the photo.

## 4.5.4 Results and Discussion

### 4.5.4.1 Experimental Results and Analysis

#### (i) Accuracy

We evaluate the accuracy and the coverage of suggested keywords by using the following formulas:

- $$AcceptableHitRate(p, k) = \frac{\sum_{j=1}^p \sum_{i=1}^k H_j(i)}{p \times k} \quad (4.2)$$

- $$CoverageRate(p, k, n) = \frac{\sum_{j=1}^p \sum_{i=1}^k H_j(i)}{p \times n} \quad (4.3)$$

Where:

- p is the total number of photos
- k is the number of suggested keywords
- n is the number of acceptable keywords expected
- $H_j(i)$  is the hit function of keyword i to photo j

+  $H_j(i) = 0$  if the keyword is not acceptable

+  $H_j(i) = 1$  if the keyword is acceptable

Fig. 4.9(A) shows that the acceptable hit rates of proposed names of people and organization drop gradually from 31% (Who) and 27% (Org.) to 19% and 9% respectively when the number of names is suggested from 1 to 5. The first name suggested of both categories can hold about 30% of being acceptable. However, by integrating all the 5 suggested names together, Fig. 4.9(B) suggests that 99% of photos will have at least 1 acceptable person name and about 49% of photos will have at least 1 acceptable name of organization.

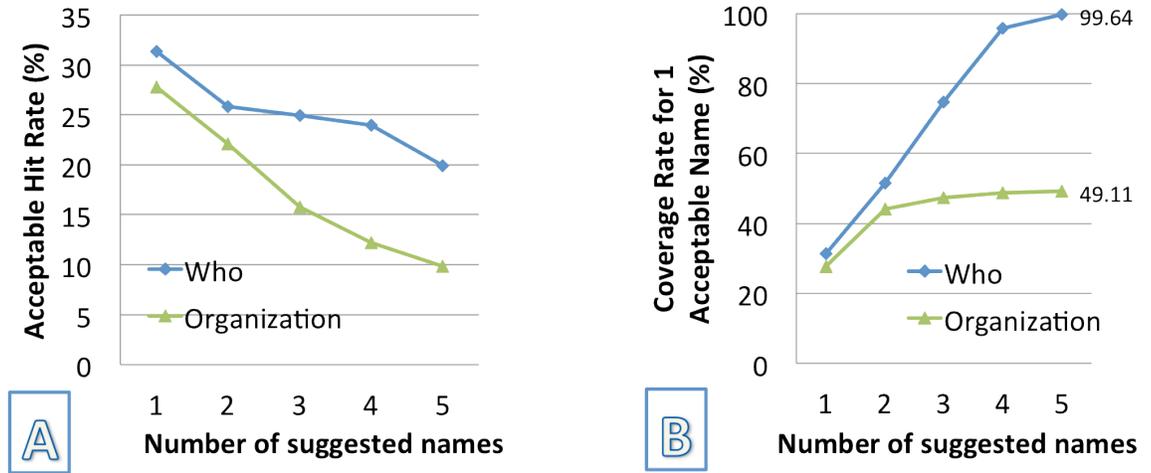


Figure 4.9: (A) Acceptable hit rate of Who (People’s name) and Org. (Organization names) keywords; (B) Coverage rate for 1 acceptable keyword of Who (People’s name) and Org. (Organization names)

*The Figure is taken from Figure 8 of the author’s paper [J2]*

Fig. 4.10(A) discusses the accuracy of automatically suggested statistical keywords. We can see that the hit rate reaches its peak level (60%) when we suggest 4 or 5 keywords. This means we shall get 3 acceptable keywords if we suggest 5 keywords to users. This is significant. However, Fig. 4.10(B) shows that, if we automatically suggest 30 keywords, the average number of acceptable keywords of the photos is 8. To further analyze, if we calculate the coverage rate for 8 acceptable keywords to the photos which is the percentage of photos that are correctly suggested by at least 8 acceptable keywords, we come up with the result in Fig. 4.10(C). It shows that to achieve 80%, 90% or 100% of coverage, we need to supply 15, 20 and 29 keywords respectively. These results are very encouraging.

## (ii) Time

We arrive at the following result. Fig. 4.11(A) shows that 9 out of 10 subjects gain benefit from this approach. Fig. 4.11(B) depicts that in average we gain an overall of 33% in annotation time over the traditional manual annotation. This is

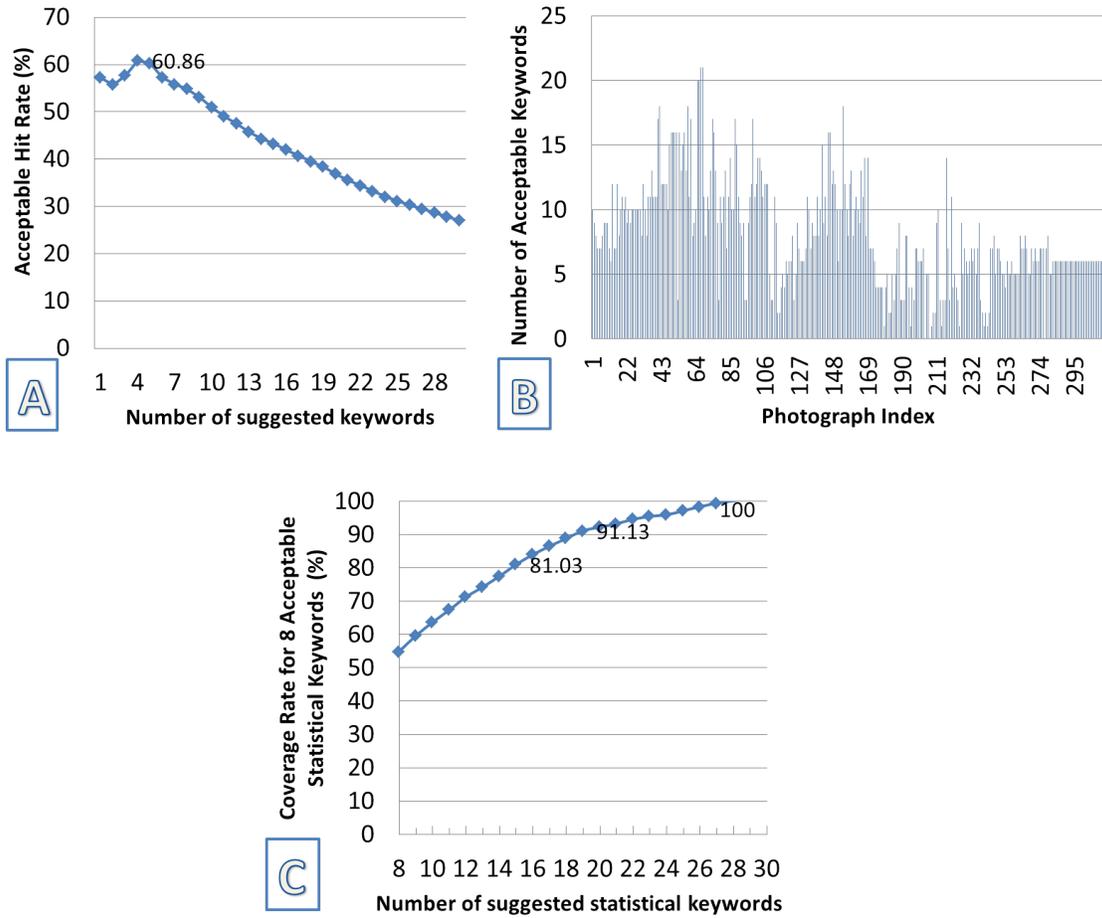


Figure 4.10: (A) Acceptable hit rate of statistical keywords; (B) Number of acceptable keywords of each photo; (C) Coverage rate for at least 8 acceptable keywords of statistical keywords

*The Figure is adapted from Figure 9 of the author's paper [J2]*

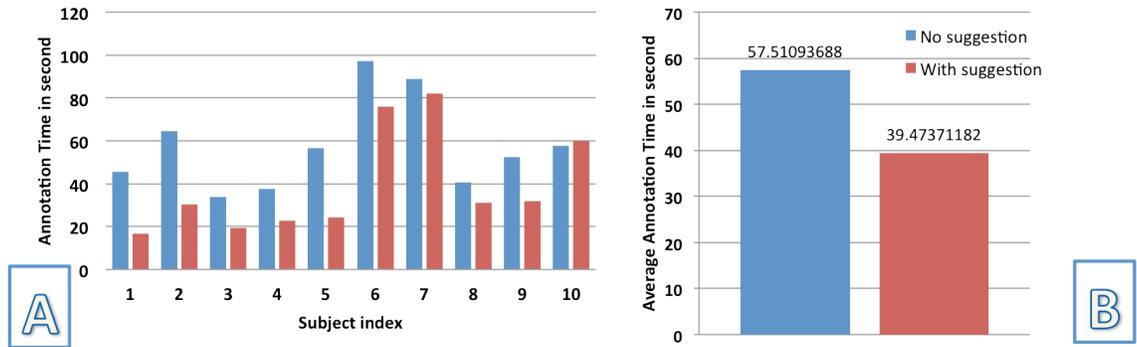


Figure 4.11: (A) Manual Annotation and Annotation with Keyword Suggestion Features of Each Subject; (B) Average Annotation Without and With Keyword Suggestion Features

*The Figure is taken from Figure 10 of the author's paper [J2]*

significant to the users.

## Analysis of Results

- In overall, our approach has allowed us to obtain good accuracy rate and time gain, despite a large diversity of photos and the relative subjectivity of our subjects. However, we should not neglect these influences. For instance, in Fig. 4.11(A), our system cannot overcome the problems of subject number 10. This is due to the fact that the majority of his photos are scenery from trips to different places and include no individual or organization names. In this case, the subject had to take time to edit the incorrectly suggested NE keywords or blank fields (when there is no keywords found by the automated system). In addition, he had to think of new keywords to attribute to his photos manually. We have noted that the type, size and numbers of files generated by a user most often link to that user's habits. This ultimately influenced the results of this study. We found that the average size of contributed data is always less than 100KB. Relevant files bigger than this size generally produce more noise. Furthermore, it is the personal information that contributes most to the acceptable keywords. Public information contributes only in the case that

the event is a breaking news event or happened in a popular place or time (such as New Year's, Christmas, at the Tokyo Dome). Events such as a simple dinner gathering do not create the same impact. Therefore, we shall establish a threshold in order to adjust to these variants.

- Obviously, there is also a strong correlation between the accuracy rate and the annotation time. However, we recognize that designing a better interface can save more time. In our case, subjects have to first click on the magnifying icon then click to select the other keywords from the list of keywords, and this process takes time. It would be more effective to show users the list of keywords in the interface directly so that they can drag and drop into the text field of each respective keyword category. In addition, by default our prototype system automatically inputs the top keywords into the text field of each category while some of the keywords might not be the acceptable ones. This would take users' time to edit and/or remove. Therefore, it would be better to directly show users the list of the keywords in the interface where users can drag and drop in the text field of the respective keyword category. However, not all the keyword candidates should be shown in the first place. For instance, from the above results, we found that if we suggested 5 names to the Who and Org. fields, we will get one acceptable name with the coverage rate of 99% and 49% respectively. And, for the statistic keywords, if we suggest 5 keywords we could get 3 acceptable keywords. We also found that when we suggest 29 keywords we will have 8 acceptable keywords with the coverage rate of 100%. However it is not practical to show all of these keywords. In this case, it is best to show the top 5 suggested keywords. To display other keywords, users just move the mouse pointer to the right or to the left at the end of the suggested keywords zone and other less ranked and high ranked keywords would appear respectively. Fig. 4.12 shows our proposed interface for the annotation based on our results.

- The information extraction part also takes a great amount of time as it involves lots of natural language phases. Better time gain could be achieved if we were able to perform this task offline.

#### 4.5.4.2 Discussion

There are a number of issues that the current prototype system does not focus on and they are worth addressing.

- We do not concentrate on distinguishing between photos that are taken during sub-events which occur within the same time and location, even if they are visually different. Therefore, in our case, for different photos taken on the same date and place, even they are visually different, the same relevant files will be generated. Thus, the same candidate keywords will be suggested. However, since we generate a lot of keywords, users can select among the proposed keywords to suit each of the photo in the sub-event accordingly. We believe that this is a powerful solution and will make it easier for users to distinguish and recall the events that happen on the same date with automated keywords. Additionally, there are already a number of research efforts in these problem areas such as Naaman et al. that propose algorithms to discover sub-events (like a birthday party). Furthermore, using observation and conversation with subjects has allowed us to learn that often subjects do not know which keywords they will eventually attribute to photos. Our system helps users with this task by not only suggesting keywords to associate with photo but also helping them to recall other relevant keywords. In parallel, this can also be regarded as a drawback because users tend to pick keywords from our suggested terms instead of generating the best new keywords for a given photo.
- Privacy is also a concern. A Google Desktop Search, for example, merely indexes all the files that it has access to. However, should a user with administrative rights install and run GDS within a multi-user environment, the

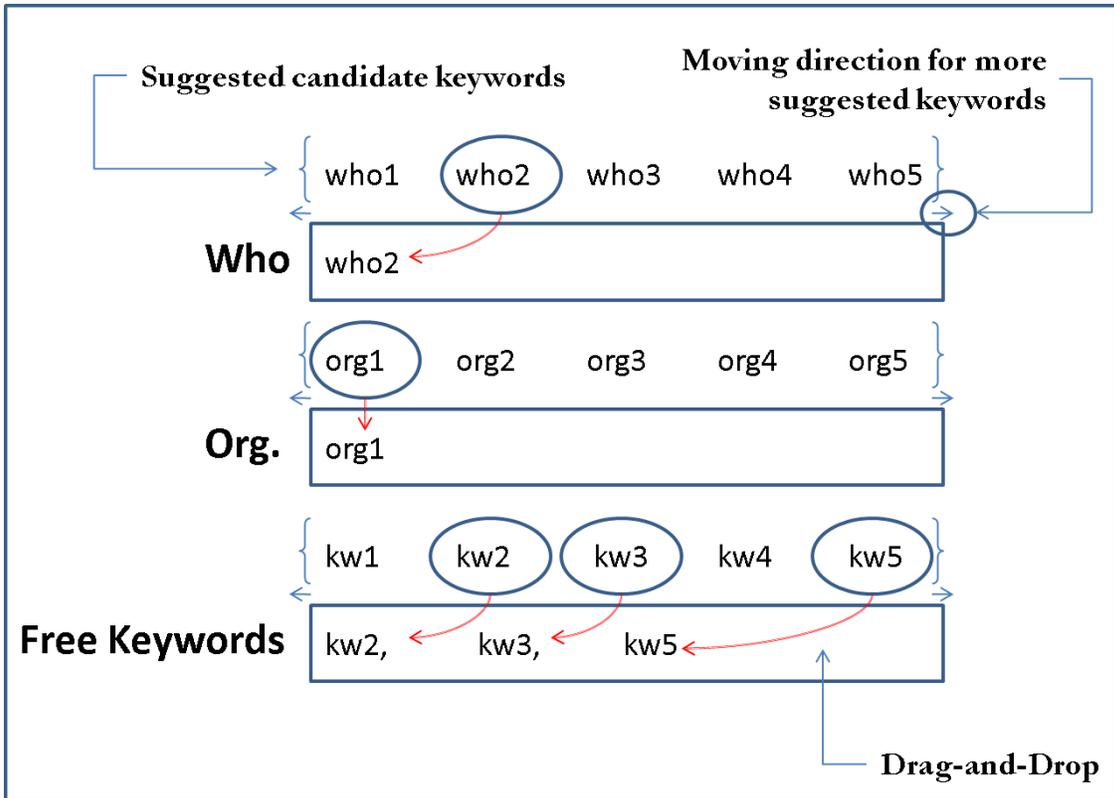


Figure 4.12: Proposed conceptual annotation interface layout for future implementation

*The Figure is taken from Figure 11 of the author's paper [J2]*

program indexes and searches all files regardless of their owner. We aim to address this problem in our future work.

- To build a faster prototype, we need to rely partially on a number of open-source APIs, and we have tried to select the best ones as our performance depends on them.

## 4.6 Other Features

Besides the annotation engine, we have also built the searching and browsing engines.

### 4.6.1 Searching

We provide five kinds of search namely, by people's name, date, location, keyword and full-text. We perform query of each category against our eXist XML native database by using XQuery and XPath. By default, a full-text search is performed to match the input keywords against the entire metadata. Fig. 4.13 shows our proposed searching interface.

### 4.6.2 Browsing

We have also built an experimental browsing system based on the episodic metadata that we get from our annotation engine. We believe that we are offering a flexible browsing interface that is different from the conventional ones.

In our case, we divide the browsing categories into four: Time, Location, People's name and Keyword. Users combine the metadata of these different categories to refine the photo sets until they reach the photo that they would like to see. They can go deeper or return backwards. With our interface, navigation becomes much easier for users. The interface gives hints at every stage of the browsing process by showing the possible metadata candidates of each category. Thus, users have



Figure 4.13: Search engine

*The Figure is taken from Figure 12 of the author's paper [J2]*



Figure 4.14: Browse engine

*The Figure is taken from Figure 13 of the author's paper [J2]*

an easier overall browsing experience. Fig. 4.14 depicts our proposed browsing interface.

## 4.7 Conclusion

A computerized system that accurately suggests annotations or keywords to its users is extremely useful. If a user is too busy to create their own keywords, he or she can simply select proposed relevant keywords from a computerized list and add a few more of their own. In this chapter, we propose a novel and practical paradigm

for responding to this type of user’s demand. We generate contextual keywords for photos from readily available *public and personal sources*, modeling the belief that a user is generally the best authority for describing his or her own photographs and that these resources coming from them can usually help generate an accurate interpretation of most photos. Our experiments were conducted on 10 subjects with 313 photographs and the results have proven our theories correct. Our proposed approach contributes to this outcome in three notable ways:

1. Helps reduce semantic gaps. This is because some parts of keywords are their own keywords (personal information) and the remaining parts are those that they are familiar with, obtained from the news, encyclopedias and other sources (public information). Additionally, we introduce the use of named entities to capture the exact meaning of keywords.
2. Semi-automates the annotation task rather than working manually. This system also helps the user to recall events with suggested keywords.
3. Provides a practical implementation framework. This approach is straightforward and is entirely unsupervised. No supervised learning is required to train a prediction of metadata for annotation.

Additionally, we would like to extract more categories of metadata, such as objects (animate and inanimate), events, feeling, actions, numbers. Figure 4.15 illustrates our goal. We also would like to infer their semantic links because understanding the relationships between these keywords of different categories will enhance our existing metadata. Furthermore, the methods described in our “Related Work” section can be complementary to this work. Finally, the methodology presented in this chapter can easily be extended to the other personal media such as video, text and audio residing on one’s computer.

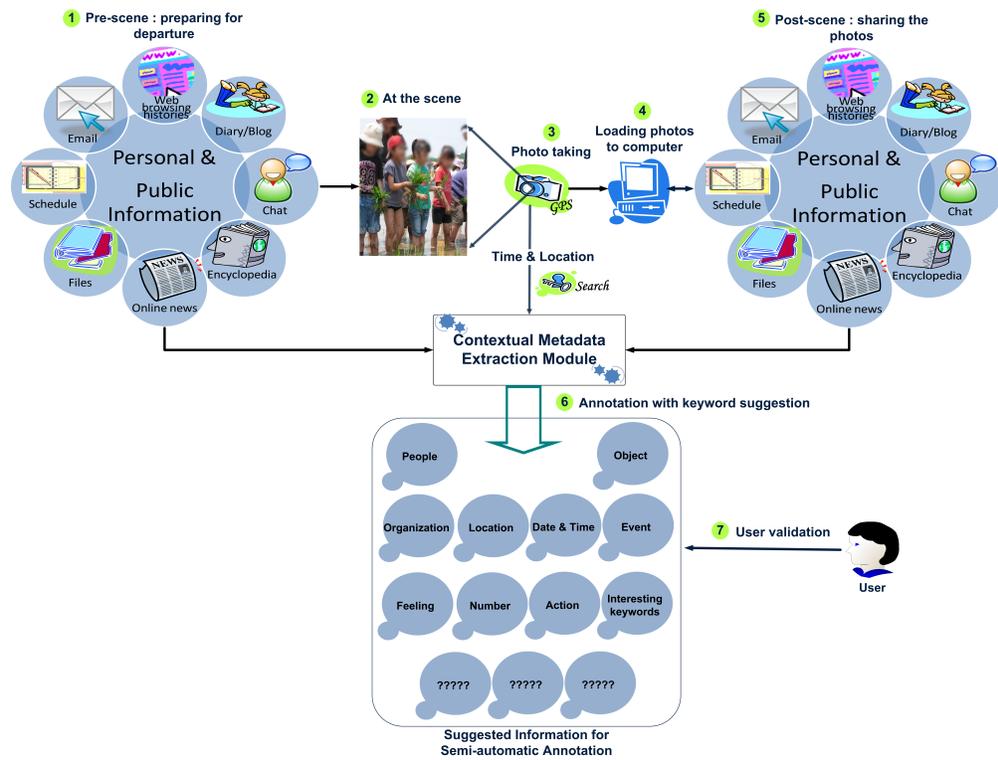


Figure 4.15: The future goal

# Chapter 5

## On Result Re-ranking in Image Retrieval Task

### 5.1 Introduction

The affordability of digital camera and the ease of use of content publishing tool have pushed for the rapid growth of everyday photographs on the web with a large percentage coming from the amateur photographers. These published amateur photographs usually come with either a short description or a few keywords. This shows potentials for image retrieval system to provide better resulting images. Unfortunately, image search engines have very limited usefulness since it is still difficult to provide different users with what they are searching for. Often times, different people issuing the same query are looking for different images. A good image search engine must not produce top results in the ranked list that contain only relevant items of a single theme, but rather diverse items representing sub-topics within the results, yet keeping high level of relevancy.

Thus, in this chapter, we present our development and contributions with the goal to promote diversity in the top ranked list of resulting images.

## 5.2 The Proposed Approach

Using surrounding text of the images or annotation as a means to interpret them is a classic research methodology. To date, however, most research efforts have only been concentrating on relevancy than diversity. The latter is also a quite important factor since the search engine usually knows nothing about the user. Furthermore, most of the time, people solve the problem through selecting some keywords and features of images to represent the photograph rather than trying to understand the semantic nature of annotation and the query. In this chapter, we approach these problems as follows:

- To enable diversity, we use commonsense knowledge as a tool for term expansion. We consider ConceptNet [54] as our commonsense knowledge database. ConceptNet is made up of a network of everyday concepts that have been automatically generated from English sentences of the Open Mind Common Sense corpus. The corpus has been handcrafted by the general public since 2000 [115]. Those concepts are connected by one or more of about twenty relationships such as IsA , PartOf, locationAt, Desires, CapableOf, UsedFor, etc. We use ConceptNet for diversity purposes because a term can be expanded to its contextually related concepts that are not necessarily its synonyms. Furthermore, those related concepts reflect the commonsense way of people’s thinking and how they relate concepts since they are input by human with a specific purpose. For instance, *drink coffee* relates to *wake up*, *yawn*, *read newspapers*, etc. However, diversity should not come as a compensation of relevancy. Therefore, we also try to maintain the level of precision by combining the former with both full-text and location matching.
- Re-ranking technique is performed subsequently to re-rank the results of the previous step by trying to eliminate duplicate and near duplicate results.

Figure 5.1 illustrates the process of our proposed approach.

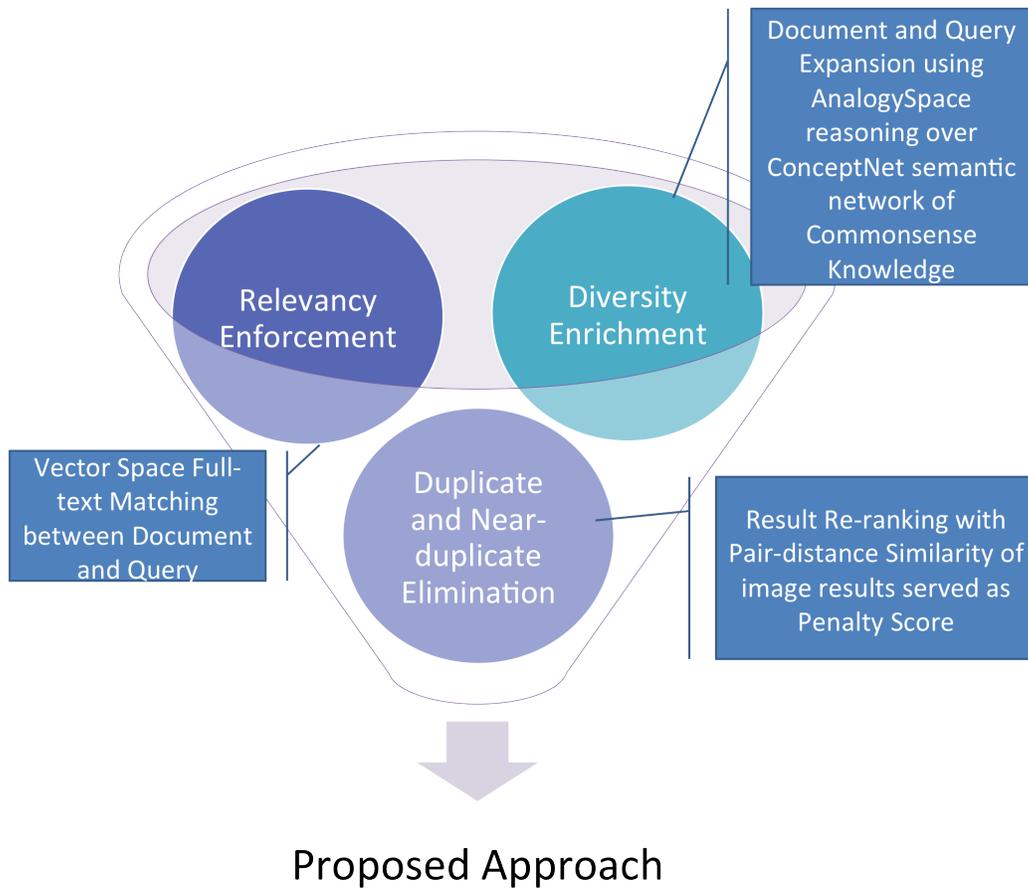


Figure 5.1: The proposed approach

## 5.3 Related Works

Usually, the methods regarding diversity are based on blind clustering whereby duplicate and near-duplicate results are eliminated or ranked using content-based similarity distance. Others simple method includes the re-ranking based on shapes, sizes, colors, etc. Google recently introduced VisualRank a method that guesses how the images would be linked together, with those being most similar having more virtual links to each other. As a result, the most "linked to" images are calculated to rank first [63]. The authors in [21] present a Bayesian retrieval approach that incorporates diversity in the retrieval with a greedy approximation for retrieval. Datta et al. have recently produced a complete survey of the current image related techniques which include methods in diversity promotion [26]. The closely related work to our is that of Hsu et al in [58]. They have used ConceptNet as tool for query and document expansion in image retrieval task. Nevertheless, in doing this, the authors only use spatial relationship function to find the concepts that co-exist in space of the real world.

## 5.4 Implementation

The overall architecture of our proposed approach can be depicted in Figure 5.2. The rest of this section describes each component. It is noted that content pair similarity re-ranking is not implemented in this implementation.

### 5.4.1 Matching

As shown in Figure 5.2, the flow can be divided into two major steps, namely, matching and re-ranking. We introduce three kinds of matching between query and annotation of the image, namely location, AnalogySpace, and full-text.

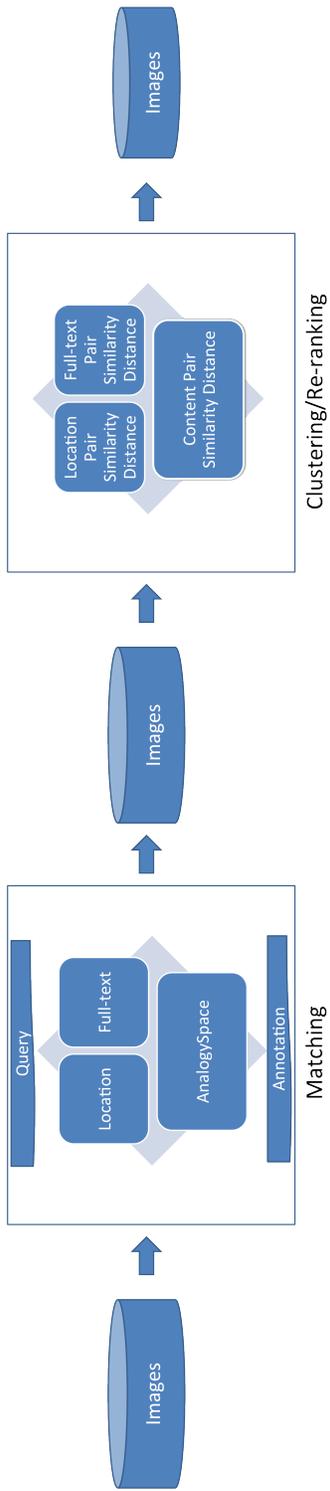


Figure 5.2: Flow diagram of the system architecture

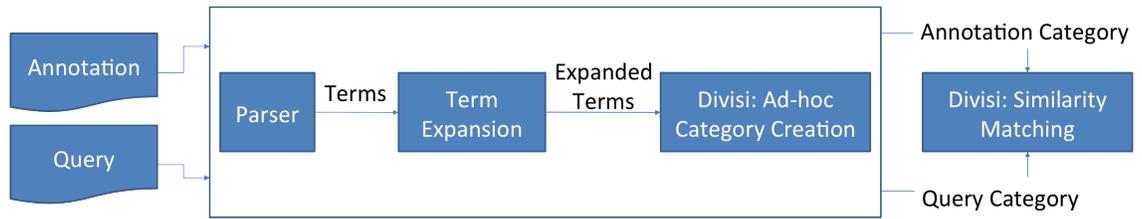


Figure 5.3: AnalogySpace matching

The Figure is adapted from Figure 2 of the author’s paper [IC6]

#### 5.4.1.1 Location matching

We begin by parsing the annotation to get location named entities. GATE is used for this purpose [22]. Then, we establish a location hierarchy from the annotation before we perform the matching. For instance, Lima is expanded to *Lima >> Peru >> South America*. Location names found in image annotations and query topics are expressed as sets with prepositions found in the query as a matching condition. To do this, we simply create two sets of prepositions namely, *include set* and *exclude set*. Prepositions in *include set* are such as ‘in’, ‘of’, ‘along’, ‘on’, ‘near’, ‘by’, ‘in’, etc., while the other set includes prepositions such as ‘out of’, ‘outside’, etc. For example, in the query “Sport stadium outside Australia”, *outside* serves as an excluding condition.

#### 5.4.1.2 AnalogySpace matching

AnalogySpace is a vector space representation of commonsense knowledge built on the top of ConceptNet using Principal Component Analysis [117]. This representation can be used as a reasoning tool as it reveals large-scale patterns in the data while smoothing over noise. In our case, we use an implementation of AnalogySpace called Divisi [2] to create an ad-hoc category for each annotation and query. We then match the query against the annotation. The degree of similarity between the two ad-hoc categories is the dot product of matrices of the shared similar concepts and features.

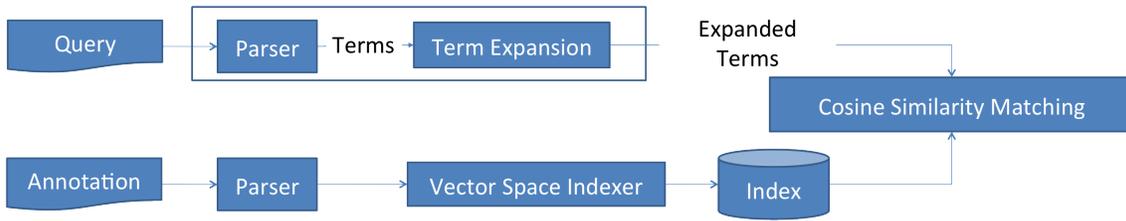


Figure 5.4: Full-text matching

*The Figure is adapted from Figure 3 of the author’s paper [IC6]*

Since ConceptNet depends on sentences contributed from human, it does not contain all the terms a dictionary has. To cope up with unknown terms, we use their synonym and hypernym. We create the set of expanded terms for the unknown term using its Wordnet’s synsets and hypernym [34] regardless of its part of speech. However, we only choose one term as our replacement for the unknown term. The best term is the term that is most uniform to other terms of the annotation. This is achieved via *dot product* of the matrix of an ad-hoc category created from a combination of other terms of the annotation, against the ad-hoc categories created from each term from the expanded set if it exists in ConceptNet. We chose the term that has the highest similarity score. Figure 5.3 shows the process.

#### 5.4.1.3 Full-text matching

Vector Space Model is used to represent the annotations and query topics. Term frequency is used for our vector space model. Each document is represented as a vector, where each dimension corresponds to the frequency of a given term. In our case, terms are reduced to their stems respectively.

Some terms from query topics might not be found in the index of the annotation documents. To cope up with this, we expand unknown query terms with their synsets and hypernym from WordNet. We select the top three terms among the set

of synonyms found. AnalogySpace is used to compute the similarity score between the unknown term and its synonyms.

The similarity distance between a document vector and a query vector is expressed as cosine distance. Figure 5.4 illustrates the technique.

Finally, we normalize each matching score according to its maximum and minimum value. The total matching score is expressed as the product of all the three matching scores. This is the simplest way to combine the scores and yet make the large difference count for even more.

## **5.4.2 Re-ranking**

In this step, the results from the first step are re-ranked according to their semantic similarity by giving penalty to the ones with high similarity between each other.

### **5.4.2.1 Pair distance similarity**

We calculate full-text and location similarity. Same as in the matching process between query topic and photograph annotation, boolean logic is used for location similarity calculation, while vector space model is used for full-text similarity calculation. We compute the total pair distance of images as the product of both distance scores.

### **5.4.2.2 Re-rank**

The similarity distance score obtained can be used to filter and re-rank the preliminary results. We use an optimization method called Hill Climbing to find a threshold of the similarity distance that can help optimize both the precision and diversity. We introduce a loop where Hill Climbing starts with a random threshold and looks for the set of solutions which are better from its neighbors. The loop goes on until

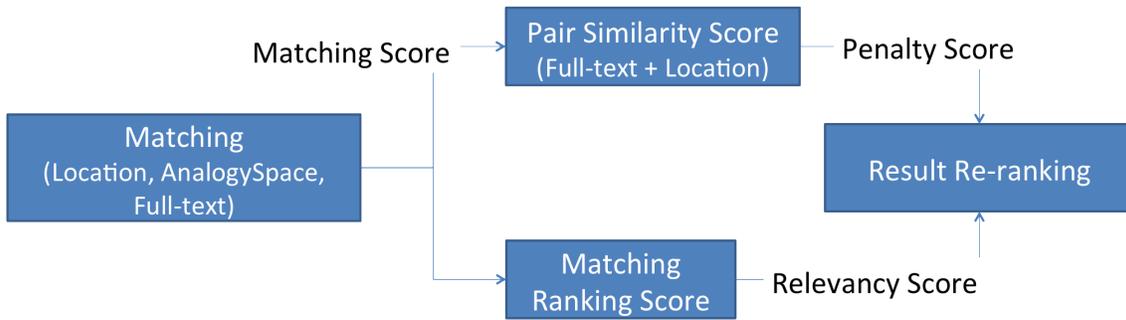


Figure 5.5: Re-ranking process

we obtain the best compromise.

## 5.5 Evaluation

### 5.5.1 Protocol

We participate in the photographic retrieval task of ImageCLEF 2008. ImageCLEF 2008 is a track running as part of the CLEF (Cross Language Evaluation Forum) campaign. It comprises five tasks on image retrieval and annotation techniques, namely, photographic retrieval, medical retrieval, general photographic concept detection, medical automatic image annotation, and image retrieval task from a collection of Wikipedia images. Organizers of ImageCLEF 2008 provide participants with a collection of annotated images, together with query topics. Participants use these resources with their retrieval systems and submit to the organizers the identifiers of the relevant documents for each query topic. Then, the organizers evaluate the result set of each submission from every participant and rank submissions according to standard evaluation measures.

### 5.5.2 Dataset

The collection of images used for ImageCLEF 2008 is the IAPR TC-12 photo collection consisting of 20,000 natural images taken from locations around the world [50].

	DocNo	annotations/07/37394.eng
	Title	The Saint Basil's Cathedral
	Description	a cathedral with crosses on many onion domes; people, trees, a statue and snow on a square in front of it; a grey sky in the background;
	Location	Moscow, Russia
	Date	February 2001
	Image	images/07/37394.jpg
	Thumbnail	thumbnails/07/37394.jpg

Figure 5.6: Example of a photograph of the collection and its attached metadata

*The Figure is taken from Figure 4 of the author's paper [IC6]*

The collection includes images of various sports and actions, photos of people, animals, cities, landscapes and many other aspects of contemporary life. Each image is also associated with an alphanumeric caption stored in a semi-structured format. These captions include the title of the image, its creation date, the location at which the photograph was taken, a semantic description of the contents of the image by the photographer and some additional notes. Figure 5.6 shows the example of a photograph and its metadata. In our system, we use only the title, description, and location parts of the metadata.

### 5.5.3 Query

There are a total of 39 queries used in this study ranging from the very specific to the very abstract ones with different levels of difficulty. Here are some of the query topics:

- "animal swimming",
- "destinations in Venezuela",

<b>Num</b>	<b>2</b>
Title	Church with more than two towers
Cluster	City
Narration	Relevant images will show a church, cathedral or a mosque with three or more towers. Churches with only one or two towers are not relevant. Buildings that are not churches, cathedrals or mosques are not relevant even if they have more than two towers.
Image	images/16/16432.jpg
Image	images/37/37395.jpg
Image	images/40/40498.jpg

Table 5.1: Example of a query topic

- "church with more than two towers",
- "sunset over water", etc.

Query topics are provided as a structured information. It is composed of the query title, cluster, narration of how relevant images should be, and some examples of relevant image files. Table 5.1 shows the example of a query topic. In our system, we use only the topic title.

#### 5.5.4 Measurement techniques

To ensure both relevancy and diversity, the evaluation is based principally on two measures, namely, precision at 20, and instance recall at rank 20 [142]. The technique is a relatively new evaluation methodology that considers results of a query as interdependence rather than a standalone. A good engine will produce results that maximize the two measurements.

## 5.6 Results and Discussions

We present below the results of the four runs.

- *AnalogySpace*: In this run, we combine location matching and AnalogySpace.
- *Full-text*: In this run, we simply use location matching and full-text search.
- *Full-text (no query expansion) + AnalogySpace*: In this run, we combine location matching, fulltext matching, and AnalogySpace matching.
- *Full-text (with query expansion) + AnalogySpace*: The same as the previous one, we combine the three matching. We further expand the terms of query topics in full-text matching with their synsets and hypernyms.

Tables 5.2, 5.3, and 5.4 show the precision, cluster recall, and other metrics, respectively. From the results, we notice that there is only a slight improvement in recall when introducing AnalogySpace. Table 5.3 shows that AnalogySpace helps to gain a little bit better cluster recall at 20 over the conventional full-text vector space model. The number of relevant images retrieved also increases as shown in Table 5.4. However, Tables 5.2 and 5.4 show that the precision at 20 and the Mean Average Precision (MAP) which is the summary of recall and precision do not produce better results with AnalogySpace. We also notice that the improvement happens only when there is no query expansion in the full-text matching. We still believe that ConceptNet could help enriching diversity in the resulting images. To our understanding, the reason why we could not achieve a more significant improvement is because of the fact that there are lots of terms that ConceptNet does not cover. When we try to expand those unknown terms using WordNet, we only introduce noise. We used synonyms from WordNet’s synsets from all its possible senses because we did not implement any sense disambiguation. We did not even check the part of speech. Therefore, most of the time, the replacement only twists the meaning of the original word since we do not select the most appropriate sense of the word. Moreover, we

<b>Runs</b>	<b>P5</b>	<b>P10</b>	<b>P15</b>	<b>P20</b>	<b>P30</b>	<b>P100</b>
AnalogySpace	0.24	0.23	0.22	0.22	0.2	0.11
Full-text	0.32	0.3	0.29	0.27	0.25	0.16
Full-text (no query expansion) + AnalogySpace	0.3	0.28	0.27	0.27	0.25	0.16
Full-text (with query expansion) + AnalogySpace	0.33	0.3	0.28	0.26	0.24	0.16

Table 5.2: Precision (P) at the top  $n$  results

Runs	CR5	CR10	CR15	CR20	CR30	CR50	CR100	CR1000
AnalogySpace	0.09	0.13	0.16	0.21	0.24	0.27	0.35	0.67
Full-text	0.14	0.21	0.25	0.28	0.35	0.46	0.55	0.81
Full-text (no query expansion) + AnalogySpace	0.14	0.21	0.25	0.31	0.38	0.46	0.54	0.82
Full-text (with query expansion) + AnalogySpace	0.13	0.19	0.22	0.25	0.33	0.42	0.52	0.8

Table 5.3: Cluster Recall (CR) at the top  $n$  results

<b>Runs</b>	<b>NumRelRet</b>	<b>NumRel</b>	<b>MAP</b>	<b>GMAP</b>	<b>BREF</b>
AnalogySpace	1247	2401	0.14	0.01	0.51
Full-text	1420	2401	0.21	0.06	0.64
Full-text (no query expansion) + AnalogySpace	1451	2401	0.2	0.06	0.65
Full-text (with query expansion) + AnalogySpace	1462	2401	0.2	0.04	0.65

Table 5.4: Other metrics: Number of Relevant Retrieved images (NumRelRet), Number of Relevant images (NumRel), Mean Average Precision (MAP), Geometric Mean Average Precision (GMAP), Blind RElevance Feedback (BREF)

limit the number of selected synonym to only one in AnalogySpace term expansion, and only up to three in our full-text query expansion. This reduces the coverage of the meanings. Moreover, content-based technology was not taken into consideration. Should we have incorporated another content-based pair similarity distance in the re-ranking step, we might be able to get better resulting images. Hence, we are planning to tackle these issues in our future works.

## 5.7 Conclusion

User’s satisfaction is not solely a function of relevancy. When nothing is known about the user, diversity plays an important role in getting the results that user would like to see. We present a novel approach to enable rich diversity in the results by incorporating commonsense knowledge expansion and result re-ranking through elimination of duplicate and near duplicate results. The presented results are just our preliminary ones. Even they are not conclusive yet, they pave the way to help us to improve our current system. We are now working to address the weak points that we have discussed earlier.



# Chapter 6

## On Categorization and Aesthetics Quality Assessment

### 6.1 Introduction

#### 6.1.1 Background and Motivation

The tremendous increase in the number of digital photographs also brings a relatively large increase of high quality and interesting photographs. Image aesthetics is still a very new area of research, though there is a growing trend in recent years. There are many applications for this area of research. Below, just to name a few, are the obvious examples:

- Media companies - especially stock photo, advertising and printing companies - usually have huge collection of high quality photographs. The task of selecting a suitable picture for a targeted theme is, and will still be, a burden, even though there are annotations in the collection. For instance, how does one select an image that depicts *freezing action*, an image that has a *great depth of field* or an image that *implies motion* for a front cover of a magazine?

- Quality is an important factor for image results filtering in addition to the popular relevancy and diversity measure for image search engine. Usually, the combination will produce better resulting images and enhance the user experience. Furthermore, image browsing and summarizing systems based on specific theme and/or quality are in demand.
- Photograph aesthetics assessment engine can be a useful tool to help both professional and amateur photographers to evaluate their work earlier. This would especially help to foster more new artists and new art works.

### 6.1.2 Problem Formulation and General Idea

Our goal is to help to solve these very obvious but difficult problems. The first question in which this research addresses is: *how should high quality photographs be classified?* We would like to look at the problem from another angle, and that is from the perspective of professional photographers and artists. In this work, we propose to study high quality photos by their *visual aesthetic primitives*. For this, we explore the role of those camera setting parameters which are increasingly available, as well as image content features. Then, we classify the photographs into *six creative exposure themes* defined by professional photographer. Our second question is on *quality assessment*. We use the defined themes for an image media quality assessment inference, rather than observe it boldly. We believe that such decomposition will give us better performance than previous efforts because each theme exhibits a different nature of content. We are also careful to make certain our work is reproducible by using public and standard available dataset; and will make our finally results available online for future comparisons.

## 6.2 Related Works

Since computational analysis of art is an emerging research, the number of research efforts in this domain is still limited. The following reviews the closely related work.

### 6.2.1 Categorization and Annotation

There have been research efforts trying to classify and annotate art works. Cutzu et al. proposed a framework for distinguishing painting from photographs based on spatial variation of colors, color edges, number of unique colors, and pixel saturation [23]. They found that a combination of these features can produce good result but no single feature could do the task alone. Other efforts in the two-class photos classification include: photos versus graphics in [15], city versus landscape in [12], indoor versus outdoor in [83,100,108], and real versus rendered in [85]. Marchenko et al. tried to annotate and classify modern and medieval artworks [89,140]. They used features such as color temperature, color palette, color contrasts, texture features, brush stroke analyze and annotation with high-level concepts with some success. Wallraven et al. studies the categorizing tasks of painting both by human and computer [135]. The study revealed that non-expert human can reliably classify painting into meaningful categories. As for computer, the author use features of computational measures sensitive to color, texture, and spatial composition to do the task. The result suggests that none of the computational measures - with the notable exception of the Gist feature - correlated with human data. Ku et al. proposed to use EXIF information for scene mode classification [72].

### 6.2.2 Aesthetic Quality Assessment

#### 6.2.2.1 Content-based approach

Earlier work on image quality assessment such as presented in [24] distinguishes original image from the degraded version without looking at its semantic value. Our

closer related work begins with Tong et al. who tried to separate snapshot from professional [125]. In [67] Ke et al. defined a number of high level features for photo classification between snapshot and photos taken by professional photographers. Datta et al. [25] proposed 56 computational features for the task of quality assessment. Based on this work, very recently an online aesthetic quality inference engine called ACQUINE was launched [6]. In a recent work, Li et al. proposed 40 features in trying to evaluate the quality of famous painting [76].

### 6.2.2.2 Subjective approach

Since image quality is highly subjective, some researchers have resorted to psychological experiment with or without the combination of content-based approach [42], [70]. In [66], Katti et al. did experiment to confirm that people can discriminate interestingness in pre-attentive ( $< 50ms$ ) time spans. The result suggests that interestingness appears to be detectable in such a short time.

## Summary

There are diverse efforts in classification and aesthetic inference. However, to the best of our knowledge, our work on classification of high quality photographs by focusing on the creative exposure themes and infer quality based on these themes is the first attempt so far. We are the first one to follow the guideline of aesthetic primitives for visualization. Research in [25,67,76] did analyze the aesthetics qualities but they did so from intuition and from their background in arts. For instance, some aesthetic properties are missing such as depth and principle axe of human body. We are also the first one to incorporate temporal and optical features for aesthetic inference. We also incorporate the global Gist features which show some success in [135] and may help in discriminating interestingness as reported earlier in [66].

## 6.3 Proposed Approach and Framework

### 6.3.1 Conceptual Approach

#### 6.3.1.1 Aesthetics and Categorization

In our attempt to classify the photographs, we first have to characterize the photograph from the artistic perspective. In the recent study [101], Peters defines six main visual aesthetic primitives that evoke pleasurable feelings namely, *colors*, *form*, *spatial organization*, *motion*, *depth*, and *human body*. She recommends the following as rules of thumb.

- Only a few strong *color* should be used; complementary contrasts are effective; utilize the dynamic range.
- *Form* should be clear and simple; silhouettes are aesthetic.
- *Spatial organization* of image elements should be clear and simple; apply the rule of the golden mean; texture and pattern can create a holistic impression; apply variations to patterns and take care for the visual rhythm induced by repetition of elements.
- *Motion* can be expressed by blur of high contrast; distinct motion phases are aesthetically appealing.
- *Depth* should illustrate linear perspective; exploit the contrast between sharpness and unsharpness; the distribution of light and shadow can also give the impression.
- Have the principle axes of the *human body* be clearly visible.

There have been lots of studies about categorization of arts. In painting, people usually do the classification by artist, historical period, or group by distinct style. For photography, there is no general agreement. In our opinion, we believe that photographer should be the one who has the authority. Therefore, in this work, we refer to

the professional photographer to define the classes of high quality photographs. In photography, exposure control being a process of controlling light striking a camera's digital sensor is the main actor to successful photography. Exposure is determined by three setting - shutter speed, lens aperture and ISO.

- The shutter speed is the duration of time that the shutter of the camera remains open, allowing light to get in and expose the sensor.
- The aperture (or *f-stop*) is the size of the adjustable lens diaphragm, which dictates the amount of light entering the camera.
- The ISO indicates the sensor's sensitivities, the sensor requires a longer exposure to get a good result, whilst at high sensitivities, less light is needed [56].

Correct combination of these three will result in a good photo - *a well exposed photo*. Obviously, there are many of such combinations that can result in a well exposed photo. However, among them only a few can give interesting photographs. In his book entitled *Understanding Exposure* [102], Peterson distinguishes seven classes of high quality photographs by exposure theme. He calls them *creative exposure themes*. Furthermore, he discusses the characteristics and the rules that can be used to produce those images. Usually, when taking a photo, photographer has in mind which type of photo he or she is going to make and configure the camera setting accordingly. This has effectively provided the basis for photo classification. In this study, we focus only on six exposure themes because we have limited number of photos that correspond to the seventh theme in the proposed dataset. The following explains each theme and Figure 6.1 shows the example images of those themes.

- *Story Telling*: when we want great depth of field with all objects inside neat and clear. It is usually done using wide angle lens and small aperture.
- *Who Cares*: when the depth of field is not a concern and when subjects are at the same distance from the lens. It is usually done with middle range aperture.

**(A) Story Telling**



 **Giorgio Giorgetti**

**(D) Freeze Action**



 **Guiuri R. Reyes**

**(B) Who Cares**



 **John Blyberg**

**(E) Imply Motion**



 **Aja Bach**

**(C) Isolation or Single Theme**



 **Mikhail Esteves**

**(F) Macro or Close-up**



 **Juergen Mangelsdorf**

Figure 6.1: Example images of the six creative exposure themes

*The Figure is taken from Figure 1 of the author's paper [IC3]*

- *Isolation or Single Theme*: when we want to focus on a specific subject. It is usually done with a large aperture open. Usually, the unfocused part is blur.
- *Freeze action*: when we want to freeze and capture the moment. This is usually done using very fast shutter speed.
- *Imply motion*: when we want to convey motion to the audiences. This is usually done using very slow shutter speed.
- *Macro or Close-up*: when we want the great detail of the subject or just part of it in close proximity. Usually, we want to record the image from 1/10 to 10 times or more of the actual size. The image often lacks of depth of field.

### 6.3.1.2 Camera Setting Context

As described above, lens aperture, shutter speed, and ISO play important roles in creating a correct exposure for each theme. Fortunately, unlike conventional camera, current modern digital cameras are equipped with many sensors. Many kinds of information are recorded at the same time when a photograph is taken. If we make an analogy of those sensors to our human eyes, this captured information represents the *intention* of the (professional) photographers. Specifically, two main things can be extracted: *photographer's intent* and the *condition in which the image is captured*. EXIF specification [1], which is universally supported by most of digital cameras, enables these settings<sup>1</sup>. Some of the important parameters which professional photographers usually refer to and which can be found in the EXIF header of the each image file are: *Lens Aperture, ISO, Exposure Time/Shutter Speed, Date and Time, Focal Length, Metering Mode, Camera Model, Exposure Program, Maximum Lens Aperture, Exposure Bias, Flash, etc.*

---

<sup>1</sup>It is noted that EXIF is supported by only JPEG and TIFF.

## 6.3.2 Research Framework

### 6.3.2.1 Framework of the Approach

We define the feature extraction as well as the classifier model to perform automatic categorization. The aesthetic characteristics defined by Peters can be found in both global and local features of the image. The EXIF metadata discussed earlier can give us information about optical and temporal context. The same type of features extracted for the exposure theme classification can also be used for aesthetics evaluation. Both the classification and aesthetics evaluation are treated as machine learning task where we separate the photos into training set and test set. Figure 6.2 illustrates the framework of our conceptual approach.

We understand from the start that aesthetics is a highly subjective task. However, we believe that using data-driven approach, to some extent, we will be able to draw some general conclusion about the quality. Moreover, by dividing the photos into different exposure themes, the performance of the quality inference model could be improved.

### 6.3.2.2 Feature Extraction

Feature extraction is an important part of this research. Here, we define features to represent the characteristics of aesthetics based on aesthetic primitives for visualization described earlier. Those features are from global, local, temporal, and optical sources. Global features give the holistic view of the photo similar to human’s first impression while local features would help to represent some most salient parts of photo. Optical and temporal features can inform extra contextual information of the scene. Some of the features are taken from the previous research efforts [25, 67, 76]. We have not finalized the list of features yet but below are the current considerations:

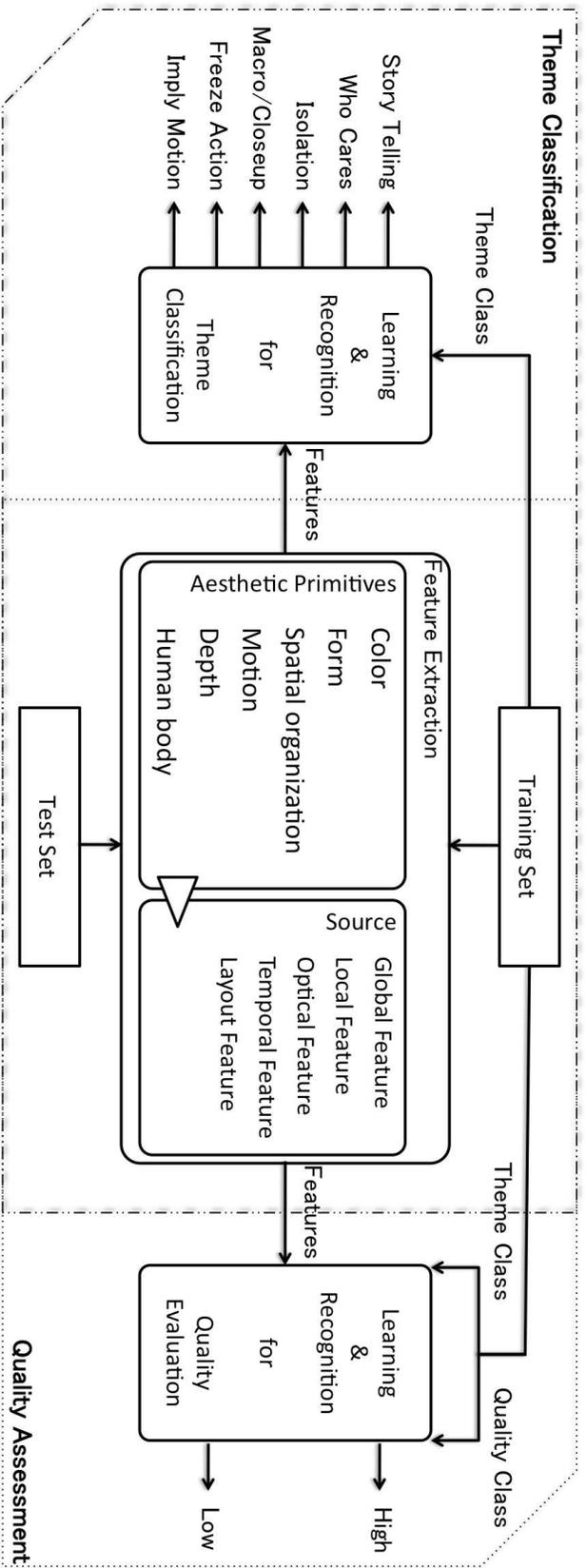


Figure 6.2: Conceptual approach of exposure theme classification and photo quality assessment  
*The Figure is taken from Figure 2 of the author's paper [IC3]*

**Color** A number of color features are considered including *color distribution* [76], *colorfulness*, *exposure of light*, *saturation* and *hue* [25], *contrast and brightness*, and *hue count* [67].

**Form** Shape recognition algorithm for simple objects like lines, circles, rectangle, and squares is considered [17]. To estimate the simplicity of the form, a Gini purity coefficient is to be calculated. Silhouette detection is also to be explored [13].

**Spatial organization** A number of features in this category are to be extracted:

- Golden mean: This is what photographers sometimes refer to as the Rule of Thirds. It is when the ratio between the sum of two quantities and the larger one is the same as the ratio between the larger one and the smaller (approximately 1.618). We can apply Datta et al.'s approach for this [25].
- Size and aspect ratio: Specific size and aspect ratio can affect how we see the image and thus can affect the rating. Size is calculated as the sum of both width and height and ratio as their scale ratio.
- Simple spatial organization: For this, we can compute the spatial distribution of edge as in [67].
- Texture and pattern: Several texture and pattern extraction algorithms are considered including wavelet-based [25].
- Other spatial properties: The spatial envelop properties or Gist of the scene, which have been used to characterized scene without object detection or recognition, are important as global features [99]. Those perceptual dimensions are: degrees of openness, naturalness, roughness, expansion, and ruggedness.

**Motion** Shutter speed and blurriness can be used to characterize motion. Shutter speed can be extracted from EXIF metadata while blurriness can be estimated from the content [67].

**Depth** Image depth can be estimated based on the whole scene image structure using the methods of Torralba et al. [127].

**Principle Axe of Human Body** For the time being, we have only considered the face detection. We intent to use the haar-like features for fast face detection [79].

### **Other features**

- **Temporal context:** Date and Time are important features. For instance, the time can implicitly tell us about the present of things like sunset, sunrise, day and night. The date can indicate the season which tells us whether it is likely to be indoor or outdoor activities. Also, it can tell in which season, the photo might be taken. With correlation with the learning data, this can help us predict the things that the low-level features cannot get.
- **Optical context:** Beside the optical context features used earlier, others can also be useful such as *focal length*, *ISO speed* and *camera type* (point-and-shoot versus digital single lens reflection).

#### **6.3.2.3 Feature Selection and Classification**

The features have to be correlated to the image. In this case, firstly, there should be a test to determine the correlation between those features and image. A correlation threshold shall be established.

Once the features are selected, given a list of features, we want a technique to combine those features. A naive method would be a weighted linear combination of the features. However, the values of the feature metrics are not linear. Therefore, it may not work. In this regards, a number of machine learning algorithms are being considered including Naive Bayes, AdaBoost, SVM, Decision Tree, etc.

## 6.4 Evaluation Protocol

### 6.4.1 Dataset

One of the key problems in this research is the dataset. It is difficult to find a large standard set of image not to mention the high quality and interesting image set for the experiment. The authors in the previous work have used different datasets [67] [25]. Datta et al. have made available the dataset that they used in [27] in form of image links and the aesthetic scores. The images in the dataset are that from Photo.net [3] and Digital PhotoChallenge [7] sites.

The last resort would be to annotate the photos of the MIR Flickr dataset by ourselves. If this is the case, then we will setup a website so that people can participate in our campaign for annotation. Furthermore, we can make this dataset available for future use. The downside is that it is a time consuming and expensive task.

Recently, a MIR Flickr 25000 test collection is available [60]. The photos in the collection are selectively taken from Flickr<sup>2</sup> based on their high interestingness rate [19]. The image collection is rich in original and high quality photography. 75% of them have the 5 major settings namely, *Aperture Number*, *Exposure Time*, *Focal Length*, *ISO Speed* and *Flash*. However, we need to classify the dataset into the six creative exposure themes as well as by quality rating for our experiment. Fortunately, the ImageCLEF Large Scale Visual Concept Detection and Annotation Task (VCDT) [98] of the Cross-Language Evaluation Forum (CLEF) 2009 annotate the dataset with different concepts for their competition task. All the concepts refer to the holistic visual impression of the photo. The complete set of the concepts is shown in Table 6.1. We can roughly define the correspondence between the six

---

<sup>2</sup>Flickr Website: <http://www.flickr.com>

Category in Ontology	Annotation
Scene Description	<i>Abstract Categories:</i> PartyLife, FamilyFriends, BeachHolidays, BuildingSights, Snow, Citylife, LandscapeNature, Desert; <i>Activity:</i> Sports; <i>Seasons:</i> Spring, Summer, Autumn, Winter, NoVisualSeason; <i>Place:</i> Indoor, Outdoor, NoVisualPlace; <i>Time of Day:</i> Day, Night, NoVisualTime, Sunny, SunsetSunrise
Landscape Element	Plants, Flowers, Trees, Sky, Clouds, Water, Lake, River, Sea, Mountains
Representation	Canvas, StillLife, Macro, Portrait, <i>Illumination:</i> Overexposed, Underexposed, Neutral
Quality	<i>Blurring:</i> MotionBlur, OutOfFocus, PartlyBlurred, NoBlur; <i>Aesthetics:</i> Fancy, OverallQuality, AestheticImpression
Pictured Object	<i>Person:</i> SinglePerson, SmallGroup, BigGroup, NoPerson, Animals, Food, Vehicle

Table 6.1: ImageCLEF VCDT Concepts

exposure themes, quality rating and the annotation concepts of ImageCLEF VCDT as shown in Table 6.2. However, the organizers of ImageCLEF VCDT are not sure whether to release the ground truth for the public after the competition, for the reason that they want to use the ground truth again for next year task.

## 6.4.2 Analysis

We plan to build our classifier using different machine learning algorithms. The results will be in the form of confusion matrix.

Let

- $TP$  : TruePositive,  $TN$  : TrueNegative,
- $FP$  : FalsePositive,  $FN$  : FalseNegative,

then, we can calculate the performance of each established model as follows:

<b>(A)</b>	
<b>Exposure Themes</b>	<b>Equivalent Annotation Concepts</b>
Story Telling	Landscape Nature <i>AND</i> NoBlur (with depth)
Who Cares	Canvas <i>OR</i> ((PicturedObject <i>OR</i> Portrait) <i>AND</i> NoBlur)
Isolation	(Person <i>OR</i> PicturedObject) <i>AND</i> PartlyBlur
Freeze Action	Sports
Imply Motion	MotionBlur
Macro/Close-up	Macro

<b>(B)</b>	
<b>Quality</b>	<b>Equivalent Annotation Concepts</b>
High	Aesthetic Impression <i>OR</i> Overall Quality <i>OR</i> Fancy
Low	Normal

Table 6.2: Correspondence between: (A) Creative exposure themes and Annotation concepts, (B) Quality and Annotation concepts

- Percentage of positive predictions that are correct

$$Precision = \frac{TP}{TP + FP} \quad (6.1)$$

- Percentage of positive labeled instances that were predicted as positive

$$Recall/Sensitivity = \frac{TP}{TP + FN} \quad (6.2)$$

- Percentage of negative labeled instances that were predicted as negative

$$Specificity = \frac{TN}{TN + FP} \quad (6.3)$$

- Accuracy (percentage of predictions that are correct)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.4)$$

## 6.5 Challenges

- We need to define more features accordingly to the aesthetic primitives. For example, salient local feature extraction and how to detect axe of human body are still under investigation.
- The variation of camera types can have influence on the optical parameters. There might be some deviations of EXIF metadata due to the different hardware specifications.
- We need to deal with multiple category issue because there are cases that photos can belong to two classes.
- Standard dataset and ground truth are in need.

## 6.6 A Preliminary Experiment

With the above considerations, there is an obvious relationship between creative exposure themes and some of the camera setting parameters. Thus, in this preliminary work, we propose to categorize the photographs into six creative exposure themes and tackle the problem computationally and experimentally using statistical learning approach by applying only the camera setting parameters.

### 6.6.1 Dataset and Extracted Features

We use the MIR Flickr 25000 test collection presented earlier. We use 5 major camera settings that are available namely, Aperture, Exposure Time, Focal Length, ISO Speed and Flash. Based on the camera model found in EXIF, we also distinguish *Point-and-Shoot* cameras with *Digital Single Lens* Reflection ones. For our study, a subset of the collection (2736 photos) is labeled into the six themes. The labeling process is done manually based on the strong correspondence of the visual expression of each of the photos to the six creative exposure themes. One problem that we faced during the labeling process is that some photos can be attributed to multiple themes. For that we put the photo to the most suitable class.

### 6.6.2 Model Building, Evaluation and Results

We divide our dataset into training (2/3) and testing sets (1/3). We carefully create the random splits within each class so that the overall class distribution is preserved as much as possible. With the training set, several machine learning algorithms such as Decision Tree, Forest, SVM and Linear combination were used to train the dataset and create the models automatically. Finally, to evaluate the models, we test them with the testing set. The confusion matrix is computed. We calculate the performance of each established model by the following measures: precision as percentage of positive predictions that are correct, recall/sensitivity as percentage of positive labeled instances that were predicted as positive, specificity as percent-

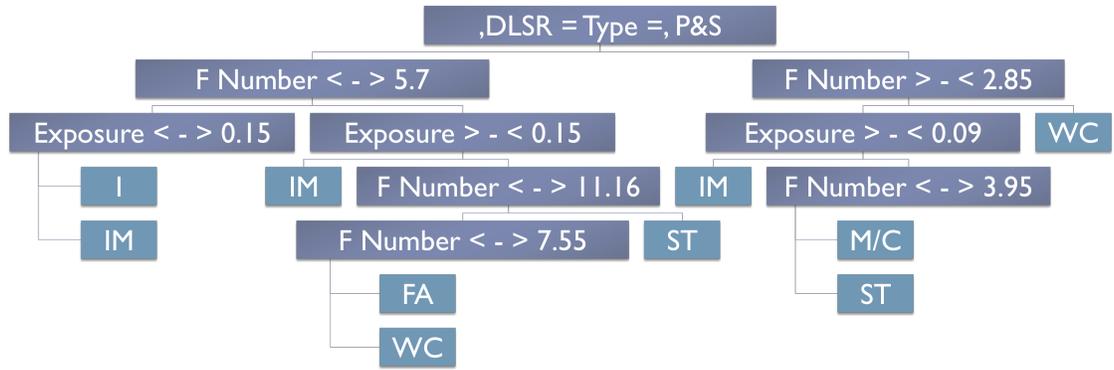


Figure 6.3: Generated decision tree model

The Figure is taken from Figure 2 of the author’s paper [IC5]

		Actual Theme					
		FA	I	IM	M/C	ST	WC
Predicted Themes	FA	15	0	1	5	0	0
	I	74	169	10	41	1	1
	IM	0	0	58	4	6	2
	M/C	6	0	2	7	0	0
	ST	30	0	5	23	127	0
	WC	28	0	13	30	1	253

Table 6.3: Confusion matrix

age of negative labeled instances that were predicted as negative, and accuracy as percentage of predictions that are correct. Decision Tree which is rather simpler than other models gives the best performance of all. Due to limited space, we show only our best result. Figure 6.3 depicts our generated model while Table 6.3 and Table 6.4 show the performance of the model.

Though we obtained a reasonable performance using very few features, for our future work, we would like to see how the integration with other type of features could help this task even more with regards to the trade-off of computational costs. For our immediate study, content-based features such as color, texture, shape, and scene description will be integrated. We also would like to perform our experiment

	FA	I	IM	M/C	ST	WC	Average
Precision	0.71	0.57	0.82	0.46	0.68	0.77	0.67
Recall	0.09	1	0.65	0.063	0.94	0.98	0.62
Specificity	0.99	0.82	0.98	0.99	0.92	0.89	0.93
Accuracy	0.84	0.86	0.95	0.87	0.92	0.91	0.89

Table 6.4: Precision, Recall/Sensitivity, Specificity and Accuracy rates (Let  $TP$  : *TruePositive*;  $TN$  : *TrueNegative*;  $FP$  : *FalsePositive*;  $FN$  : *FalseNegative*)

on larger dataset with multiple annotators to avoid any bias.

## 6.7 Conclusion

We present our research proposal targeting high quality photographs which are becoming more important as the amount of photos increases sky high and people are demanding more adaptive content. We discuss the state-of-the-arts of the research and present our new conceptual approach towards more effective techniques for the tasks of classification and quality evaluation of such images.

We have done extensive literature review. The elements of visual aesthetic primitives and the categories of creative exposure theme have been identified, and a preliminary framework for creative exposure theme-based classification and aesthetics quality inference has been formulated. We have also described our dataset and the implementation part. The result of our preliminary study on the classification task using only the camera setting features is encouraging. The proposed features from the previous chapters can easily be integrated.

We believe that in the future with the evolution of the digital camera (i.e. with more advanced settings, programmable functions, better optical precision, other sensory inputs, etc.), this research will become more relevant and important. It could be applied in either prior, real-time, and post photo taking sessions. The following are some possible examples:

- In prior photo-taking session, the photo quality inference system can help the users to learn different visual properties of a high quality photo in each category.
- In real-time photo-taking session, the proposed research can be used to automatically or semi-automatically help the users to take high quality photos.
- In post photo-taking session, as discussed earlier, it can be used to classify the photographs, infer their quality, etc.

Last but not least, this research is not limited to photographs, other visual related works can be beneficial from its finding. Some obvious examples include, painting, drawing and other creative works.

# Chapter 7

## Conclusion and Future Perspectives

### 7.1 Summary

We are now living in an image explosion era where tools for managing and digesting such overloaded number of images become extremely important for our daily life. This thesis helps alleviate the burden by proposing various mechanisms in image analysis and its methodology design from inter-disciplinary areas (i.e. social, cognitive science, computer vision, machine learning, etc.). The proposed methods contribute to the semantic understanding of image by going beyond the superficial image content analysis. They either fully exploit the holistic content analysis, contextual understanding, other related information about the image or the combination of them for this difficult task. Specifically, the thesis focuses on *automatic image annotation*, *result re-ranking*, *categorization* and *aesthetics quality assessment* tasks and it can be summarized as the following.

In *Automatic Image Annotation* (AIA), for personal digital photographs, the thesis proposes a personal photo library system with built-in annotation engine to lessen user's work. Photo with Geo-referential (GPS) information is the current tendency. Using the exact location information given by GPS together with timestamps, the

novel engine semi-automatically generates contextual metadata for each photo from different sources of information namely, the *public information* and user's *personal information*. As for general image annotation, the thesis leverages the use of salient region and background in addition to the whole original image for a holistic feature extraction and better annotation scheme. 43 diverse image features are extracted and the K Nearest Neighbor approach is used for annotation propagation. The experiments confirm that the proposed methods in these AIA tasks are efficient and effective.

In *Result Re-ranking*, the thesis concentrates on the retrieval task. Image search systems have a very limited value since it is still difficult to support different users with what they are searching for. This is because most research efforts to date have only been concentrating on relevancy rather than diversity which is also a quite important factor, given that the search engine knows nothing about the user's context. In the proposed approach, the author makes use of commonsense knowledge and its reasoning tool for document and query expansion, which aims to increase the diversity of the results. The technique combines *AnalogySpace* mapping with other two mappings namely, *location* and *full-text*. Afterward, re-ranking is employed to the resulting images from the mapping in order to eliminate duplicate and near-duplicate results. The results show that the integrated method yields better performance in terms of cluster recall and the number of relevant photographs retrieved.

In *Categorization and Aesthetic Quality Assessment*, the thesis outlines a proposed framework for the tasks. It addresses these challenges by exploring the aesthetics from the combined perspectives of the artists and the photographers. The proposal utilizes the *aesthetic primitives* of images for visualization as a guideline for high and low-level image feature extraction and to classify this high quality content into *six creative exposure themes*, which are commonly followed by the professional photographers. Subsequently, the quality assessment can be done accordingly to these

themes. A preliminary experiment using only the camera setting features is conducted and the result is encouraging.

These analysis and methodology design presented in the thesis shall contribute to *the better understanding of image beyond just the superficial analysis of image content*. Many *fully targeted applications and services* - not limited to visual related ones - could rise from these findings. Furthermore, this thesis becomes even more relevant and important with the current trend of technologies and user's behaviour (i.e. number of image is growing sky high, the advancement of digital camera, the availability of more sensory data, and the social interaction trend).

## 7.2 Future Perspectives

*If an image is worth a thousand words, then what is the combined value of a collection of images?* We are now living in the world with billions of images. The future perspectives of image understanding would be to explore the connection between those images and eventually to infer knowledge from them. In addition, it would be interesting if we can make use of this huge volume of image content to help augment the understanding of other kind of media such as video, text or audio. Below are some considerations of how we can make sense from the large collection of images.

### 7.2.1 Structuring

Once we have gathered all the images and the related information, we need to make the structure out of those images. The processing steps could be as follows:

1. First, it is imperative to give meaning to each of the image. One way is to associate each image with some meaningful keywords, their category types, their quality properties, etc. This could be built upon the research findings in this thesis.

2. Recently, there have been many efforts in building a semantic lexical network in different forms: Japan's NICT Concept Dictionary [120], Princeton's Wordnet [91], MIT's ConceptNet [54], Cyc [75], etc. . These large semantic networks of concepts are particularly useful. We would like to map the image with the corresponding concept in the lexical database through its annotation generated earlier. The existing relationships from the lexical database will be helpful to reinforce the data. For example, this will help us to further refine our annotations by eliminating noise (wrong annotations) since all the concepts are linked together with meaningful relationships. Moreover, it is intriguing to investigate other image datasets that have some built-in relationships. One example is the ImageNet which is organized by WordNet hierarchy [28].

### 7.2.2 Making sense

There are many potential research works that we could explore from the structured image contents. The following are just some possible examples:

1. *Image Distance*: Assume that all the images are semantically annotated, we can look for ways to help consumers explore their collection of images efficiently, effectively and joyfully. The focus will be on finding the multi-dimensional relationship of images. For instance, browsing, searching and sharing would be much more interesting and efficient if multi-dimensional relationship of images (or *Distance of images*) is well defined and established. The approach of the research can be based on the combination of one or more of the following items: contextual information, content features, semantic lexical dictionaries and other related resources.

A mathematical model or measurement of similarity shall be established. Currently, most of the work in similarity is based on content based technology. A number of distance metrics have been introduced for this purpose such as the

Mahalanobis distance [53], the intersection distance [122], the earth mover's distance (EMD) [105], etc. In [92], a similarity measure is defined from subjective experiments and multidimensional scaling based on the human's perception model in understanding color patterns. There are also works combining text and visual information such as that the probability-based similarity scheme introduced by Barnard et al. [16] and Google Image Search [44]. Nevertheless, the research efforts so far are still superficial. Mathematically, defining a similarity measure is equivalent to defining the distance between points in high-dimensional feature space. The basic idea is to establish a model on how to effectively represent the images in the vector space with both contextual, content and other related features. Subsequently, a distance measurement between two sets of feature points of respective images shall be provided.

There can be many possible applications and visualization methods when the *image distance* is realized.

2. *Community-based clustering or identification*: all of us belong to one or more communities, meaning that usually we are not the only one who experiences any event that we are participating and the contextual information is spread within or across communities. This is also the case for the images taken by us. It would be interesting to categorize the images and identify the communities that they belong to. For this, we could explore many theories including the small-world theory and content distribution research [118,119].
3. *Extracting knowledge from images*: Exploring the possible relations between contextualization and personalization is of particular interest. If the annotation and the relationships are accurate and meaningful enough, we should be able to establish the semantic links between images and the world of information (e.g. user's information inside their own computer and/or from elsewhere such as those from the World Wide Web). Browsing images is typically a very



Figure 7.1: Extracting knowledge from images

enjoyable experience. It would even be better if we could map the enjoyment and engagement in mind, which is to use images to recall, explore knowledge and as memory aid tools to human. This is the ultimate purpose. For instance, we could imagine using the image collection to help us find some objects inside your house that you want to look for. Figure 7.1 illustrates the general concept.

# Appendix A

## List of publications

### Journal

J1 - Supheakmongkol SARIN, Michael FAHRMAIR, Matthias WAGNER, and Wataru KAMEYAMA, "Leveraging Salient Regions and Background for Automatic Image Annotation," IPSJ Journal of Information Processing, Special Issue on IT Infrastructures for Info-plosion, vol. 20, no. 1, pp. 250-266, 2012 (doi:10.2197/ipsjip.20.250)

J2 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "On the Design and Exploitation of User's Personal and Public Information for Semantic Personal Digital Photograph Annotation," Journal of Advances in Multimedia, vol. 2008, issue 3, pp. 1-16, 2008. (doi:10.1155/2008/592690)

### International Conference/Symposium/Workshop

IC1 - Supheakmongkol SARIN, Michael FAHRMAIR, Matthias WAGNER, and Wataru KAMEYAMA, "Holistic Feature Extraction for Automatic Image Annotation," in Proc. of International Conference on Multimedia and Ubiquitous Engineering (MUE 2011), pp. 59-66, IEEE ComSoc, June 28-30, 2011, Loutraki, Greece

IC2 - Kok-Meng ONG, Supheakmongkol SARIN, and Wataru KAMEYAMA, "Affective and Holistic Approach at TRECVID 2010 Task - Semantic Indexing (SIN)," In Notebook papers of TRECVID 2010, Gaithersburg, MD, USA, Nov 15-17, 2010

IC3 - Supheakmongkol SARIN and Wataru KAMEYAMA, "Classification and Quality Assessment of High Quality Digital Photographs," in Proceedings of ACM Multimedia 2009, pp. 1137-1138, Beijing, China.

IC4 - Supheakmongkol SARIN and Wataru KAMEYAMA, "Joint Equal Contribution of Global and Local Features for Image Annotation," in Proceedings of CLEF 2009 Workshop. Corfu, Greece, 2009.

IC5 - Supheakmongkol SARIN and Wataru KAMEYAMA, "Classifying High Quality Photographs by Creative Exposure Themes," in Proceedings of 3<sup>rd</sup>BCS-IRSG Symposium on Future Directions in Information Access (FDIA 09), pp. 128 - 130, Padua, Italy. Electronic Workshops in Computing (eWiC) Series.

IC6 - Supheakmongkol SARIN and Wataru KAMEYAMA, "Targeting Diversity in Photographic Retrieval Task with Commonsense Knowledge," in Proceeding of CLEF 2008 Workshop. ISSN: 1818-8044. ISBN: 2-912335-43-4, Aarhus, Denmark, 2008.

IC7 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "Exploiting Users' Personal and Public Information for Personal Photo Annotation," in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2007), pp. 564-567, Beijing, China, July 2007.

IC8 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "On Automatic Contextual Metadata Generation for Personal Digital Photographs," in Proceedings of the IEEE International Conference on Advanced Communication Technology (ICACT 2007), pp. 66-71, Phoenix, Korea, February 2007

## **National Conference/Workshop**

DC1 - Supheakmongkol SARIN and Wataru KAMEYAMA, "On Employing Content-aware Retargeted Image in Automatic Image Annotation," to appear in Proceedings of the IEICE General Conference. March 2012, Okayama, Japan.

DC2 - Supheakmongkol SARIN and Wataru KAMEYAMA, "Holistic Image Features Extraction for Better Image Annotation," in Proceedings of the IEICE General Conference, D-11-103. March 2010, Sendai City, Miyagi, Japan.

DC3 - Supheakmongkol SARIN and Wataru KAMEYAMA, "Raw Relationships in Aesthetic Photos by Optical Features," in Proceedings of the IEICE General Conference, D-13-38. March 2009, Ehime, Matsuyama, Japan.

DC4 - Supheakmongkol SARIN and Wataru KAMEYAMA, "Enabling Diversity in Image Retrieval Task using Clustering Approach with Commonsense Knowledge," in Proceedings of the 7<sup>th</sup> Forum of Information Technology (FIT2008), H-015. September 2008, Kanazawa, Japan.

DC5 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "Bridging Semantic Gap in Personal Digital Photo Annotation by Leveraging Personal and Public Information," 4<sup>th</sup> GITI/GITS Workshop, October 2007. Honjo, Japan (Poster Session)

DC6 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "Semi-automatic Annotation of Personal Digital Photographs with W6H2 Metadata," in Proceedings of the 6<sup>th</sup>Forum of Information Technology (FIT2007), 4N-5. September 2007, Nagoya, Japan.

DC7 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "Maximizing Context with Desktop Search Integration for Semi-Automatic Metadata Generation of Digital Photographs," in Proceedings to the 69<sup>th</sup>Information Processing Society of Japan Annual Conference (IPSJ 2007), 1D-7. March 2007, Tokyo, Japan.

DC8 - Tadashi MIYOSAWA, Toshinori NAGAHASHI, Supheakmongkol SARIN, and Wataru KAMEYAMA, "A Semi-automatic Metadata Generation for Electric Photo Album and Evaluation," in Proceedings of Institute of Television Engineer General Conference. December 2006.

DC9 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "SemiANNOTATE: A Semi-automatic Approach to Personal Photo Album Annotation using Public and Personal Information," in Proceedings of the 5<sup>th</sup>Forum of Information Technology (FIT2006), J-056. September 2006, Fukuoka, Japan.

DC10 - Supheakmongkol SARIN, Toshinori NAGAHASHI, Tadashi MIYOSAWA, and Wataru KAMEYAMA, "A Personal Digital Photo System with Built in Semi-automatic Annotation Engine," 3<sup>rd</sup>GITI/GITS Workshop, October 2006. Honjo, Japan. (Poster Session)

DC11 - Tadashi MIYOSAWA, Toshinori NAGAHASHI, Supheakmongkol SARIN, and Wataru KAMEYAMA, "Semi-automatic Metadata Generation for Electric Photo Album," in Proceedings of Institute of Television Engineer (ITE) Workshop. July 2006, Japan.

DC12 - Supheakmongkol SARIN, Toshinori Nagahashi, Miyosawa Tadashi, and Wataru Kameyama, "A Personal Photo Album with Semi-Automatic Metadata Generation," in Proceedings to the 68<sup>th</sup>Information Processing Society of Japan Annual Conference (IPJSJ 2006), 7D-6. March 2006, Tokyo, Japan.

## **Others**

O1 - Supheakmongkol SARIN, "Leveraging Salient Regions and Background for Automatic Image Annotation," in Presentation at Waseda-Hanyang University ICT PhD Academy 2011, November 2011, Hanyang University, Seoul, South Korea

O2 - Supheakmongkol SARIN, "Semantic Annotation of Digital Photographs," in Presentation at Waseda-Hanyang University ICT PhD Academy 2009, October 2009, Hanyang University, Seoul, South Korea

O3 - Supheakmongkol SARIN, "Semantic Annotation of Digital Photographs," in Presentation at ICT PhD Academy of Research Festival 2009, October 2009, Waseda University, Tokyo, Japan

O4 - Supheakmongkol SARIN, " Semi-Automatic Personal Digital Photograph Annotation using Personal and Public Information," Master's Thesis, September 2007, GITS, Waseda University



# Bibliography

- [1] EXIF Specification. <http://www.exif.org>, 2008.
- [2] Divisi: Commonsense reasoning over semantic networks. <http://csc.media.mit.edu/divisi>, 2009.
- [3] Photo.net. <http://www.photo.net>, 2009.
- [4] Facebook Photo Statistics. <http://blog.facebook.com/blog.php?post=206178097130>, 2010.
- [5] Flickr Photo Statistics. <http://blog.flickr.net/en/2010/09/19/5000000000>, 2010.
- [6] ACQUINE. <http://acquine.alipr.com/>, 2011.
- [7] Digital Photography Challenge. <http://www.dpchallenge.com>, 2011.
- [8] A.E. Abdel-Hakim and A.A. Farag. CSIFT: A SIFT descriptor with color invariant characteristics. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1978–1983. IEEE, 2006.
- [9] Acdsee pro. <http://www.acdsystems.com/>, 2008.
- [10] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned Salient Region Detection. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

- [11] Adobe photo album. <http://www.adobe.com>, 2008.
- [12] Aditya Vailaya Anil, Anil Jain, and Hong Jiang Zhang. On image classification: City images vs. landscapes. *Pattern Recognition*, 31:1921–1935, 1998.
- [13] Wang Ao-yu, Tang Min, and Dong Jin-xiang. A survey of silhouette detection techniques for non-photorealistic rendering. In *Proc. Third International Conference on Image and Graphics*, pages 434–437, 18–20 Dec. 2004.
- [14] Asahi English News. <http://www.asahi.com/english/>.
- [15] Vassilis Athitsos and Michael J. Swain. Distinguishing photographs and graphics on the world wide web. pages 10–17, 1997.
- [16] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.
- [17] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, April 2002.
- [18] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(4):712–727, 2008.
- [19] Daniel S. Butterfield, Caterina Fake, Callum James Henderson-Begg, and Serguei Mourachov. Interestingness ranking of media objects, US 2006/0242139 A1. US Patent.
- [20] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. 29(3):394–410, 2007.
- [21] Harr Chen and David R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, pages 429–436, 2006.

- [22] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*, 2002.
- [23] Florin Cutzu, Riad Hammoud, and Alex Leykin. Estimating the photorealism of images: Distinguishing paintings from photographs. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:305, 2003.
- [24] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik. Image quality assessment based on a degradation model. *Image Processing*, 9(4):636–650, 2000.
- [25] R. Datta, D. Joshi, J. Li, and J.Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the ECCV'06*, pages III: 288–301, 2006.
- [26] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.
- [27] Ritendra Datta, JianBing Li, and James Z Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, October 2008.
- [28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [29] Y. Deng, BS Manjunath, and H. Shin. Color image segmentation. In *cvpr*, page 2446. Published by the IEEE Computer Society, 1999.
- [30] Dublin Core Metadata Initiative. <http://dublincore.org/>, 2008.

- [31] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pages 97–112, London, UK, 2002. Springer-Verlag.
- [32] Exif and related resources. <http://www.exif.org>, 2008.
- [33] eXist XML Database. <http://exist.sourceforge.net>, 2008.
- [34] Christiane Fellbaum, editor. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May 1998.
- [35] Shaolei Feng, Raghavan Manmatha, and Victor Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR (2)*, pages 1002–1009, 2004.
- [36] A. Friedman. Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, 108:316–355, 1979.
- [37] Chen T. Gallagher, A.C. Using context to recognize people in consumer images. *IPSSJ Transactions on Computer Vision and Applications*, 1:115–126, 2009.
- [38] John F. Gantz, David Reinsel, Christopher Chute, Wolfgang Schlichting, John Mcarthur, Stephen Minton, Irida Xheneti, Anna Toncheva, and Alex Manfrediz. The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010. *IDC White Paper*, March 2007.
- [39] Google Desktop Search. <http://desktop.google.com/>, 2008.
- [40] GDS Java API. <http://sourceforge.net/projects/gdapi/>, 2007.
- [41] Geonames. <http://www.geonames.org/>, 2008.

- [42] Gi-Yeong Gim, Hyun-Chul Kim, Jin-Aeon Lee, and Whoi-Yul Kim. Subjective image-quality estimation based on psychophysical experimentation. *PSIVT*, pages 346–356, 2007.
- [43] Andreas Girgensohn, John Adcock, and Lynn Wilcox. Leveraging face recognition technology to find and organize photos. In *MIR '04: Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pages 99–106. ACM Press, 2004.
- [44] Google image search. <http://images.google.com>, 2010.
- [45] Google Image Labeler. <http://images.google.com/imagelabeler/>, 2008.
- [46] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1768–1783, 2006.
- [47] L. Grady and E.L. Schwartz. Isoperimetric graph partitioning for image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(3):469–475, 2006.
- [48] David Grangier and Samy Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1371–1384, August 2008.
- [49] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. 2005.
- [50] Michael Grubinger, Paul Clough, Henning Muller, and Thomas Deselaers. The iapr benchmark: A new evaluation resource for visual information systems. In *International Conference on Language Resources and Evaluation*, Genoa, Italy, May 2006.
- [51] Matthieu Guillaumin. *Exploiting Multimodal Data for Image Understanding*. PhD thesis, Université de Grenoble, sep 2010.

- [52] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *International Conference on Computer Vision*, sep 2009.
- [53] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack. Efficient color histogram indexing for quadratic from distance functions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(7):729–736, July 1995.
- [54] Catherine Havasi, Robert Speer, and Robert Speer. Conceptnet 3 : a flexible , multilingual semantic network for common sense knowledge. *Structure*, 11:33–38, 2007.
- [55] Tomer Hertz, Aharon Bar-Hillel, and Daphna Weinshall. Learning distance functions for image retrieval. In *Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition*, CVPR’04, pages 570–577, Washington, DC, USA, 2004. IEEE Computer Society.
- [56] Ross Hoddinott. *Digital Expoure Handbook*. Photographers’ Institute Press, 2008.
- [57] Xiaodi Hou and Liqing Zhang. Saliency Detection: A Spectral Residual Approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR ’07*, pages 1–8, 2007.
- [58] Ming-Hung Hsu and Hsin-Hsi Chen. Information retrieval with commonsense knowledge. In *SIGIR*, pages 651–652, 2006.
- [59] Httrack. <http://www.httrack.com/>, 2007.
- [60] Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In *MIR ’08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, NY, USA, 2008. ACM.

- [61] Laurent Itti, Christof Koch, and Ernst Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.
- [62] L.M. Jeon, V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Seventeenth Annual Conference on Neural Information Processing Systems (NIPS)*. MIT Press, 2003.
- [63] Yushi Jing and Shumeet Baluja. Pagerank for product image search. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 307–316, New York, NY, USA, 2008. ACM.
- [64] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2106–2113. IEEE, 2010.
- [65] Hyunmo Kang and Ben Shneiderman. Visualization methods for personal photo collections: Browsing and searching in the photofinder. In *IEEE International Conference on Multimedia and Expo*, 2000.
- [66] H. Katti, Kwok Yang Bin, Tat Seng Chua, and M. Kankanhalli. Pre-attentive discrimination of interestingness in images. In *Proc. IEEE International Conference on Multimedia and Expo*, pages 1433–1436, June 23 2008–April 26 2008.
- [67] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 419–426, 17–22 June 2006.
- [68] Naaman M. Kennedy, L. Generating diverse and representative image search results for landmarks. pages 297–306, 2008.

- [69] Naaman M. Ahern S. Nair R. Rattenbury T. Kennedy, L. How flickr helps us make sense of the world: Context and content in community-contributed media collections. pages 631–640, 2007.
- [70] Jin-Seo Kim, Maeng-Sub Cho, and Bon-Ki Koo. Experimental approach for human perception based image quality assessment. In *ICEC*, pages 59–68, 2006.
- [71] ByoungChul Ko, Hae-Sung Lee, and Hyeran Byun. Image retrieval using flexible image subblocks. In *Proceedings of the 2000 ACM symposium on Applied computing - Volume 2*, SAC '00, pages 574–578, New York, NY, USA, 2000. ACM.
- [72] William Ku, Mohan S. Kankanhalli, and Joo-Hwee Lim. Using camera settings templates ("scene modes") for image scene classification of photographs taken on manual/expert settings. In *PCM*, pages 10–17, 2007.
- [73] Jorma Laaksonen, Markus Koskela, and Erkki Oja. Content-based image retrieval using self-organizing maps. In *VISUAL*, pages 541–548, 1999.
- [74] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.
- [75] Douglas B. Lenat. Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, 38:33–38, November 1995.
- [76] Congcong Li and Tsuhan Chen. Aesthetic visual quality assessment of paintings. *IEEE Journal of Selected Topics in Signal Processing*, 3(2):236–252, April 2009.

- [77] Jia Li and James Z. Wang. Real-time computerized annotation of pictures. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 911–920, New York, NY, USA, 2006. ACM.
- [78] Henry Lieberman, Elizabeth Rozenweig, and Push Singh. Aria: An agent for annotating and retrieving images. *Computer*, 34(7):57–62, 2001.
- [79] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proc. International Conference on Image Processing 2002*, volume 1, pages I–900–I–903, 22–25 Sept. 2002.
- [80] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [81] Chang-Tien Lu, M. Shukla, S. H. Subramanya, and Yamin Wu. Performance evaluation of desktop search engines. In *Proc. IEEE Int. Conf. Information Reuse and Integration IRI 2007*, pages 110–115, 2007.
- [82] Apache Lucene. <http://lucene.apache.org>, 2008.
- [83] Jiebo Luo and Andreas E. Savakis. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. *ICIP*, pages 745–748, 2001.
- [84] Mathias Lux, Jutta Becker, and Harald Krottmaier. Semantic annotation and retrieval of digital photos. 2003.
- [85] Siwei Lyu and Hany Farid. How realistic is photorealistic? *IEEE Transactions on Signal Processing*, 53(2-2):845–850, 2005.
- [86] Mainichi News. <http://mdn.mainichi-msn.co.jp/>, 2008.
- [87] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for Image Annotation. *International Journal of Computer Vision*, pages 1–18, 2010.

- [88] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A New Baseline for Image Annotation. In *ECCV (3)*, pages 316–329, 2008.
- [89] Yelizaveta Marchenko, Tat-Seng Chua, and Ramesh Jain. Ontology-based annotation of paintings using transductive inference framework. *MMM*, pages 13–23, 2007.
- [90] Carlo Meghini, Fabrizio Sebastiani, and Umberto Straccia. A model of multimedia information retrieval. *J. ACM*, 48:909–970, September 2001.
- [91] G. Miller. WordNet: a Lexical Database for English. *Communications of the ACM*, Volume 38(Number 11), November 1995.
- [92] A. Mojsilovic, J. Kovacevic, J. Hu, R. J. Safranek, and S. K. Ganapathy. Matching and retrieval based on the vocabulary and grammar of color patterns. *IEEE Trans. on Image Processing*, 9(1):35–54, January 2000.
- [93] Florent Monay and Daniel Gatica-Perez. On image auto-annotation with latent space models. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 275–278, New York, NY, USA, 2003. ACM.
- [94] Florent Monay and Daniel Gatica-Perez. Plsa-based image auto-annotation: constraining the latent space. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 348–351, New York, NY, USA, 2004. ACM.
- [95] MPEG-7 Overview. <http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm>, 2008.
- [96] Mor Naaman, Hector Garcia-Molina, Andreas Paepcke, and Ron B. Yeh. Leveraging context to resolve identity in photo albums. In *Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 178–187, New York, NY, USA, 2005. ACM Press.

- [97] Mor Naaman, Andreas Paepcke, and Hector Garcia-Molina. From where to what: Metadata sharing for digital photographs with geographic coordinates. In *10th CoopIS*, 2003.
- [98] Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 Large Scale Visual Concept Detection and Annotation Task. In *CLEF working notes*, Corfu, Greece, 2009.
- [99] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
- [100] Andrew Payne and Sameer Singh. Indoor vs. outdoor scene classification in digital photographs. *Pattern Recognition*, 38(10):1533 – 1545, 2005.
- [101] G. Peters. Aesthetic primitives of images for visualization. In *Proc. 11th International Conference Information Visualization IV '07*, pages 316–325, 4–6 July 2007.
- [102] Bryan Peterson. *Understanding Exposure [Revised Edition]*. AMPHOTO Book, 2004.
- [103] Picasa. <http://www.picasa.com>, 2008.
- [104] M. C Potter. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory* 2, pages 509–522, 1976.
- [105] Y. Rubner, L. J. Guibas, and C. Tomasi. The earth mover’s distance, multi-dimensional scaling, and color-based image retrieval. in *Proc. DARPA Image Understanding Workshop*, pages 661–668, May 1997.
- [106] P. Salembier and J. Smith. *Overview of Multimedia Description Schemes and Schema Tools*. Addison-Wesley, 2001.

- [107] Risto Sarvas, Erick Herrarte, Anita Wilhelm, and Marc Davis. Metadata creation system for mobile images. In *Proceedings of the 2nd MobiSys*, pages 36–48. ACM Press, 2004.
- [108] Andreas Savakis. A computationally efficient approach to indoor/outdoor scene classification. *ICPR '02: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02)*, 2002.
- [109] Raimondo Schettini, Gianluigi CIOCCA, Silvia Zuffi, Istituto Tecnologie, and Infomatiche Multimediali. A survey of methods for colour image indexing and retrieval in image databases. In *In Color Imaging Science: Exploiting Digital*, pages 9–1. Media, John Wiley, 2001.
- [110] Google Desktop Search - Search Across Computer. <http://desktop.google.com/features.html>, 2008.
- [111] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.
- [112] Sundaram H. Xie L. Shevade, B. Modeling personal and social network context for event annotation in images. In *Proceedings of the ACM International Conference on Digital Libraries*, pages 127–134, 2007.
- [113] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2002.
- [114] Mei-Ling Shyu, Shu-Ching Chen, Min Chen, Chengcui Zhang, and Kanoksri Sarinnapakorn. Image database retrieval utilizing affinity relationships. In *Proceedings of the 1st ACM international workshop on Multimedia databases, MMDB '03*, pages 78–85, New York, NY, USA, 2003. ACM.
- [115] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open mind common sense: Knowledge acquisition from the

- general public. In *On the Move to Meaningful Internet Systems, 2002 - DOA/CoopIS/ODBASE 2002 Confederated International Conferences DOA, CoopIS and ODBASE 2002*, pages 1223–1237, London, UK, UK, 2002. Springer-Verlag.
- [116] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.
- [117] Robert Speer, Catherine Havasi, and Henry Lieberman. Analogyspace: Reducing the dimensionality of common sense knowledge. In Dieter Fox and Carla P. Gomes, editors, *AAAI*, pages 548–553. AAAI Press, 2008.
- [118] Pao Sriprasertsuk and Wataru Kameyama. Information distribution analysis based on human’s behavior state model and the small-world network. *IEICE Transactions*, pages 608–619, 2009.
- [119] Pao Sriprasertsuk, Akiko Seki, and Wataru Kameyama. On content distribution model and analyzing distribution effectiveness. *IPSJ Digital Courier*, 3:492–505, 2007.
- [120] D.S. Stijin, T. Kentaro, K. Jun’ichi, O. Kiyonori, V. Isrvan, and Y. Yulan. The nict concept dictionary. In *Universal Communication Symposium (IUCS), 2010 4th International*, page 403, oct. 2010.
- [121] Yanfeng Sun, Hongjiang Zhang, Lei Zhang, and Mingjing Li. Myphotos: a system for home photo management and processing. In *MULTIMEDIA ’02: Proceedings of the tenth ACM international conference on Multimedia*, pages 81–82. ACM Press, 2002.
- [122] M. J. Swain and B. H. Ballard. Color indexing. *Int. J. Comput. Vis.*, 7(1):11–32, 1991.

- [123] Richard Szeliski. Computer vision : Algorithms and applications. *Computer*, 5:832, 2010.
- [124] Tele Tan, Jiayi Chen, Philippe Mulhem, and Mohan Kankanhalli. Smartalbum: a multi-modal photo annotation system. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, pages 87–88, New York, NY, USA, 2002. ACM Press.
- [125] Hanghang Tong, Mingjing Li, HongJiang Zhang, Jingrui He, and Changshui Zhang. Classification of digital photos taken by photographers or home users. In *PCM (1)*, pages 198–205, 2004.
- [126] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1958–1970, November 2008.
- [127] Antonio Torralba and Aude Oliva. Depth estimation from image structure. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1226–1238, 2002.
- [128] Chih-Fong Tsai, Ken McGarry, and John Tait. Image classification using hybrid neural networks. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, SIGIR '03*, pages 431–432, New York, NY, USA, 2003. ACM.
- [129] Mischa Tuffield, Stephen Harris, David P. Dupplaw, Ajay Chakravarthy, Christopher Brewster, Nicholas Gibbins, Kieron O’Hara, Fabio Ciravegna, Derek Sleeman, Yorick Wilks, and Nigel R. Shadbolt. Image annotation with photocopain. In *Proceedings of the First International Workshop on Semantic Web Annotations for Multimedia (SWAMM)*, May, 2006.
- [130] Benjamin Turnbull, Barry Blundell, and Jill Slay. Google desktop as a source of digital evidence. *IJDE*, 5(1), 2006.

- [131] K.E.A. Van De Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.
- [132] J. Van de Weijer, T. Gevers, and A.D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 150–156, 2006.
- [133] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM.
- [134] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. 2003.
- [135] Christian Wallraven, Roland Fleming, Douglas Cunningham, Jaume Rigau, Miquel Feixas, and Mateu Sbert. Categorizing art: Comparing humans and computers. *Computers & Graphics*, In Press, 2009.
- [136] Xin-Jing Wang, Lei Zhang, Feng Jing, and Wei-Ying Ma. Annosearch: Image auto-annotation by search. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1483–1490, Washington, DC, USA, 2006. IEEE Computer Society.
- [137] Liu Wenyin, Yanfeng Sun, and Hongjiang Zhang. Mialbum - a system for home photo managemet using the semi-automatic image annotation approach. In *MULTIMEDIA '00: Proceedings of the eighth ACM international conference on Multimedia*, pages 479–480, New York, NY, USA, 2000. ACM Press.
- [138] Wikipedia. <http://en.wikipedia.org/>, 2008.
- [139] J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. *illumination*, 17:21.

- [140] M. Yelizaveta, C. Tat-Seng, and A. Irina. Analysis and retrieval of paintings using artistic color concepts. In *Proc. IEEE International Conference on Multimedia and Expo ICME 2005*, pages 1246–1249, 2005.
- [141] C.T. Zahn. Graph-theoretical methods for detecting and describing gestalt clusters. *Computers, IEEE Transactions on*, 100(1):68–86, 2006.
- [142] Cheng Xiang Zhai, William W. Cohen, and John Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '03, pages 10–17, New York, NY, USA, 2003. ACM.
- [143] R. Zhao and W. Grosky. From features to semantics: Some preliminary results. page TAS3, 2000.
- [144] Xiangdong Zhou, Mei Wang, Qi Zhang, Junqi Zhang, and Baile Shi. Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 25–32, New York, NY, USA, 2007. ACM Press.
- [145] Yahoo! Research Berkeley, ZoneTag, 2007. <http://zonetag.research.yahoo.com/>.
- [146] Sundaram H. Xie L. Zunjarwad, A. Contextual wisdom: Social relations and correlations for multimedia event annotation. In *Proceedings of the ACM International Multimedia Conference and Exhibition*, pages 615–624, 2007.