



HOKKAIDO UNIVERSITY

Title	Extracting Semantic Relations Between Concepts Based on Textual and Visual Features of Tagged Images
Author(s)	桂井, 麻里衣
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11492号
Issue Date	2014-06-30
DOI	https://doi.org/10.14943/doctoral.k11492
Doc URL	https://hdl.handle.net/2115/59430
Type	doctoral thesis
File Information	Marie_Katsurai.pdf



Extracting Semantic Relations Between Concepts Based on Textual and Visual Features of Tagged Images



Marie Katsurai

Graduate School of Information Science and Technology
Hokkaido University

This dissertation is submitted for the degree of
Doctor of Philosophy in Information Science and Technology

June 2014

Acknowledgements

Undertaking this PhD research has been a truly valuable experience for me and it would not have been possible to complete without the support and guidance that I received from others.

To begin, I would like to express my deepest gratitude to my dissertation advisor, Prof. Miki Haseyama, for her outstanding guidance, care, and patience in conducting research. She provided me with a lot of valuable experiences including participating in conferences and performing several research-related activities. I appreciate from the bottom of my heart that she taught me the importance of having a clear vision and a deep sense of purpose.

Great thanks to my co-advisors, Prof. Kenji Araki and Prof. Tsuyoshi Yamamoto, for their encouragement and comments, which were invaluable and constructive in improving my thesis. Also, thanks to Prof. Yasutaka Ogawa, Prof. Toshio Nojima, Prof. Yoshikazu Miyanaga, Prof. Kunimasa Saito, Prof. Takeo Ogane, Prof. Yoshinori Dobashi, Prof. Hideyuki Imai, and Prof. Hajime Igarashi, who served on the thesis committee and provided many useful comments on my thesis.

In addition, I thank Prof. Masanori Koshiba, whose impressive insight and attitude towards research inspired me greatly. I also wish to offer my thanks to Prof. Mineichi Kudo, who gave me multiple opportunities to get to know and collaborate with many people.

I would like to thank Dr. Takahiro Ogawa, whose meticulous comments on my papers were an enormous help to me. I also wish to offer my thanks to all of my colleagues at the Laboratory of Media Dynamics for their care and help. In particular, I would like to thank Michael Penkov, who provided useful comments on the technical aspects of conducting experiments. I also owe gratitude to all of my friends pursuing doctoral degrees in the Graduate School of Information Science and Technology for their valuable discussions and for making our time spent together enjoyable.

During my PhD study, I also enjoyed the interactions I had with many talented researchers and students. This thesis has greatly benefited from their advice and sugges-

tions when participating in conferences. My research motivation has also been quite influenced by discussing and exchanging opinions with them through online social media websites. It made me appreciate the success and benefits of Internet technologies, and in addition, I deeply believe that being able to effectively use them in my PhD life was really positive for my work.

I am particularly thankful to Prof. Toshihisa Tanaka at Tokyo University of Agriculture and Technology, who gave me, since I was a master's student, many opportunities to communicate with researchers and students. He also provided invaluable comments, help, and encouragement during my PhD study in numerous ways.

Furthermore, I gratefully acknowledge the financial support from the Japan Society for the Promotion of Science (JSPS) that I have received since April 2013. It has supported my participation in conferences and provided an excellent research environment.

I would like to give a heartfelt thank you to my parents and younger brother for always being so supportive and encouraging me to follow my dream.

Finally, I would like to especially thank Dr. Yuhei Ishizaka at Laboratory of Information Communication Photonics, who has been hugely supportive throughout my eight years at Hokkaido University. Long discussions with him have significantly inspired many research directions and expanded my horizons. It was indeed a great pleasure to have spent many enjoyable times with him. Thank you with all my heart.

Abstract

This dissertation is about extracting semantic relations between concepts based on the textual and visual features of tagged images.

The semantic relations between concepts hold essential information for improving the performance of the computational techniques in artificial intelligence, including information retrieval and automatic annotation. For the purpose of facilitating these various applications, knowledge bases for computers have been manually constructed by experts, and many methods have been used to quantify the inter-concept relations. However, the manual creation of lexical ontologies requires considerable time and effort, and consequently, the number of concepts in the ontologies is much smaller than the number of concepts that are used on the Web.

To automatically determine semantic relations without pre-defined ontologies, several methods have focused on the utilization of data on the Web as an information source. Social media websites such as Flickr enable us to enjoy sharing multimedia data taken with own digital cameras or smartphones, which accelerates the increase of volume on the Web. On these websites, users can freely describe updated content using a set of concepts commonly known as tags. This is recognized as one of most important features in the Web 2.0 era. Using a tag-based image search, we can easily obtain a multimedia data collection that describes a target concept. Recent methods benefit from this availability and have attempted to extract inter-concept relations based on tagged images. Although the tagged images provide us with rich information such as pairs of text documents and their corresponding visual information, conventional methods cannot deal with different types of modalities simultaneously; they use either textual or visual features. This dissertation seeks to investigate whether exploiting both of these modalities is effective for measuring inter-concept relations.

Throughout this thesis, I have presented approaches to extract semantic relations between concepts based on textual and visual features of tagged images. First, I have presented a classifier-based approach for quantifying the semantic similarity between concepts, assuming the use of an accurately labeled image collection. This approach aims

to investigate the effectiveness of the collaborative use of both the textual and visual features. Second, I have presented a cross-modal approach that enables using the two features of tagged images on the Web for extracting the inter-concept relations. Specifically, the proposed method projects the two features to the same space and then quantifies the following information: the semantic relatedness between concepts, and the abstraction levels of concepts. Experiments conducted on tagged images collected from a photo sharing website show that the proposed approach is more effective in order to capture the inter-concept relations as compared with conventional methods that exploit single modality. Finally, I have presented an analysis of the temporal dynamics of tagged images for a target concept in order to extract new information about concept characteristics. Experiments based on tagged images associated with time stamped information show that the presented analysis provides a new idea for the future generation of multimedia data mining.

Contents

Acknowledgements	iii
Abstract	v
Contents	vii
List of Figures	xi
List of Tables	xv
1 Introduction	1
1.1 Background	1
1.2 Thesis Statement	3
1.3 Terminology	4
1.4 Structure of Thesis	4
2 Literature Review	7
2.1 Methods based on external knowledge bases	8
2.1.1 WordNet-based measures	9
2.1.2 Wikipedia-based measures	13
2.1.3 Drawback of the measures using knowledge bases	14
2.2 Methods based on Web text documents	14
2.2.1 Web search engine-based measures	14
2.2.2 Probabilistic topic model-based measures	15
2.2.3 Measures on social tags	17
2.2.4 Drawback of text-based measures	17
2.3 Methods based on visual features	18
2.3.1 Measures that focus on the semantic gap	18
2.3.2 Measures based on visual similarity	20

2.3.3	Drawbacks of visual feature-based measures	20
2.4	Applications of semantic measures to multimedia related-tasks	21
2.4.1	Multimedia annotation	21
2.4.2	Multimedia retrieval	25
2.4.3	Tag recommendation	25
3	A Classifier-Based Approach for Extracting Inter-Concept Relationships from An Image Database	27
3.1	Introduction	27
3.2	Feature Selection Using Classifiers	29
3.2.1	Training logistic regression	29
3.2.2	Feature selection using logistic regression	30
3.3	Extracting Inter-Concept Relationships Based on Shared Visual Features . .	31
3.3.1	Removal of relevant features	31
3.3.2	Determination of Relevance Presence Between concepts	33
3.3.3	Measurement of Relevance Between Concepts	33
3.3.4	Discussion of classifiers	34
3.4	Experimental Results	34
3.4.1	Visualizing Concept Relationships	37
3.4.2	User evaluation of extracted relationships	39
3.4.3	Application to automatic image annotation	42
3.5	Conclusion and Future Work	45
4	A Cross-Modal Approach for Extracting Inter-Concept Relationships From Tagged Images	51
4.1	Introduction	51
4.2	Canonical Correlation Analysis and Its Probabilistic Interpretation	52
4.2.1	An overview of canonical correlation analysis	53
4.2.2	Probabilistic interpretation of canonical correlation analysis	53
4.3	A CCA-based Approach for Extracting Semantic Relations Between Concepts	55
4.3.1	Calculation of the representative distribution of a concept	57
4.3.2	Measures of semantic relationships	59
4.3.3	Computational complexity	62
4.4	Experiments	63
4.4.1	Experimental settings	63
4.4.2	Visualizing Concept Relations	65

4.4.3	Evaluating the measures of semantic relatedness	66
4.4.4	Evaluation of measures on abstraction levels	75
4.5	Conclusion	78
5	Exploiting Temporal Information for Finding Tag Relationship Changes	89
5.1	Introduction	89
5.2	Related Work	90
5.2.1	Temporal dynamics of text documents	90
5.2.2	Temporal dynamics of Web image collections	91
5.3	Constructing Gaussian Mixture Models for Tags	92
5.4	Finding Time-varying Tags Based on Visual Features	93
5.4.1	Recursive update of GMM parameters	94
5.4.2	Extracting semantic relationships between tags based on the updated GMMs	95
5.5	Experiments	96
5.5.1	Dataset	97
5.5.2	Analyzing a tag's visual changes over time	97
5.5.3	Finding tag relationship changes	98
5.6	Conclusion and Future Work	103
6	Conclusions and Discussion	105
6.1	Key Observations and Contributions	105
6.2	Future Directions	106
	References	111
	List of Publications	123

List of Figures

1.1	An example of tagged images on the Web.	2
2.1	An illustration of different methods for extracting semantic relationships between concepts. Conventional methods can be divided on the basis of the source of information that they exploit: external knowledge bases, text documents on the Web and visual features of images.	8
2.2	An excerpt of noun relations defined in WordNet [1].	9
2.3	Graphical model representation of probabilistic topic models.	15
2.4	An example of calculating semantic relatedness between concepts based on visual features.	19
2.5	An illustration of general framework that introduces semantic relationships between concepts for image annotation refinement.	21
2.6	An example of tag recommendation that exploits semantic relatedness between concepts.	25
3.1	An illustration of the proposed approach. Given a training image, each of its labels is first associated with visual features of the image, and then the degree of sharing visual features among labels is measured as the visual feature-based relevance.	28
3.2	Examples of visual feature removal relevant to “bloom”: (a)-(d) Original images A-D, (e)-(h) Images in which visual features relevant to “bloom” are removed using Eq. (3.4) from Original images A, B, C, and D.	32
3.3	Concept relationships extracted from IAPR-TC12 dataset by proposed method and comparative methods A, B, and C.	38
3.4	Concept relationships extracted from Hokkaido landscape dataset by proposed method and comparative method A, B, and C.	40

3.5	Image annotation results of proposed method and comparative method A for the IAPR-TC12 dataset. The top seven concepts are provided.	43
3.6	Image annotation results of proposed method and comparative method A for the Hokkaido landscape database. The top seven keywords are provided.	48
3.7	Average precision ($m = 1, 2, 3, 4, 5$, and 10) on testing images of the IAPR-TC12 dataset.	49
3.8	Average precision ($m = 1, 3, 5, 7, 10$, and 20) on testing images of the Hokkaido landscape dataset.	49
4.1	A graphical model for canonical correlation analysis [2].	54
4.2	The proposed steps that are taken for measuring concept relationships: (a) feature extraction from tagged images, (b) sample projection using CCA, (c) calculation of representative distributions in the latent space, and (d) semantic relationship extraction.	56
4.3	An illustration of Hypothesis 1 (semantic relatedness), which represents how semantically similar two concepts are. This measure is based on the distance between their representative distributions.	60
4.4	An illustration of Hypothesis 2 (abstraction levels of concepts), which represents whether a target concept subsumes several semantics or appears in a certain context. This measure is based on the entropy of the concept's representative distribution.	61
4.5	The frequency distribution of tags in the MIRFLICKR-1M dataset.	64
4.6	The number of tags per image in the MIRFLICKR-1M dataset.	65
4.7	Concept network constructed by the proposed method based on the dataset. Each node represents a concept and each edge between concepts represents the strong semantic relatedness. Node color indicates the abstraction levels, which are set according to the color bar in the right side.	66
4.8	Concept networks whose edges are calculated using the proposed method (P-H) and the conventional methods [3, 4]. Each node represents a concept, and the edges between the concepts represent the strongest 60 relations.	68
4.9	Average user scores for the most relevant concept pairs extracted using different methods. A high score means that the method can accurately extract the top relevant concept pairs.	69

4.10 Precision of clustering concepts using different measures of semantic relatedness: (a) results for the first clustering experiment (strongly correlated concepts), where text-based measures (TCD and PText) obtained high precisions, (b) results for the second clustering experiment (weakly correlated concepts), where visual-based measures (GMM and PVis) performed better than text-based measures.	70
4.11 Relative improvement of image annotations with different semantic measures. The baseline MAP was computed based on initial annotations. A high MAP gain means that the method refined annotations better.	72
4.12 The computation time for each procedure when changing q in the proposed method and for each procedure in the conventional methods [4, 5]: (a) the computation time for the calculation of the representative distributions of 850 concepts, (b) the computation time for the measurement of the pairwise relatedness among 850 concepts.	84
4.13 Performance comparison in different experiments when changing q in the proposed method (P-H): (a) results for the first clustering experiment (strongly correlated concepts), (b) results for the second clustering experiment (weakly correlated concepts), and (c) results for image annotation.	85
4.14 Precision of comparing the abstraction levels between two concepts, where the ground truth is determined using the WordNet hierarchy.	85
4.15 Performance on measuring the abstraction level. A high $P@m$ means that a method accurately locates specific concepts in the bottom m concepts in the sort of the scores.	88
4.16 Performance comparison in WordNet-based evaluation of the abstraction levels between two concepts when changing q in the proposed method. . .	88
5.1 An overview of the proposed method.	94
5.2 Results obtained using the first measure for tags “apple” and “sunset”, in which their example images at each time period are shown.	98
5.3 Concept networks around “blackberry,” whose edges are calculated using the proposed method and the comparative method. Each node represents a tag, and the edges between the concepts represent the strongest relationships.	100

- 5.4 Concept networks around “jaguar,” whose edges are calculated using the proposed method and the comparative method. Each node represents a concept, and the edges between the concepts represent the strongest relationships. 101
- 5.5 Concept networks around “jaguar,” whose edges are calculated using the proposed method and the comparative method. Each node represents a concept, and the edges between the concepts represent the strongest relations. 102

List of Tables

3.1	Examples of images and their labels in IAPR TC-12 dataset [6].	35
3.2	Examples of images and their labels in Hokkaido landscape dataset.	35
3.3	User assessments of the IAPR-TC12 dataset.	41
3.4	User assessments of the Hokkaido landscape database.	41
3.5	Precision, recall, and F-measure of the conventional methods [3, 4, 7, 8] and the proposed method for the IAPR-TC12 dataset, in which $m = 5$	44
3.6	Precision, recall, and F-measure of the conventional methods [3, 4, 7, 8] and the proposed method for the Hokkaido landscape database, in which $m = 12$	45
4.1	The computational complexities of the proposed method (Hellinger distance) and the conventional methods [4, 5] for each procedure.	62
4.2	Statistics of the MIRFLICKR-1M dataset [9].	63
4.3	Examples of subjective evaluation for the most relevant pairs extracted using different methods. Each of the correlation scores are normalized from 1 (“semantically unrelated”) to 5 (“highly correlated”). The correlation scores that are closest to the user scores are highlighted in bold.	80
4.4	Examples of the categories and concepts used for the first clustering experiment. Concepts that belong to the same category correlate strongly with one another.	81
4.5	Examples of the categories and concepts used for the second clustering experiment. Concepts that belong to the same category do not explain one another compared with the strongly correlated concepts used in the first clustering experiment.	81
4.6	Examples of testing images and automatic annotation using four different measures, where the top seven annotations are provided to each image. The correct annotations are highlighted in bold.	82

4.7	Evaluation results for tag recommendations with different semantic measures. Each column shows the corresponding measure's performance. The best result in each row is marked in bold.	83
4.8	Results of abstraction level estimation for each of the concepts, in which the top 15 (abstract) and bottom 15 (specific) concepts are shown. Concepts that represent objects, scenes, or activities are highlighted in bold for reference.	86
4.9	Some of the 1,082 concept pairs used for the objective evaluation of abstraction levels. In each of the pairs, the right concept is a hypernym of the left concept.	87
4.10	Category labels of concepts and examples in each category.	87
5.1	Tags and their numbers of images in the dataset.	97

Chapter 1

Introduction

1.1 Background

With the rapid advancement of Internet technologies, the popularity of image capturing devices, and the success of Web 2.0 applications, we have seen a dramatic increase in our ability to collect digital information in various situations and share them on the Web. According to an IDC survey [10], the amount of digital information created and replicated throughout the world was about 0.8ZB (0.8 trillion GB) in 2009, and it will increase to almost 40ZB by 2020. This means the data will just about double every two years from now until 2020. With the goal of efficiently managing quite a large and complex collection of data, several computational techniques in the research field of artificial intelligence have been presented, such as information retrieval, automatic annotation, and recommendation. A common fundamental feature of these techniques is to carry out their tasks on the basis of the knowledge about the world, i.e., rich and complex structures of concepts that humans have formed naturally [11]. To establish knowledge bases for computers, lexical ontologies such as WordNet [1] have been manually constructed by experts. Many methods have used them to quantify the inter-concept semantic relations, for example, semantic relatedness between concepts and abstraction levels of concepts. However, the manual creation and maintenance of lexical ontologies requires considerable time and effort, which results in the following problems: (i) the number of concepts defined in the ontologies is much smaller than the number of concepts found in our daily lives; and (ii) given a new concept, the existing inter-concept relations cannot be easily updated.

To automatically determine inter-concept relations without pre-defined knowledge bases, several methods have exploited a collection of text documents on the Web as an underlying information resource. Most of them utilize statistical information regard-

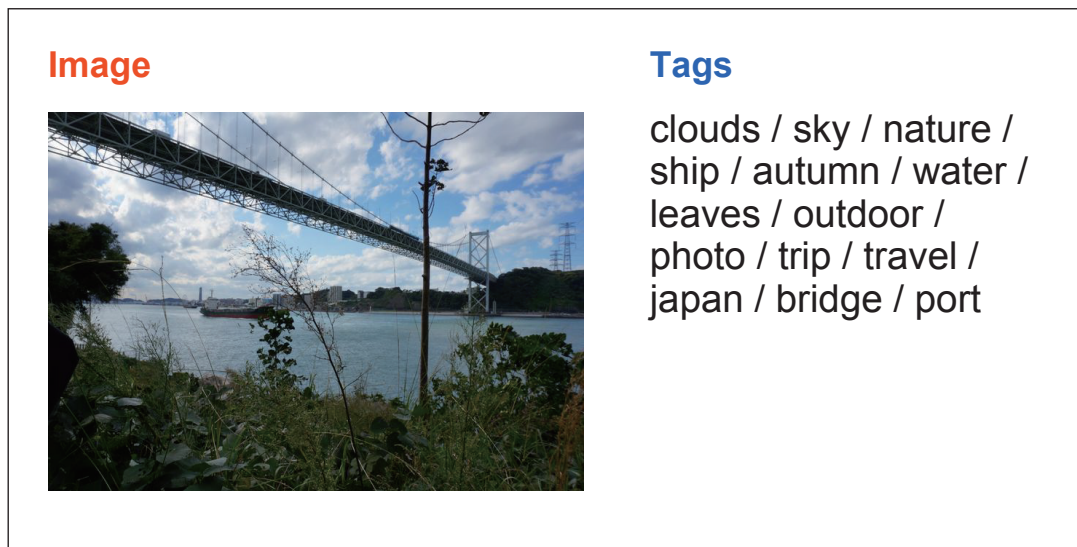


Fig. 1.1 An example of tagged images on the Web.

ing the use of concepts in a collection of text documents returned from a Web search engine [12–15]. Some methods assume that a pair of concepts that frequently co-occur in the same documents are semantically correlated, while others consider a topic representation of each concept and then calculate the similarity between concepts.

The data for analyzing inter-concept relations recently have changed from Web text documents to user-generated multimedia contents. As initially mentioned in this section, due to an increase of the transmission capacity of networks and the widespread use of photo-taking devices, communication on the Web is no longer limited to solely text document-based - we can record individual and collective experiences in our daily lives as digital images anytime and anywhere, and share them with other people through social media websites. For example, Flickr¹, a popular photo sharing website, hosted over six billion photos on August 2011, and uploads have been increasing 20% per year [16]. As a Web 2.0 characteristic, Flickr provides functions that allow users to annotate their uploaded image with a set of concepts (called tags), as shown in Fig. 1.1. These tags play an important role on social media websites as a textual description of the data. For example, the image in Fig. 1.1 will appear in a search result whose query keyword is “clouds”, “sky”, or “ship”. These photos, returned by inputting a concept as a query, can be valuable information that conveys the visual aspects of that concept. By using tag-based image retrieval in practice, it has now become easy to obtain a very large number of visual features of concepts. Recent methods benefit from this availability, and have

¹<https://www.flickr.com/>

attempted to extract concept relationships by evaluating the distribution of concepts in a visual feature space [4, 5].

Although tagged images on social media websites can provide rich information, the above-mentioned conventional methods can deal with only a single modality, i.e., they use either textual or visual features. As a result, there could be a lack of information since methods that use only textual features cannot capture visual similarity [4], while methods that use only visual features cannot effectively show the properties of abstract concepts due to the well-known semantic gap [17].

1.2 Thesis Statement

Considering the above-mentioned drawbacks of the conventional methods, I argue in this thesis that both of these modalities are essential to capturing the semantics of the concepts and that they act in a complementary way. If not single but multiple types of modalities are used effectively, we can extract inter-concept relations that are coherent with human cognition, which is the motivation behind this dissertation.

In this thesis, I present approaches for extracting semantic relations between concepts, which enable the collaborative use of textual and visual features from tagged images. First, I present a framework for quantifying the semantic similarity between two concepts using the two types of features of images, in which the images are accurately labeled. In the proposed method, we select visual features that are relevant to each concept through classifier training. Because just comparing the relevant features among concepts is equivalent to the utilization of visual features only, we reference a list of labels associated to an image and measure a degree of sharing visual features between two concepts that co-occur in the same images, as this corresponds to the utilization of textual features. The calculated degree is used to measure the semantic similarity between concepts. Experiments conducted on image collections show that exploiting the two features is effective for extracting inter-concept relations. Second, I present a cross-modal approach for extracting semantic inter-concept relations from tagged images on a social media website. To use different types of features simultaneously, we consider a latent space that connects to two feature spaces. In the latent space, we model a concept as a probabilistic distribution parameterized by a mean and a covariance matrix. Based on the distribution estimated for each concept, we present two types of semantic measures: semantic relatedness between concepts and abstraction levels of concepts. Third, I discuss in this thesis the benefits of using timestamps of the tagged images for extracting

knowledge about concepts. Experiments conducted on a large-scale image dataset collected from Flickr show that exploiting time information associated to image data has the possibility of extracting new information about concepts.

1.3 Terminology

In the literature related to quantifying the relevance between concepts, at least three different terms are used [18–22]: **semantic relatedness**, **semantic similarity**, and **semantic distance**. A measure of semantic similarity quantifies how much these two concepts are alike in terms of hierarchical *hyponymy/hypernymy* relations. Semantic relatedness is a more general notion than semantic similarity, and indicates how much two concepts are related in a taxonomy by using all relations between them (i.e., hyponymy/hypernymy, meronymy, co-occurrence, and any kind of functional relations including *has-part*, *is-made-of*, *is-an-attribute-of*). The term semantic distance has been used to refer to an inverse of semantic relatedness or semantic similarity. That is, concepts that are semantically similar or related are considered to be semantically close to each other. In this thesis, the first approach introduced in Chapter 3 instead focuses on hyponymy/hypernymy and meronymy relations, while the second approach presented in Chapter 4 calculates the semantic relatedness. Through this thesis, several distance measures are alternatively exploited to derive similarity/relatedness measures.

Inter-concept relations are not only horizontal but also vertical, and among concepts, the granularities of semantics are different from each other [23]. Chapter 4 also presents a measure of **abstraction level** of a concept, which can be called a granularity level of the concept. For example, if a hierarchical structure like an ontology is constructed, this represents how close a target concept is located near the root. Thus, a broad concept that is comprised of several concepts has a high abstraction level.

1.4 Structure of Thesis

The structure of this thesis is as follows: The background and objective of the thesis is given in this chapter. Chapter 2 provides a review of the related studies, which includes conventional methods that use different knowledge sources for extracting semantic relations between concepts and their applications to multimedia-related tasks. In Chapter 3, we present a classifier-based method for measuring semantic similarity between concepts, assuming that a given dataset is accurately labeled by a set of concepts. In Chapter

4, we present a cross-modal approach for extracting semantic relations between concepts using tagged images on a photo sharing website. In Chapter 5, we exploit timestamps of tagged images as additional features for extracting new knowledge about concepts. Finally, this thesis is summarized, and some possible directions for future work are suggested in Chapter 6.

Chapter 2

Literature Review

Acquiring knowledge on concepts and their relations has a long history in research fields including artificial intelligence and biomedicine. Many methods have been presented to quantify the semantic relationships between concepts. The conventional methods can be roughly divided into three categories on the basis of the source of information they exploit: external knowledge bases, text documents on the Web, and visual features of Flickr images. The three types of conventional methods and the proposed method are summarized in Fig. 2.1. In this chapter, we provide a brief literature review of these methods and indicate their drawbacks, which should be resolved. Compared with conventional information sources, we believe that tagged images, which are outcomes of the development of Web2.0 technologies, can be a new source for knowledge extraction. Our approaches presented in this thesis provide solutions to effectively use the features that can be acquired from tagged images on the Web. Furthermore, as shown in Fig. 2.1, the inter-concept relations have been introduced to several multimedia-related applications; thus, their conventional methods should also be introduced as related work.

The remainder of this chapter is organized in the following manner. Section 2.1 describes measures based on external knowledge bases such as WordNet and Wikipedia. Section 2.2 explains measures that exploit statistics in text documents on the Web, and Section 2.3 introduces measures that focus on distributions of concepts in a visual feature space. Furthermore, conventional methods in multimedia-related tasks, which exploit semantic relationships between concepts to improve their performance, are described in Section 2.4.

Objective: Extraction of semantic relations between concepts

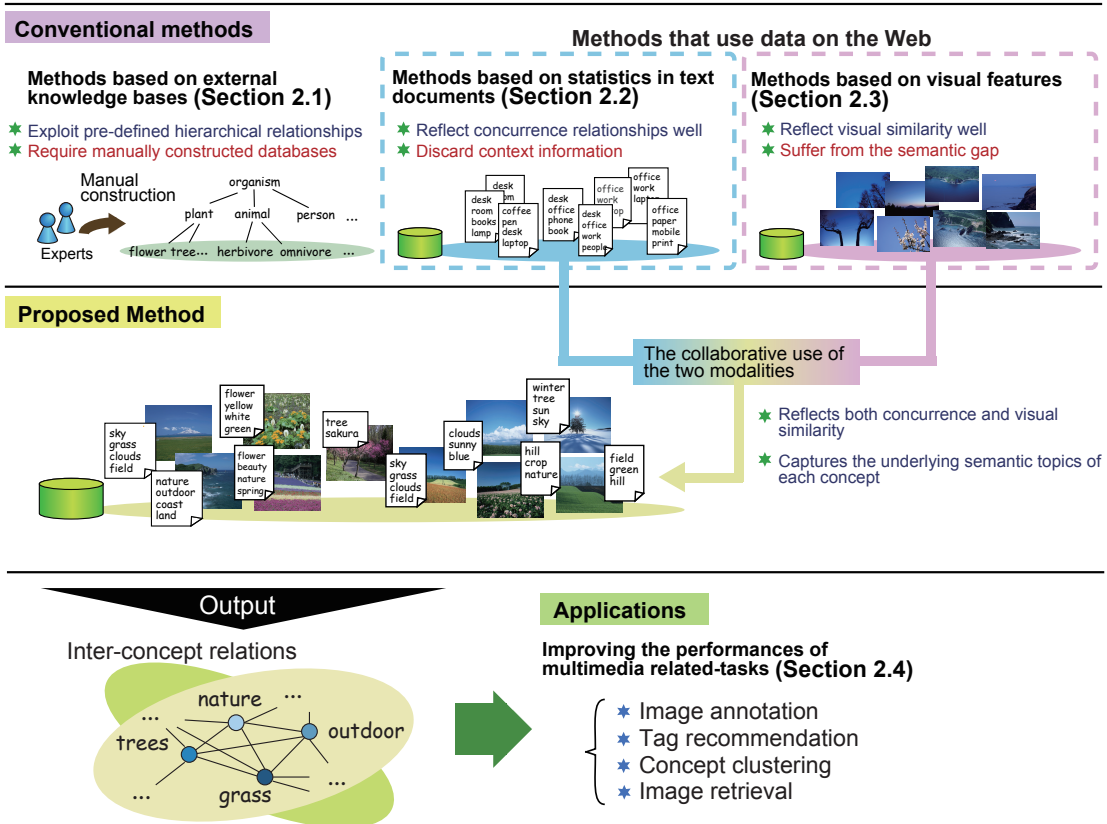


Fig. 2.1 An illustration of different methods for extracting semantic relationships between concepts. Conventional methods can be divided on the basis of the source of information that they exploit: external knowledge bases, text documents on the Web and visual features of images.

2.1 Methods based on external knowledge bases

Human-created knowledge bases including dictionaries, thesauri, and encyclopedic resources have played an important role in providing computational techniques with the semantic knowledge on concepts. These knowledge bases usually form a graph of concepts, and concepts are connected by edges or links that represent certain type of semantic relations. In this section, I describe how conventional methods are used to extract semantic relations between concepts from representative knowledge bases WordNet and Wikipedia.

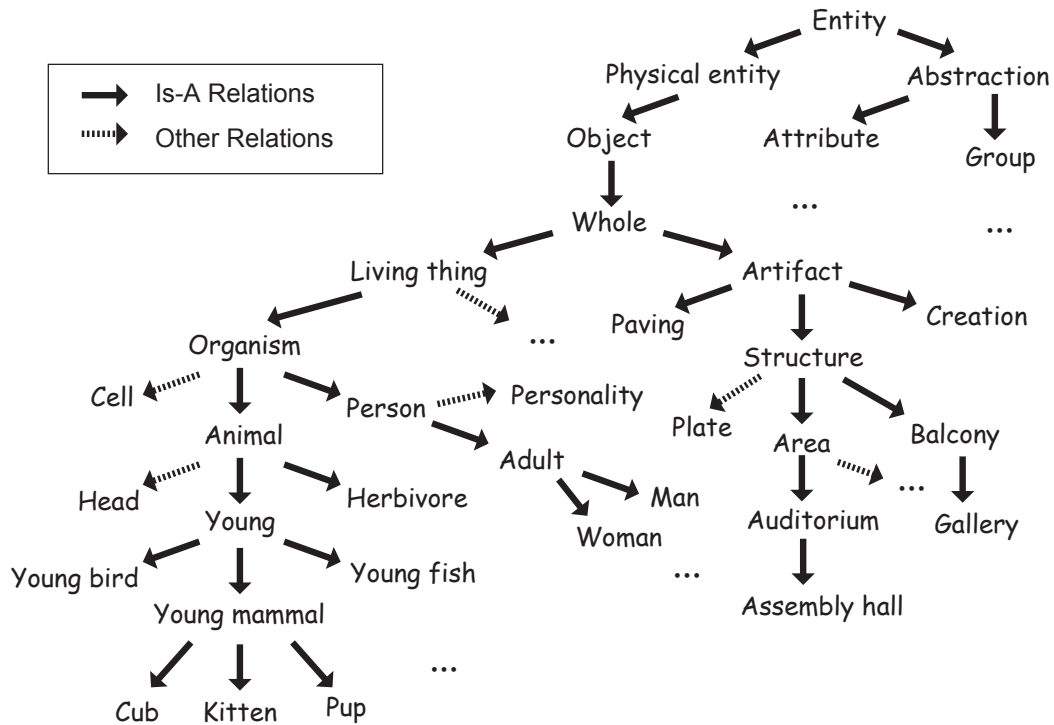


Fig. 2.2 An excerpt of noun relations defined in WordNet [1].

2.1.1 WordNet-based measures

WordNet [1] is a lexical database of English that has been developed under the direction of G. A. Miller. In WordNet, nouns, verbs, adjectives, and adverbs are grouped into sets of synonyms (synsets), each expressing a distinct concept. Each concept contains a *gloss* that expresses its semantics by means of a textual description and a list of words that can be used to refer to it. There are several types of relations used to connect the different synsets. Most of the connections consist of *is-a* relations (*hypernymy/hyponymy*), while others represent *part-of* relations (*holonymy/meronymy*), *antonymy* relations, *instance of*, and *has instance*. Note that WordNet is designed to provide general and common English words; it is not specialized for specific domains. An excerpt of noun relations defined in WordNet is illustrated in Fig. 2.2. In this subsection, we present a brief review of conventional WordNet-based measures.

Path-based similarity measures

Many methods have quantified information from WordNet based on the location of concepts in a hierarchy and the paths among them. The most intuitive method is using path counting, shortest path, and the depth in an ontology. Rada et al. [24] proved that the minimum number of edges separating two concepts is a metric for measuring the semantic distance between them. This measure can be converted to semantic similarity, which is equal to the inverse of the shortest path length between two concepts, which is expressed in the following manner:

$$\text{sim}_{\text{path}} = \frac{1}{\text{path_length}(c_1, c_2)}, \quad (2.1)$$

where $\text{path_length}(c_1, c_2)$ is a function to count the minimum number of edges separating concepts c_1 and c_2 . The above PATH measure cannot reflect semantic similarity that is coherent to human cognition, because it assumes that all the edges correspond to identical semantic distance.

To consider the granularity of a concept in the ontology, Wu and Palmer [25] extended the PATH measure by incorporating the depth of a concept that subsumes two concepts, the least common subsumer (LCS). Their measure is given by

$$\text{sim}_{\text{wup}}(c_1, c_2) = \frac{2 \times \text{depth}(\text{lcs}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}, \quad (2.2)$$

where $\text{lcs}(c_1, c_2)$ represents the lowest concept in the hierarchy that subsumes the concepts c_1 and c_2 .

Leacock and Chodorow [26] also modified the PATH measure to reflect the fact that edges that are lower down in the hyponymy/hypernymy relation correspond to smaller semantic distance than the ones that are higher up. For example, synsets corresponding to “station wagon” and “car” are much more similar than those corresponding to “transport” and “instrumentation” which are located higher up in the hierarchy, even though both pairs of concepts are separated by just one edge in the hierarchy. This formula is given by:

$$\text{sim}_{\text{lch}}(c_1, c_2) = -\log \frac{\text{path_length}(c_1, c_2)}{2 \times \max_{c \in \text{WordNet}} \text{depth}(c)}. \quad (2.3)$$

In both the above measures [25, 26], the path length between concepts is scaled in some manner by the overall depth. An edge between two very general concepts may

imply a reasonably large difference between the concepts, while one between two very specific concepts might represent a small difference. Note that path-based measures focus on semantic similarity rather than semantic relatedness due to their formulas that use direct paths between concepts.

Information content (IC)-based similarity measures

The limitation of path-based measures is that the degree of semantic similarity implied by a single edge is not consistent in an overall hierarchy. This problem can be solved by exploiting a corpus-based statistic known as *Information Content* (IC) [27], which is essentially a measure of the specificity of a concept. IC is defined as the negative log for the probability of a concept in the following manner:

$$IC(c) = -\log p(c), \quad (2.4)$$

where $p(c)$ is the probability of encountering concept c in a given corpus, which is determined by the frequency of occurrence of the concept and its subconcept in the corpus. If one assumes that general concepts have high probabilities since they occur frequently, then such concepts have low IC values. Resnik [27] exploited IC for measuring semantic similarity as follows:

$$\text{sim}_{\text{res}}(c_1, c_2) = IC(\text{lcs}(c_1, c_2)) = -\log(P(\text{lcs}(c_1, c_2))) \quad (2.5)$$

Note that Resnik's measure does not range from 0 to 1, and any two pairs of concepts having the same LCS are considered to have the same similarity.

To modify Resnik's measure, Lin [28] considered that the semantic similarity between two concepts can be measured by the ratio between the amount of information needed to state the commonality of the two concepts and the information needed to fully describe the two concepts. This formula is given below:

$$\text{sim}_{\text{lin}}(c_1, c_2) = \frac{2 \times IC(\text{lcs}(c_1, c_2))}{IC(c_1) + IC(c_2)} \quad (2.6)$$

On the other hand, Jiang and Conrath [29] defined the distance between two concepts to be the sum of the IC of the two concepts minus twice the IC of the concepts' LCS.

$$\text{sim}_{\text{jcn}}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(\text{lcs}(c_1, c_2))} \quad (2.7)$$

Budanitsky and Hirst [18] showed that this measure performs better than all measures described above. Because the information content is calculated from a target corpus, this type of similarity measure can be adapted to a particular domain, for example, biomedicine.

Gloss vector-based measures

The above approaches, that is, path-based measures and IC-based measures, are used to calculate the semantic similarity between concepts. To capture not only semantic similarity but also semantic relatedness, several methods refer to the glosses that are provided to concepts. The underlying assumption of this approach is that the inter-concept relationship is often implied by the shared words in their glosses. For example, Lesk [30] selected the sense of a word in a text, that has the highest word overlap between its definition and its context in the text.

Other approaches using glosses of concepts represent a concept using a feature vector derived from glosses. Patwardhan and Pedersen [31] extended Lesk's measure using second-order co-occurrence vectors, known as gloss vectors. In this method, a vector is created for each word in the concept's definition containing words that co-occur with it in a WordNet corpus. These word vectors are averaged to create a single co-occurrence vector for the concept. The similarity between the concepts is generally calculated by taking the cosine similarity between the concepts' second-order vectors.

Other measures on WordNet

Other measures exploit the conventional WordNet-based measures to provide a new concept space that is as compact and complete as possible. Wei and Ngo [32] presented Ontology-enriched Semantic Space (OSS), which is constructed on the basis of a linear combination of semantic relatedness computed from WordNet. Its procedure can be described in the following manner: given a set of pre-defined n concepts, the pair-wise similarity of each concept pair is computed using a WordNet-based measure (e.g., Wu and Palmer's measure [25]); this forms an $n \times n$ matrix. Then, to minimize the redundancy among concepts, n concept vectors in the matrix are grouped into $m < n$ clusters, and then one medoid from each cluster is selected to form the set of basis vectors. Thus, the OSS is spanned by m medoid concepts, which forms an $m \times m$ matrix. In OSS, a given concept can be easily represented as a vector in m dimensions, by calculating the WordNet-based measure between the concept and m medoids.

In [33], the authors further developed OSS and presented Ontology-enriched Orthog-

onal Semantic Space (OS^2). Each basis vector in OS^2 is the linear combination of concept vectors and orthogonal to each other.

2.1.2 Wikipedia-based measures

The online encyclopedia Wikipedia is a vast multilingual database of concepts and their semantic relationships. Wikipedia contains text pages for concepts and each page provides a detailed description of the corresponding concept. These pages are edited by online users and constitute collaborative knowledge resources. Compared with WordNet, Wikipedia provides entries on a vast number of named entities and very specialized concepts. In addition to richly interlinked textual information, Wikipedia provides categories to each article, which creates a taxonomy. In this subsection, I describe some approaches that use Wikipedia as a knowledge resource.

Measures based on text documents in Wikipedia

Strube and Ponzetto [34] were the first to compute semantic relatedness using Wikipedia. Their measure, called WikiRelate, follows conventional approaches that are applied to WordNet. Among the existing WordNet-based measures, Leacock and Chodorow's measure achieves the best performance. In addition, Zesch and Gurevych also adapt several WordNet-based measures to Wikipedia articles [35].

Gabrilovich and Markovich [36] present an approach called Explicit Semantic Analysis (ESA). ESA follows a similar approach to gloss vector-based measures on WordNet, explained in Section 2.1.1 Specifically, ESA is based on the vector space model that represents each term as a vector of Wikipedia concepts, where each element in the vector corresponds to a concept and the dimension is the number of Wikipedia articles. Each element is weighted by the tf-idf score of the word in the associated Wikipedia article, and the weight denotes the degree of relevance between the word and the concept. The semantic relatedness between concepts is calculated as the cosine similarity of the corresponding concept vectors. ESA captures the distributional similarity of the terms over Wikipedia.

Milne and Witten [37] employed the hyperlink structure of Wikipedia rather than its category hierarchy or textual content for representing a concept as a vector of concepts linking to it or a vector of concepts it links. Their method utilizes the normalized Google distance (NGD) [14], in which they substituted the frequency of concepts as the number of incoming links leading to a concept article and co-occurrence frequency as the

number of incoming links leading to both concept articles.

2.1.3 Drawback of the measures using knowledge bases

Among the existing methods that are based on knowledge bases, WordNet-based measures have displayed effectiveness for improving the performance of various tasks. For example, Varelas et al. [38] exploited WordNet-based semantic similarity measures for information retrieval tasks such as term expansion and document similarity computation. Wikipedia-based measures have also recently been used in text retrieval [39–42] due to the availability in several languages and its constant development by the community. These studies showed that introducing inter-concept relationships can improve the performance of several tasks. Because measures on the knowledge bases require a manual and well-defined hierarchy or encyclopedia, they result in limited coverage: for example, the number of concepts in WordNet is far lower than the number used to describe multimedia data on the Web [4, 43]. Furthermore, the number of concepts is increasing everyday: given a new concept, the semantic relations surrounding it should be flexibly extracted. To solve this problem, measures that automatically determine inter-concept relationships using the data on the Web as the background information corpus have been presented; these measures are described in the following sections.

2.2 Methods based on Web text documents

This section describes conventional methods that aim to extract semantic relations using text documents on the Web as an information source.

2.2.1 Web search engine-based measures

There have been several efforts to quantify inter-concept relations based on frequency and co-occurrence statistics in a large volume of text documents on the Web. The underlying assumption is that concepts (words) that tend to co-occur in similar contexts are semantically similar. Web search engines provide an efficient interface for obtaining a set of query-related documents. Several methods use information returned by inputting query keywords to web search engines like Google. Chen et al. [12] proposed a double-checking model based on text snippets returned by a Web search engine to compute semantic similarity. Bollegala et al. [13] proposed a word similarity measure that exploits

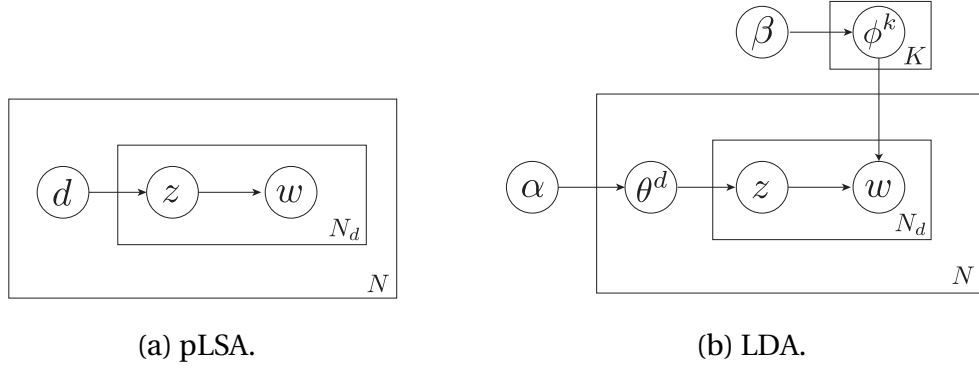


Fig. 2.3 Graphical model representation of probabilistic topic models.

various similarity scores that are based on pages and text snippets¹ returned by a Web search engine. Their method is based on supervised classification trained on positive and negative examples, such as SVMs, to find the optimal combination of different similarity scores. Cilibrasi and Vitanyi [14] developed normalized Google distance (NGD), which uses page counts returned by Google. The formula is well-founded on normalized information distance [44], which is defined using Kolmogorov complexity theories. The formula is given by

$$NGD(c_1, c_2) = \frac{\max\{\log h(c_1), \log h(c_2)\} - \log h(c_1, c_2)}{\log N - \min\{\log h(c_1), \log h(c_2)\}}, \quad (2.8)$$

where $h(c)$ denotes the number of documents in which concept c occurs, $h(c_1, c_2)$ denotes the number of documents that contain both concepts c_1 and c_2 , and N refers to the total number of documents. Note that NGD only focuses on page counts and does not consider multiple occurrences of the concept in a single document. Moreover, NGD is often converted to the semantic similarity in the range of $[0, 1]$, as shown below:

$$rel(c_1, c_2) = \exp\left\{-\frac{1}{\gamma} NGD(c_1, c_2)\right\}. \quad (2.9)$$

The parameter γ is usually set to the mean value of NGD of all possible combinations.

2.2.2 Probabilistic topic model-based measures

To capture the semantic relatedness between concepts, it is important to focus on their contexts: related concepts tend to appear in similar contexts. Topic models have been

¹A brief window of text extracted by a search engine around the query term in a document.

effectively used to reduce the dimensionality of a feature space. A basic assumption of topic models is that each item (document) is generated from latent topics. The most successful and well known of these techniques is probabilistic latent semantic analysis (pLSA) [45]. The graphical model of pLSA can be depicted as shown in Fig. 2.3 (a). The generative process of the pLSA is defined as: (i) a document d is chosen with a prior probability $P(d)$, (ii) a latent topic z is chosen from the document with probability $P(z|d)$, and (iii) a word w is chosen from the latent topic distribution with probability $P(w|z)$. Parameter estimation in the pLSA is based on maximum likelihood estimation (MLE) and approximated using an Expectation-Maximization (EM) algorithm. Specifically, the expectation step calculates the posterior probabilities of the latent variables. The equation can be obtained using Bayes's theorem:

$$P(z|d, w) = \frac{P(z)P(w|z)P(d|z)}{\sum_{z'} P(z')P(w|z')P(d|z')} \quad (2.10)$$

In the maximization step, the parameters are updated in the following manner:

$$P(w|z) = \frac{\sum_d n(d, w)P(z|d, w)}{\sum_{d, w'} n(d, w')P(z|d, w')}, \quad (2.11)$$

$$P(d|z) = \frac{\sum_w n(d, w)P(z|d, w)}{\sum_{d', w} n(d', w)P(z|d', w)}, \quad (2.12)$$

$$P(z) = \frac{\sum_{d, w} n(d, w)P(z|d, w)}{\sum_{d, w} n(d, w)}, \quad (2.13)$$

where $n(d, w)$ represents the number of times that a word w appears in a document d . While the pLSA can effectively reduce the dimensions of feature space, it cannot provide information on how the mixture weights $p(z|d)$ are generated. Thus, the number of parameters that need to be estimated grows linearly with the size of the corpus, which leads to the overfitting problem [46]. The Latent Dirichlet Allocation (LDA), presented by the authors in [46], uses a Dirichlet prior on document-topic distribution. The generative process in the LDA is similar to that in the pLSA: each word w in a document d is generated by sampling a topic z from topic distribution, and then sampling a word from topic-word distribution. This generative process can be illustrated as shown in Fig. 2.3 (b) and is represented in the following manner:

$$P(w_j) = \sum_{k=1}^K P(w_j|z_j = k)P(z_j = k) \quad (2.14)$$

where $P(w_j|z_j = k)$ is the probability of sampling w_j under the k -th topic, and $P(z_i = k)$ is the probability that the k -th topic is sampled for the j -th word. As shown in Fig. 2.3 (b), hyperparameters α and β are Dirichlet priors on $\phi^k = P(w|z = k)$ and $\theta^d = P(z)$, respectively.

Wang et al. [15] applied these topic models, pLSA and LDA, to a set of text documents to learn latent topics represented as probabilistic distributions over words. The authors in [15] exploited the Kullback-Leibler (KL) divergence to measure the semantic relatedness between two concepts and determine which concept is broader. Similarly, Ni et al. [47] applied LDA to a specific domain corpus to derive the target domain's thesauri. In their method, the relevance of two terms is measured as the similarity between the conditional topic distributions for the two terms.

The probabilistic topic model-based measures enable the calculation of the similarity between two concepts in a reduced feature space. Motivated from this, our approach presented in Chapter 4 also reduces the dimensionality of a feature space, that is found by using both textual and visual features.

2.2.3 Measures on social tags

The measures described in Section 2.2.1 are not limited to a specific search engine, and they can also work in a set of social tags. For example, the measures can also be applied to tagged images on Flickr, considering each tag list to be a document. Schmitz et al. [43] extracted concept hierarchical relationships using tag-frequency statistics on Flickr. Liu et al. [3] presented the concept of Tag Concurrence Distance (TCD), which uses the NGD framework by regarding tags of images as a single text document. This approach is very simple and has been used in several multimedia-related tasks, as described subsequently in Section 2.4.

Compared with the measures on raw text documents, tag-based measures do not need NLP preprocessing, such as POS tagging. Implementing these measures is easier, and the extracted inter-concept relations are suitable to target media, for example, images or videos.

2.2.4 Drawback of text-based measures

The use of text documents on the Web has the advantage of being able to encompass many more concepts than WordNet. On the other hand, the following limitations are associated with measures that are based on statistics of social tags: (i) some relations

such as synonyms might be overlooked because the measures do not account for the semantic contexts in which the tags co-occur, and (ii) the measures occasionally suffer from the incompleteness of tag co-occurrences [4]. Although probabilistic topic model-based measures effectively reduce the dimensionality of tag space, tags usually occur only once in an image, not multiple times, and this does not contribute to the frequency feature of the tag in the formula of the topic model. In addition, these measures might not be able to capture the similarity between concepts in a visual aspect.

To complement such a lack of information of tags, recent methods exploit visual features calculated from images on the Web. In the next section, we describe state-of-the-art measures based on visual features.

2.3 Methods based on visual features

As initially described in Chapter 1, the popularity of various digital imaging devices and the advances of social media websites have resulted in proliferation of digital images on the Web. Some successful photo-sharing websites, such as Flickr, allow users to upload personal images and assign tags to describe the image contents. In recent studies, the utilization of visual features extracted from the images has been investigated to explore inter-concept relations. In this framework, for each concept, a collection of images is obtained from the photo sharing website using the concept as a query keyword. Then, visual features are extracted from the images, and the concept's distribution in a visual feature space is calculated. In the following subsections, we introduce two types of measures based on visual features: measures that quantify the semantic gap (See Section 2.3.1) and measures that evaluate distributional similarity (See Section 2.3.2).

2.3.1 Measures that focus on the semantic gap

Several methods focus on the fact that the abstraction levels of concepts are relevant to the semantic gap [48, 49] between low-level visual features and high-level concepts and quantify the semantic gaps through data-driven approaches. Yanai and Barnard [50] presented a pioneering approach to quantify “visualness” of concepts, that is, the extent to which concepts have visual characteristics. Their method selects regions associated with a certain concept from a labeled image and computes the entropy of the selected regions based on a Gaussian mixture model (GMM) for the regions. Lu et al. [17] presented a method to identify concepts with small semantic gaps for image categorization.

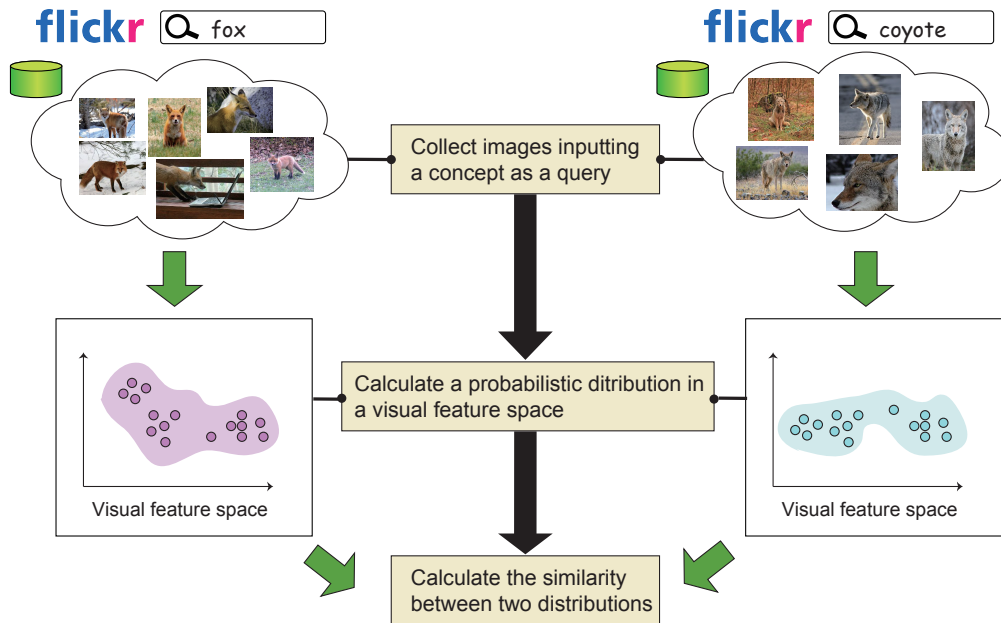


Fig. 2.4 An example of calculating semantic relatedness between concepts based on visual features.

The authors use Web images to develop a sorted list of vocabularies based on semantic gap: the top ranked concepts in the list are easier to learn using visual features than the bottom-ranked ones. Furthermore, Sun and Bhowmick [51] calculated the compactness of the distribution for each concept in the visual feature space to quantify the tag's "visual-representativeness". The calculated score for each tag characterizes the effectiveness of the tag in describing the visual content of the set of images annotated with the tag. Fan et al. [52] also presented a method for quantitative characterization of semantic gaps, which is based on a kernel-based visual similarity between two images. The authors in [52] identified the concepts with small semantic gaps and trained their one-against-all SVM classifiers independently. For concepts with large semantic gaps, they used a greater number of images and trained their inter-related SVM classifiers jointly. Xu et al. [53] also quantified the visualness of a concept using a large-scale image dataset. These measures produce useful information for image annotation and retrieval. For example, one should select the most visual-representative concepts to construct a vocabulary for automatic annotation.

2.3.2 Measures based on visual similarity

Recently, the semantic relatedness between concepts has been measured using visual features. An example of calculating the semantic relatedness is depicted in Fig. 2.4. Fan et al. [54] calculate inter-concept visual similarity using a kernel-based visual similarity between two images, which is also used in [52]. The constructed visual concept network was used for designing concept classifiers of images. The method in [54] was extended by [55], in which visually similar concepts were automatically grouped and inter-group separation was performed. They compared the performance of their visual concept network with other measures using WordNet or concept co-occurrence in an image classification task. Wu et al. [4] presented a novel measure, called Flickr distance, for calculating the visual similarity between concepts. For each concept, a latent topic visual language model (LTVLM) is constructed using spatial information of texture histograms that are extracted from a set of images that are tagged with the concept. Then, the Jensen-Shannon (JS) divergence between two LTVLMs is computed as a visual similarity between two concepts. This approach is similar to conventional methods that use probabilistic topic models, described in Section 2.2.2. Similarly, Wang [5] constructed a GMM for each concept in a visual feature space and calculated the KL divergence between two GMMS.

2.3.3 Drawbacks of visual feature-based measures

The measures described in Section 2.3.2 can capture the visual similarity between concepts that represent specific objects or scenes. For example, as shown in Fig. 2.4, relevance of a concept pair such as “coyote”-“fox” can be effectively extracted due to its similarity in a visual context. On the other hand, the relationships around abstract and general concepts might not be accurately extracted because it is difficult to model their semantics using only visual features. This difficulty has been revealed by conventional studies, as noted in Section 2.3.1.

Although tagged images on Flickr provide valuable information such as text features and their corresponding visual features, the above-mentioned conventional measures exploit a single modality, that is, either text or visual features only. Several studies on multimedia annotation and retrieval have revealed that multimedia content can be recognized better using both text features and visual features rather than using only one of the feature modalities [56–58]. Thus, a cross-modal approach using text features and visual features is expected to present measures that are more coherent with human cogni-

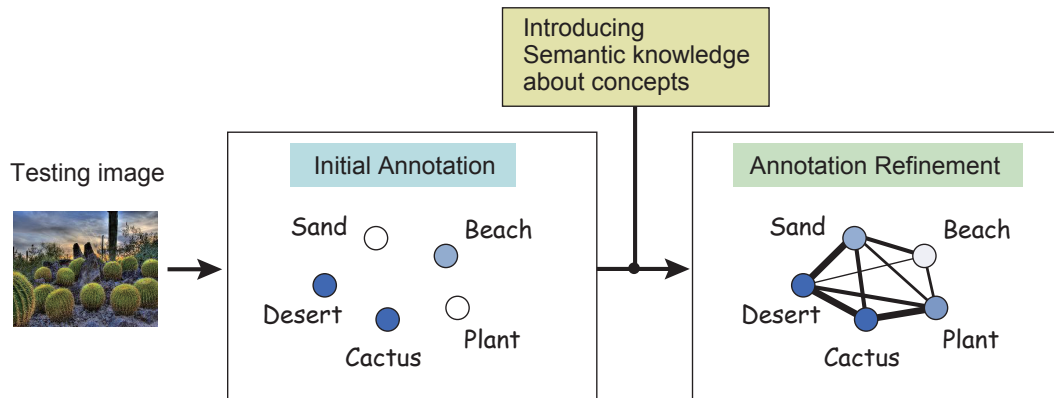


Fig. 2.5 An illustration of general framework that introduces semantic relationships between concepts for image annotation refinement.

tion compared with conventional approaches; the resulting advantage is the motivation for this thesis.

2.4 Applications of semantic measures to multimedia related-tasks

The semantic relations between concepts have been introduced to improve performance of multimedia related-tasks such as multimedia annotation, retrieval, and tag recommendation. In this section, I provide a review of the applications. In these applications, the semantic relationships between concepts are generally regarded through semantic graphs, in which a node represents each concept and an edge is weighted by the semantic relatedness between concepts.

2.4.1 Multimedia annotation

Multimedia annotation is one of the most important applications for facilitating information retrieval. In this research field, Large-Scale Concept Ontology for Multimedia (LSCOM) [59], which includes 834 concepts and training images for 449 out of the 834 concepts has been used as a source of video news annotation. However, this vocabulary is limited and, hence, automatic construction of such an ontology is necessary. Another important resource in this research field is ImageNet [60], a large scale ontology of im-

ages built upon WordNet information. This database contains 500 - 1000 clean and full resolution images for each selected synset of WordNet. The labels of each image are manually corrected using the Amazon Mechanical Turk (AMT)². Since ImageNet depends on a fixed ontology, the automatic extraction of inter-concept relations facilitates the construction of such a database. Although these resources are limited in terms of scalability due to manual construction, several studies have shown the effectiveness of using inter-concept relations in multimedia annotation. Below, I describe conventional methods that utilize inter-concept relations for multimedia annotation.

Introducing hierarchical relations to annotation models

Semantic hierarchies such as WordNet have been utilized to improve performance of image annotation. Jin et al. [61] utilized the semantic similarity between concepts to eliminate irrelevant keywords. The authors used several WordNet-based measures and finally fuse outcomes of all the measures to make a final decision. Liu et al. [62] presented a Dual Cross-Media Relevance Model (DCMRM), which introduces the WordNet-based measure of semantic relatedness between labels. The aim of DCMRM is to predict the joint probability of a testing image I_q and a keyword w_k as follows:

$$w^* = \arg \max_{w_k \in V} p(w_k, I_q), \quad (2.15)$$

where V is a set of keywords and w^* denotes an optimal keyword for annotation. The DCMRM assumes that the probability of observing both the annotation keyword w_k and the image I_q are mutually independent given a keyword w_j , so that the relevance model is represented as follows:

$$w^* = \arg \max_{w_k \in V} \sum_{w_j \in V} P(I_q|w_j)P(w_k|w_j)P(w_j), \quad (2.16)$$

where $P(w_k|w_j)$ denotes the probability of a keyword w_k given a keyword w_j (See (1)) and $P(I_q|w_j)$ denotes the probability of image I_q given a keyword w_j (See (2)). In the method presented in [62], they calculate the keyword relevance as distance $dist(w_k|w_j)$ by using the conventional co-occurrence approach [3] and calculate the conditional probability $P(w_k|w_j)$ as $P(w_k|w_j) = \exp \{-\gamma \cdot dist(w_k|w_j)\}$, where γ is an adjustable param-

²<https://www.mturk.com/mturk/welcome>

eter. In the DCMRM, we first calculate image-based similarity as follows:

$$S(I_q, R_k) = \sum_{I_i \in R_k} \alpha_i \exp \left\{ -\frac{d(I_i, I_q)}{\sigma_l} \right\}, \quad (2.17)$$

where R_k is a set of images annotated with keyword w_k , α_i is an adjustable parameter that aims to support similar image-pairs and penalize dissimilar image-pairs, and $d(I_i, I_q)$ is a certain distance metric between image I_i and query image I_q . We set the average distance to parameter σ_l . The final word-to-image relation can be approximated by this measure, which is given as

$$S_{WIR}(I_q, w_k) = [S(I_q, R_k)]^\eta, \quad (2.18)$$

where η is a parameter. In Chapter 3, we introduce the extracted inter-concept relationships to this model for quantitative evaluation.

Several approaches exploit hierarchical information of an ontology for learning image categories. Marszalek and Schmid [63] extract the subgraph around given class labels from WordNet to construct the classifiers. They classify images top-down, beginning from the root node, down to a leaf node in a hierarchy. Fan et al. [64] presented a hierarchical approach that learns the ensemble classifier of some specific concepts for high-level image concepts. They use both WordNet information and visual similarity between concepts to build an ontology, and allow image annotation using concepts that have different abstraction levels. Furthermore, Torralba et al. [65] categorizes images at different abstraction levels using a hierarchical vote.

Annotation refinement with a semantic graph

Another strategy for multimedia annotation is to modify initial annotations to be semantically consistent by introducing inter-concept relations as a semantic graph. An example of image annotation refinement frameworks is depicted in Fig. 2.5. As illustrated in Fig. 2.5, concepts are first estimated using only visual features of a given image; thereafter these words are refined on the basis of the semantic relatedness between concepts. For the annotation refinement task, many methods exploit random walk with restarts (RWR), which is a special case of the Markov Chain. The most famous work of RWR is PageRank [66], in which a random surfer makes random transitions between Web pages in a Web graph following hyperlinks and occasionally jumps to another page because of boredom. We denote V as a vocabulary of concepts, and $|V|$ is the vocabulary size.

Let $\mathbf{A}$ be a row-normalized adjacency matrix whose (i, j) -th entry denotes semantic relatedness between the i -th concept and the j -th concept of V . Given a testing image, a $|V|$ -dimensional initial annotation score vector $\mathbf{v}$ is first provided using its visual features by assuming that each concept is independent from the others, where the j -th entry of $\mathbf{v}$ denotes a detection score of the j -th concept of V . Then, the annotation confidence scores of the concepts are updated at the t -th iteration using the following equation:

$$\mathbf{u}_{t+1} = (1 - \alpha)\mathbf{A}\mathbf{u}_t + \alpha\mathbf{v} \quad (2.19)$$

where $\mathbf{u}_t$ denotes a $|V|$ -dimensional annotation score vector at the t -th iteration, and $\alpha \in [0, 1]$ is a parameter that balances the contribution of semantic relatedness and initial annotations. In [67], RWR is exploited for image annotation refinement. In [68], RWR is employed to refine annotation results of a given news video, in which WordNet-based measures and co-occurrences are used as inter-concept relations.

Similarly, many methods construct a semantic graph using inter-concept relations and use it to refine image annotation. Wang and Gong [69] modeled the image annotation refinement problem using conditional random field (CRF). The refined annotation is given by inferring the most likely states of the nodes in the graph. Zhu et al. [70] formulated the image annotation refinement problem as a decomposition of the user-provided tag matrix into a low-rank refined matrix and a sparse error matrix. Ji et al. [71] use WordNet-based measures to introduce inter-concept relationships as correlation among nodes in Hidden Markov Random Field. Furthermore, Liu et al. [72] formulated the image annotation refinement problem as an optimization scheme that simultaneously takes into account visual consistency of images and semantic consistency of tags. These methods use tag co-occurrences as semantic relatedness [3, 14]. Sang et al. [73] presented a ranking based multi-correlation tensor factorization to jointly model the ternary relations among user, image, and tag. To construct a tag affinity graph in their scheme, both tag co-occurrence and WordNet taxonomy are utilized. In [8, 74], the authors presented domain adaptive semantic diffusion (DASD). This approach refines image annotation results using a graph diffusion technique, which corresponds to lead the consistency and smoothness of the annotation over a semantic graph. During the annotation refinement, DASD also adapts the concept affinities to new testing images. Thus, two outputs of DASD can be obtained: refined annotation results and concept relationships in the domain of testing images.

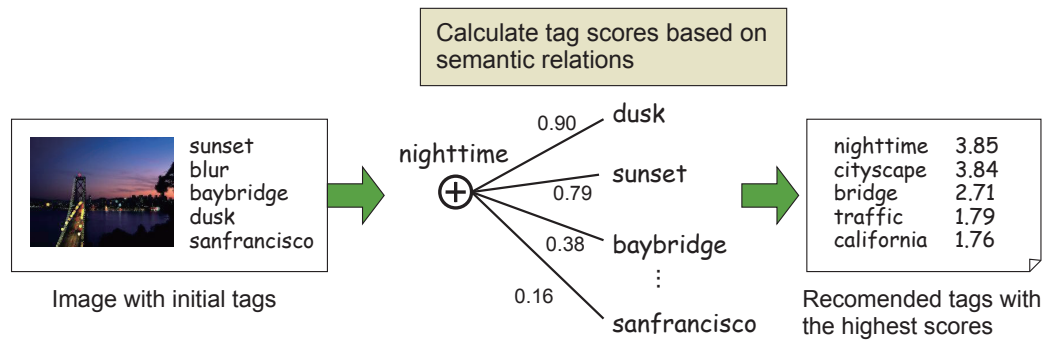


Fig. 2.6 An example of tag recommendation that exploits semantic relatedness between concepts.

2.4.2 Multimedia retrieval

Semantic hierarchy has also been exploited in multimedia retrieval. Popescu et al. [75] associated the use of WordNet and visual features in image retrieval task. Ferecatu et al. [76] also combined the WordNet information and visual features for relevance feedback of image retrieval. Wei et al. [33] construct a semantic space using WordNet-based measures to facilitate the selection of concept detectors for video retrieval. Furthermore, Lienhart et al. [77] exploit WordNet hierarchy to enrich the annotation of images in image retrieval tasks. For each image, parent words of its tags are derived from WordNet and added to the image, which facilitates the pLSA training. Deng et al. [78] showed that using a hierarchy that relates semantic labels during training improve the performance of similar image retrieval. The authors noted that there are other possible ways to obtain semantic similarity between concepts or attributes, like Web data mining, when a manually constructed hierarchy is not available. Deselaers and Ferrari [79] presented a distance function between two images for image retrieval based on visual nearest neighbors in ImageNet.

2.4.3 Tag recommendation

Tag recommendation is also an important application of semantic relatedness. This task aims to assist a user's tagging by suggesting a set of candidate tags that are related to the image. An example of tag recommendation frameworks is presented in Fig. 2.6. In [80], several strategies for tag recommendation are presented, in which tag co-occurrences are utilized. Given a testing image with initial tags, a set of indexes of tags that are related to

the j -th initial tag based on a measure of semantic relatedness is obtained. Then, for example, a recommendation score of the k -th tag of the vocabulary, $score(k)$, is calculated, using the following equation (called *sum strategy*):

$$score(k) = \sum_{j \in U} (rel(j, k), \text{if } k \in \Omega_j), \quad (2.20)$$

where U denotes a set of indexes of initial tags and $rel(j, k)$ denotes semantic relatedness between the j -th and k -th tags of the vocabulary. Tags that have high scores are finally recommended to the image. In this thesis, we apply the semantic relations extracted by methods to these multimedia-related applications for performance evaluation.

Heymann et al. [81] use association rules to create a set of candidate tags. Garg and Weber [82] propose a personalized and interactive tag recommendation system. When a user adds a tag to an image, the system recommends a ranked list of candidate tags based on the user's past tagging. Wu et al. [83] presented a learning-based multi-modal algorithm to suggest relevant tags. This framework utilizes both tag and visual correlation, and applies Rankboost algorithm to learn the optimal combination of the ranking features from the two modalities. In [84], the authors utilize multiple textual features in several heuristic methods for tag recommendation in social media websites.

Chapter 3

A Classifier-Based Approach for Extracting Inter-Concept Relationships from An Image Database

3.1 Introduction

Concept relationship extraction from an image database has attracted much attention as a means of facilitating image annotation and retrieval. As described in Chapter 2, many conventional methods regard each training image in a target database as a document containing concepts and use the co-occurrence frequency of two concepts to measure their relevance [8, 14, 67]. This scheme is independent of visual features and there is, therefore, no guarantee that relevant concepts always co-occur in the same image. In [4], semantic relatedness is measured by calculating the similarity between two probabilistic models that are trained using visual features. This method uses visual features effectively, but is independent of context information from instances of concept co-occurrence. In [8], Jiang et al. focused on two similarities in the target database by employing the co-occurrence frequency of concepts and visual similarities obtained by comparing classifier outputs to construct a concept relationship graph for image annotation. However, the extracted relationships do not always represent the same objects or scenes because visual similarities and concept co-occurrences were used separately. This chapter aims to investigate whether the simultaneous use of both textual and the visual features offers an advantage in extracting inter-concept relationships. To simplify the problem, in this chapter, we assume that a given image database is labeled accurately

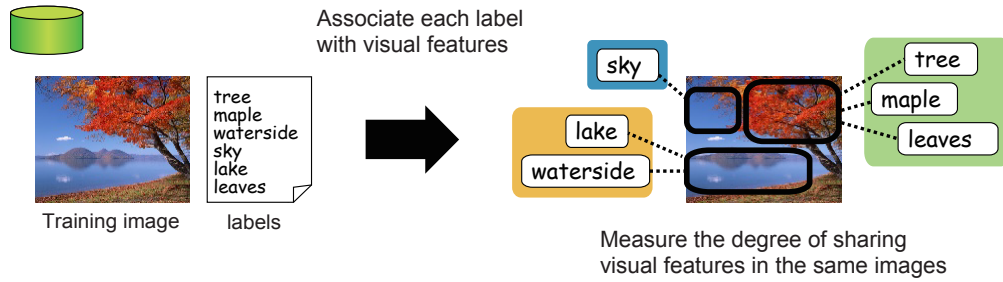


Fig. 3.1 An illustration of the proposed approach. Given a training image, each of its labels is first associated with visual features of the image, and then the degree of sharing visual features among labels is measured as the visual feature-based relevance.

with a predefined set of concepts by experts.

To capture semantic similarity rather than co-occurrence relations, we need to know the visual features that are shared by these concepts in images. A schematic illustration of our idea is shown in Fig. 3.1. According to this figure, given a labeled training image, we associate each concept label with visual features of the image. This corresponds to referencing concept co-occurrences, i.e., context information. If visual features are shared by two concepts in the same image, the two concepts tend to share a hyponymy/hypernymy or meronymy relationship.

In this chapter, we present a classifier-based approach for extracting inter-concept relationships from an image database, to realize the above-mentioned idea. First, a classifier is trained for each concept using visual features extracted from positive and negative samples of the concept. Specifically, we choose logistic regression (LR) for this task. Logistic regression is a linear classifier, and we can easily find the features that contribute to the estimation of a target concept by focusing on its parameters. Then, we present novel schemes that exploit both context information and visual similarity for calculating semantic similarity between the concepts. The method presented in this chapter focuses on quantifying hyponymy/hypernymy and meronymy relations, rather than co-occurrence relations.

The main contribution of this chapter is two-fold: (a) we focus on the degree to which two concepts share visual features in the same image to calculate the semantic similarity between said concepts, and (b) we compare the proposed method with conventional methods that exploit context information or visual features only to show the effectiveness of using both features.

The rest of this chapter is organized as follows: Section 3.2 describes how to select visual features relevant to each concept through classifier training. Section 3.3 presents the aforementioned novel schemes for extracting inter-concept relationships using not only visual features but also contextual information. In Section 3.4, results of the experiments conducted for verifying the effectiveness of the proposed method are presented. Finally, the chapter is summarized, and some possible directions for future work are suggested in Section 3.5.

3.2 Feature Selection Using Classifiers

In this section, we show how to select visual features relevant to each concept. Given a set of accurately labeled images I_i ($i = 1, 2, \dots, N$), we first extract their visual features $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,d_x}]^T$. Let C_k be a concept in the database. We set the binary variable $l_i^k = 1$ if image I_i is labeled with concept C_k and $l_i^k = 0$ otherwise. Thus, the set of training data is represented by $S = \{\mathbf{x}_i, l_i^k | i = 1, 2, \dots, k = 1, 2, \dots, K\}$. Below, we describe how to train each classifier and use it for feature selection. Furthermore, in Section 3.3.4, we discuss classifiers for the proposed framework.

3.2.1 Training logistic regression

This subsection describes the training of logistic regression, which was chosen for feature selection in the proposed method. Logistic regression is one of the popular classifiers and has been used in many areas owing to its simplicity and effectiveness. Its applications for information retrieval include text categorization [85], image annotation [86, 87], and video concept detection [88]. Using logistic regression, the probability that image I_i contains concept C_k is calculated as follows:

$$y_i^k = p(l_i^k = 1 | \mathbf{x}_i) = \frac{1}{1 + \exp(-\boldsymbol{\beta}_k^T \mathbf{x}_i)}, \quad (3.1)$$

where $\boldsymbol{\beta}_k = [\beta_{k,1}, \beta_{k,2}, \dots, \beta_{k,d_x}]^T$ is a parameter vector. The above equation gives us a conditional generative model for l_i^k given $\mathbf{x}_i$. Logistic regression is a linear classifier because its decision boundary is determined by a linear function of the input features. Assuming binary classification such that an estimate of the existence of a label for concept, $p(l_i^k = 1 | \mathbf{x}_i)$ can be used for determining the class label using a threshold of 0.5.

The parameter vector $\boldsymbol{\beta}_k$ can be determined by minimizing the following negative

log-likelihood function:

$$\begin{aligned}
-\log P(S|\boldsymbol{\beta}_k) &= -\sum_{i=1}^N \log P(\mathbf{x}_i, t_i^k) \\
&= -\sum_{i=1}^N \left[t_i^k \log y_i^k + (1 - t_i^k) \log(1 - y_i^k) \right].
\end{aligned} \tag{3.2}$$

Because there exists no closed form solution for Eq.(3.2), the parameter vector is generally estimated using the Newton-Raphson update or the stochastic gradient descent (SGD) algorithm. In this chapter, we use the SGD algorithm given its efficient performance even when the number of data points is very large.

3.2.2 Feature selection using logistic regression

Logistic regression is able to not only classify images but also detect the relevant features to each concept. This is because logistic regression is a linear discriminative model, and Eq. 3.1 has a decision boundary in which $\boldsymbol{\beta}_k^T \mathbf{x}_i$ is zero. The parameters obtained for a target concept represents the visual features that are effective for estimating the concept's existence. Assume that each element of the visual feature vector is positive, e.g., histograms. In the case that the d -th visual feature of an input image is related to the target concept C_k , the corresponding parameter $\beta_{k,d}$ should be positive so that the visual feature $x_{i,d}$ contributes to the annotation of concept C_k . Thus, in this chapter, we assume that the d -th feature corresponding to a parameter such that $\beta_{k,d} > 0$ is a relevant feature for concept C_k .

If the selected features are compared among concepts, we can obtain the visual similarity between the concepts. There are several methods for calculating the similarity between two vectors. For example, we calculate the cosine similarity between two parameter vectors $\boldsymbol{\beta}_j, \boldsymbol{\beta}_k$ as follows:

$$sim_{cos}(C_j, C_k) = \frac{\boldsymbol{\beta}_j \cdot \boldsymbol{\beta}_k}{\|\boldsymbol{\beta}_j\| \|\boldsymbol{\beta}_k\|}, \tag{3.3}$$

where $\cdot$ represents the inner product. A high value of $sim_{cos}(C_j, C_k)$ means that the two concepts C_j and C_k are quite similar visually and vice versa. The comparison among selected features corresponds to simply calculating visual similarity between concepts, which is the same as the approach presented in conventional studies [4, 5]. To effectively use the co-occurrence information of labels, a novel framework is introduced in the next

section.

3.3 Extracting Inter-Concept Relationships Based on Shared Visual Features

In this section, we present an approach for extracting inter-concept relationships based on shared visual features. If concept C_k ($k = 1, 2, \dots, K$) is semantically similar to concept C_j ($k \neq j$), it is clear that they share the same visual features in images from the database. To find such concepts, we reference the concept labels associated with the training images. Below, our approach is divided into three steps that are executed sequentially.

1. Removal of relevant features

We remove the visual features related to a concept from all images containing the concept and monitor how difficult it becomes to estimate the other concepts (See 3.3.1).

2. Determination of relevance between concepts

If the performance of concept C_j worsens upon the removal of features related to concept C_k , concept C_j has a visual feature-based relevance with concept C_k (See 3.3.2).

Measurement of relevance between concepts We consider the change in the performance of concept C_j as a strength of the visual feature-based relevance from concept C_k to concept C_j (See 3.3.3).

In the following subsections, the details of the abovementioned three steps are described.

3.3.1 Removal of relevant features

After training LR for concept C_k , we can detect the visual features relevant to concept C_k using the obtained parameter, as described in Section 3.2. We first remove the visual features relevant to concept C_k from the training image I_i containing concept C_k by using the obtained parameter vector β_k and generate a new visual feature vector

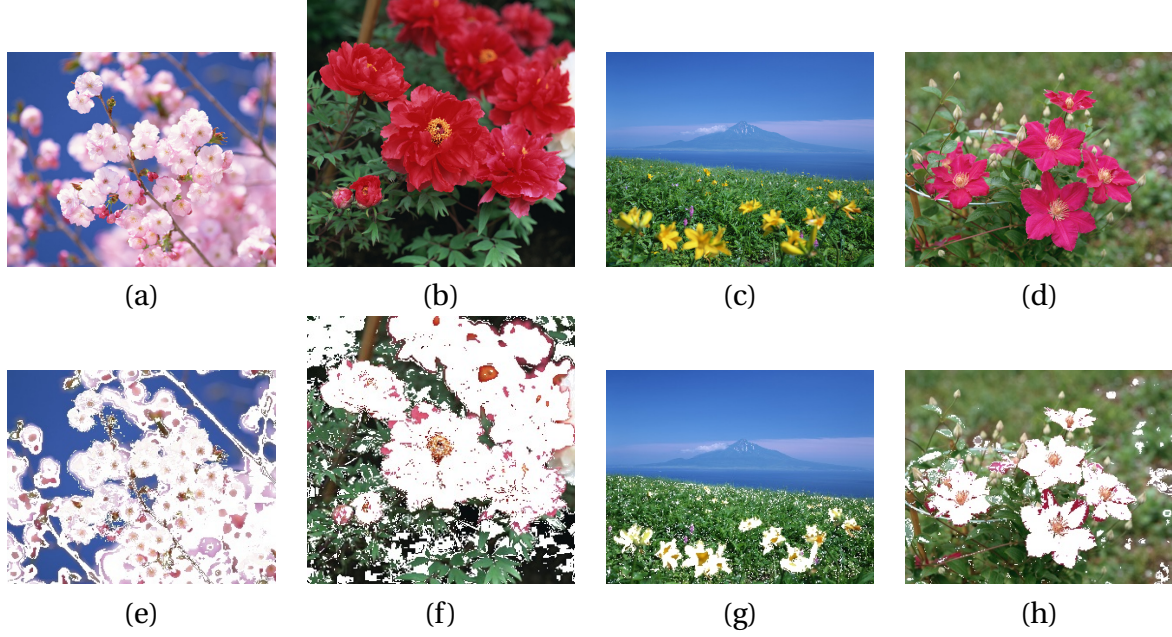


Fig. 3.2 Examples of visual feature removal relevant to “bloom”: (a)-(d) Original images A-D, (e)-(h) Images in which visual features relevant to “bloom” are removed using Eq. (3.4) from Original images A, B, C, and D.

$\mathbf{x}_i^k = [x_{i,1}^k, x_{i,2}^k, \dots, x_{i,D}^k]^T$ as follows:

$$x_{i,j}^k = m_j^k x_{i,j}, \quad (3.4)$$

$$m_j^k = \begin{cases} 0 & \beta_{k,j} > 0 \\ 1 & \text{otherwise.} \end{cases}$$

This equation provides a new training set, $S^k = \{\mathbf{x}_i^k, t_i^k \mid i = 1, 2, \dots, N, k = 1, 2, \dots, K\}$, from which the visual features related to concept C_k are removed. The proposed method removes visual features related to a target concept from images annotated with that concept. This corresponds to using concept co-occurrence information because such removal would affect the performance of concepts that co-occur with the target concept. For an explanation of this removal process, some examples are shown in Fig. 3.2, in which the target concept is “bloom.” Figure 3.2 shows that the visual features representing the concept “bloom” are effectively removed using Eq. (3.4). Also, we can see that removal of visual features related to concept “bloom” affects the performance of concept “flower” that co-occurs with “bloom” in the images. This is equivalent to detecting the visual feature-based relevance among concepts. If we remove features related to the target

concept from all images, it affects the estimation of concepts even if they do not co-occur with the target concept, which means that only the visual similarities are computed for a pair of concepts.

3.3.2 Determination of Relevance Presence Between concepts

Based on the new training set S^k , we train the other LR for concept C_j ($j \neq k$). This training provides a new parameter vector β_j^k . From the obtained results, we monitor the performance of concept C_j . Specifically, we utilize

$$\begin{aligned} d_k(j) &= -\log P(S^k | \beta_j^k) - [-\log P(S | \beta_j)] \\ &= -\sum_{i=1}^N \left[t_i^j \left\{ \log z_i^{k,j} - \log y_i^j \right\} + (1 - t_i^j) \left\{ \log(1 - z_i^{k,j}) - \log(1 - y_i^j) \right\} \right], \end{aligned} \quad (3.5)$$

as the criterion for determining the relevance from concept C_k to concept C_j , where

$$z^{k,j} = \frac{1}{1 + \exp(-\beta_j^{k,T} \mathbf{x})}.$$

The above criterion $d_k(j)$ represents the difference in performance of concept C_j between the two LR whose parameter vectors are β_j and β_j^k . Since the LR is trained to minimize the estimation error, we compare the performance of the two models using this criterion. If $d_k(j) > 0$, the performance of concept C_j worsens owing to the removal of visual features relevant to concept C_k . This means that the removed visual features are necessary for estimating concept C_j . In contrast, if $d_k(j) < 0$, the performance of concept C_j becomes better owing to noise removal. We denote R_k as a set of concepts C_j that satisfy $d_k(j) > 0$, i.e., related to concept C_k based on visual features.

3.3.3 Measurement of Relevance Between Concepts

We measure the degree of visual feature-based relevance from concept C_k to concept C_j by using the following equation:

$$D(C_j | C_k) = -\sum_{i=1}^N \left\{ y_i^j \log z_i^{k,j} + (1 - y_i^j) \log(1 - z_i^{k,j}) \right\}. \quad (3.6)$$

Using the above equation is equivalent to calculating the degree of change in performance after the removal of visual features related to concept C_k . A large value of the cal-

culated degree $D(C_j|C_k)$ implies that concept C_j shares most of its visual features with concept C_k . We normalize the degree for each concept C_k and calculate the relevance scores as follows:

$$r(C_j|C_k) = \frac{D(C_j|C_k)}{D(C_k|C_k)}. \quad (3.7)$$

This conversion is useful for comparing the strengths of similarities among concept pairs. The score $r(C_j|C_k)$ approaches one when the relevance is strong. Thus, visual feature-based concept relationship extraction is realized by using the proposed method.

3.3.4 Discussion of classifiers

In this subsection, we discuss the classifiers used in the proposed framework. Our approach requires a classifier that enables us to execute the following two steps:

- (i) Detection of visual features relevant to each keyword,
- (ii) Evaluation of performance in the learning stage for measuring the degree of relevance.

Because logistic regression can satisfy these requirements in a simple manner, we apply it to the framework proposed in this paper. If other classifiers can satisfy these requirements, e.g., accurately classify images or effectively detect visual features better than the LR, it is expected that the concept relationship extraction performance will improve. For example, non-linear methods such as support vector machines (SVMs) and kernel logistic regression have been used widely in many applications. These methods can be used for our framework as well by introducing some procedures. The issue of selecting a suitable classifier for the proposed framework will be discussed in our future work.

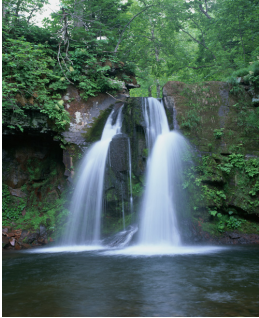
3.4 Experimental Results

In this section, experimental results are presented for verifying the effectiveness of the proposed method. To show whether the proposed method can extract concept relationships adaptively from a target image database, we used two different databases: IAPR-TC12 dataset [6] and Hokkaido landscape dataset. Detailed descriptions of these databases are given below.

Table 3.1 Examples of images and their labels in IAPR TC-12 dataset [6].

			
building, car, child, column, front, house, palm, square, tree, woman	beach, chair, lawn, people, tree	bed, blanket, floor, pillow, room, towel, wall	llama, mountain, range, terrace

Table 3.2 Examples of images and their labels in Hokkaido landscape dataset.

			
nature, land, waterside, night, sky, vegetation, tree, buildings, street, lake, illumination, street light, lighting, romantic, summer, outdoor	nature, land, rock, waterside, river, forest, waterfall, tree, vegetation, leaves, solid, liquid, outdoor	snow, cute, winter, night, outdoor	land, pasture, animal, mammal, herbivore, cow, work, farmwork, agriculture, outdoor

IAPR-TC12 dataset

This database was introduced in [6] for cross-language information retrieval, and it consists of 17,665 training images and 1,962 testing images. Each image is manually annotated with an average of 5.7 labels (standard deviation of 2.6) out of 291 noun concepts. The vocabulary consists of several concepts such as nature-related (e.g., “sky,” “mountain,” “meadow”), architecture-related (e.g., “church,” “stadium,” “corridor”), and clothes-related (e.g., “sock,” “skirt,” “pant”). Examples of images and their labels in this database are listed in Table 3.1.

Hokkaido landscape dataset

This database contains 22,000 color images of Hokkaido’s landscapes, which were provided by Datacraft Co., Ltd¹. There are 230 concepts in the database, and each image is manually labeled with 10-20 concepts of the 230 concepts as ground truth. Examples of images and their labels in this database are listed in Table 3.2. We used 20,000 images for training and 2,000 images for testing.

For feature extraction from the two databases, each image in the databases is divided into about 20 regions by using a multi-resolution implementation of the recursive shortest spanning tree (RSST) method [89]. Then, global features representing color and texture are extracted from each region in the same manner as in [90, 91]. Those extracted features are used to form a 200-dimensional feature vector $\mathbf{x}_i$ by following the bag-of-words approach.

In the following subsections, we compare our concept relationship extraction method with the following methods.

- **Comparative method A:** an approach presented by the authors in [3], which uses only concept co-occurrence.
- **Comparative method B:** an approach introduced by the authors in [4], which uses only visual features.
- **Comparative method C:** an approach presented for image annotation in [8], which derives inter-concept relations during graph diffusion. This corresponds to the separate use of visual features and concept co-occurrence.

¹<http://www.datacraft.co.jp/>

Section 3.4.1 visualizes parts of the results of concept relationship extraction from each of the databases, and Section 3.4.2 shows user evaluation results of the extracted relationships. Furthermore, in Section 3.4.3, for quantitative evaluation, we apply the extracted relationships to automatic image annotation of the tested images.

3.4.1 Visualizing Concept Relationships

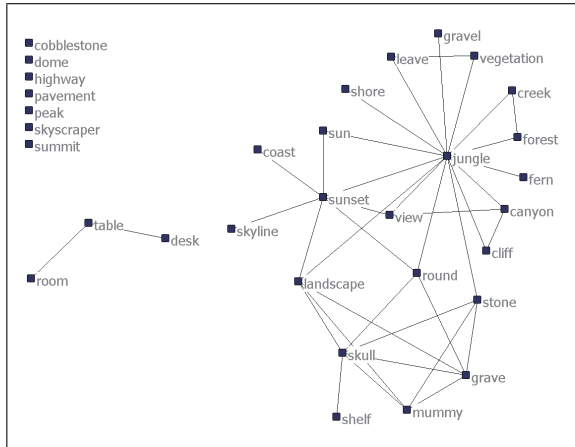
In this subsection, we use the proposed method to extract concept relationships from each of the target databases and present the extracted relationships. The proposed method and the aforementioned conventional methods are applied to the same image database, and the extracted concept relationships are depicted as networks by using NetDraw [92]. Because it is difficult to show the relationships of all concepts in the database, we randomly select 32 concepts from each database to extract the respective relationships. Moreover, strong similarities are chosen by thresholding for easier viewing.

Results of IAPR-TC12 dataset

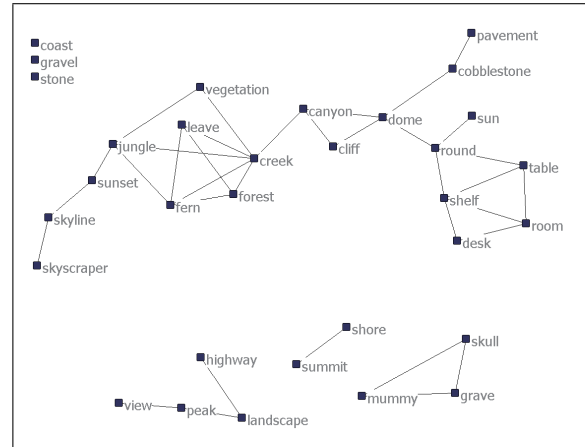
The extracted relationships are shown in Fig. 3.3, where each node represents a concept, and each edge between the concepts represents concept relevance. As shown in Fig. 3.3, although all networks are constructed from the same database, their concept relationships are different from one another. We can find the following from Fig. 3.3:

- The proposed method finds semantic similarity well, not reflecting only concept co-occurrences.
- Comparative method A extracts the concept pairs that tend to co-occur in the image database, such as “creek” - “canyon”.
- Comparative method B finds visually similar concept pairs. Because this method does not have a scheme that detects the features relevant to each concept, the similarities might be affected by some visual features derived from other concepts in images.
- Comparative method C finds relevant concept pairs by combining concept co-occurrences and visual similarities. This method can represent either co-occurrence probability or visual similarity.

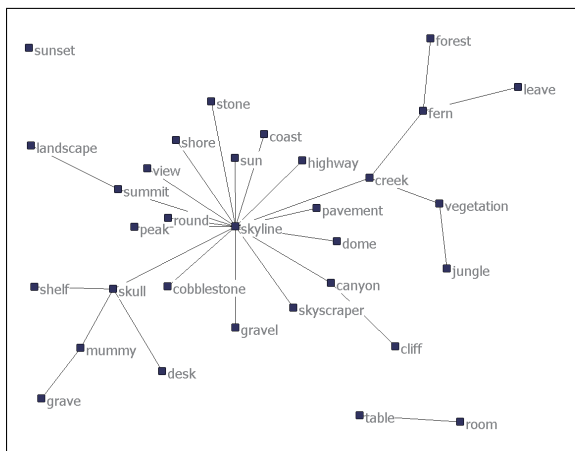
In contrast, we can find that the proposed method has the following weakness. From Fig. 3.3, the proposed method fails to describe the relationship corresponding to the



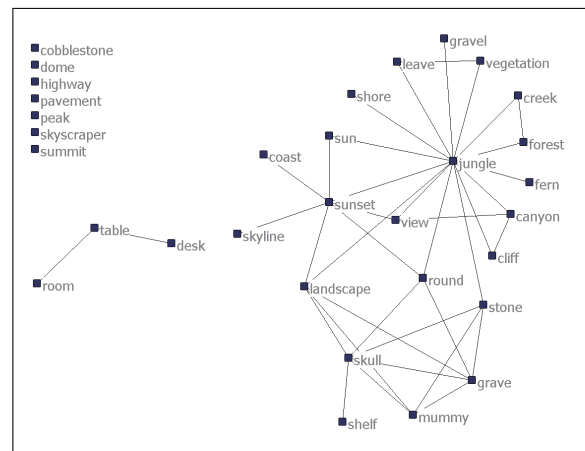
(a) Concept relationship extracted by the proposed method.



(b) Concept relationship extracted by Comparative method A.



(c) Concept relationship extracted by Comparative method B.



(d) Concept relationship extracted by Comparative method C.

Fig. 3.3 Concept relationships extracted from IAPR-TC12 dataset by proposed method and comparative methods A, B, and C.

concept “round.” In this experiment, the features used do not include shape information. Thus, if the features cannot model each concept well, the relationships around the concepts cannot be extracted. This weakness is common to other conventional methods that use visual similarities. For these concepts, the method that uses only concept co-occurrences [3] can extract the relationships better than the other methods can. In future works, we should investigate various types of visual features for overcoming this problem.

Results of Hokkaido landscape dataset

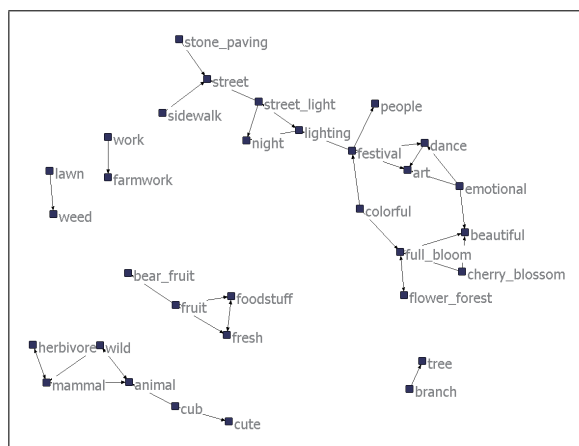
The extracted relationships are shown in Fig. 3.4, where each node represents a concept, and each edge between concepts represents concept relevance. Similar to the results shown in Fig. 3.3, all concept relationships shown in Fig. 3.4 are different from one another. We can find the following from Fig. 3.4:

- The proposed method finds the visual feature-based relationships that are derived from the same source, thus reflecting not only concept co-occurrences.
- Comparative method A extracts the co-occurrence relations well, but does not reflect the visual content in images. For example, “night” and “dance” are connected, but these seem to not be semantically similar.
- Comparative method B finds visually similar concept pairs; however, the extracted relationships might be strongly affected by the visual context information. For example, “full bloom” and “weed” are connected, but they are not from the same objects.
- Comparative method C finds relevant concept pairs by diffusing co-occurrence and visual similarity, but some relationships such as “work” - “animal” seem to not be semantically similar.

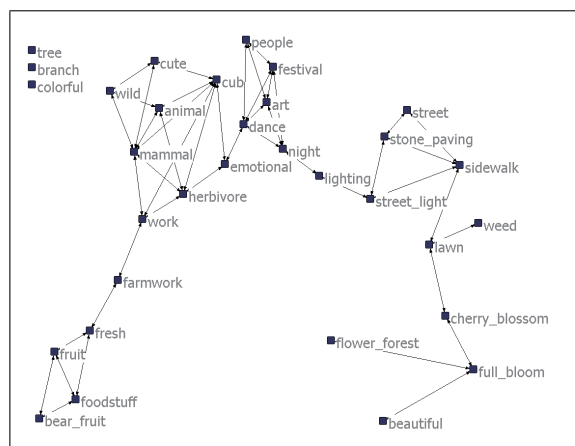
These results show trends similar to those of the IAPR-TC12 dataset tests.

3.4.2 User evaluation of extracted relationships

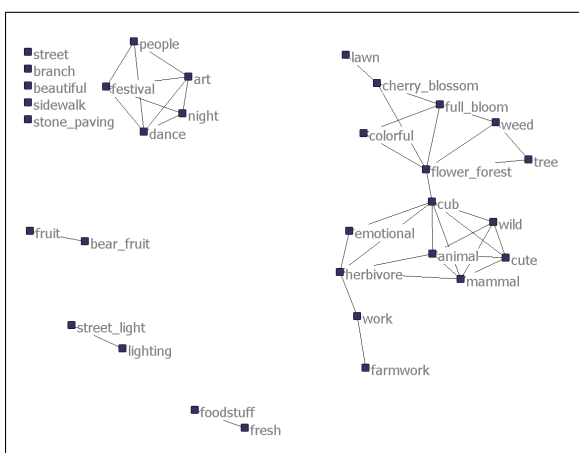
The strong relationships extracted using each method are manually evaluated by users as well. We present 11 users with a list of 55 concept pairs having the highest relevance from each database. For each concept pair, the user is required to assign score of 1 - 5. Note that the higher the score, the stronger is the relationship between two concepts. By



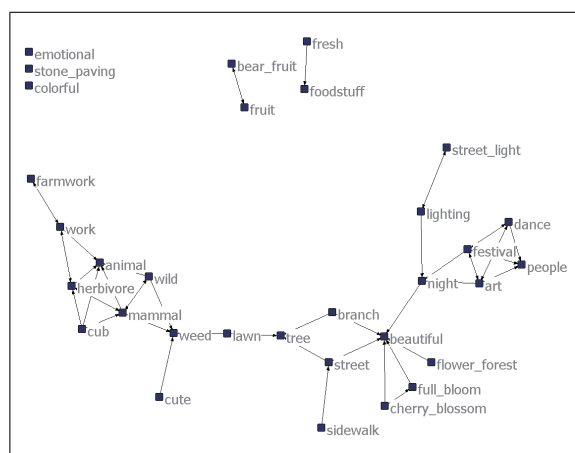
(a) Concept relationship extracted by the proposed method.



(b) Concept relationship extracted by Comparative method A.



(c) Concept relationship extracted by Comparative method B.



(d) Concept relationship extracted by Comparative method C.

Fig. 3.4 Concept relationships extracted from Hokkaido landscape dataset by proposed method and comparative method A, B, and C.

Table 3.3 User assessments of the IAPR-TC12 dataset.

Proposed method		Comparative method C	
Example of concept pair	Score	Example of concept pair	Score
cyclist - bicycle	4.81	lake - cloud	2.63
wave - sea	4.91	waterfall - rock	3.82
dune - beach	2.64	table - wall	3.45
desk - shelf	4.10	portrait - man	3.18
fern - jungle	3.55	kid - classroom	3.10
gravel - cobblestone	4.37	horizon - cloud	3.54
hill - ground	3.91	stadium - shirt	2.09
cloth - cape	2.64	jungle - stone	2.00
highway - skyscraper	3.18	wood - wall	2.10
shirt - pullover	4.10	lagoon - cloud	2.27
Average	3.17	Average	2.87

Table 3.4 User assessments of the Hokkaido landscape database.

Proposed method		Comparative method C	
Example of keyword pair	Score	Example of keyword pair	Score
waterside - river	4.64	herbivore - work	1.82
maple - autumn	3.64	crop - agriculture	4.88
sea - horizon	3.73	horse - lawn	3.55
tree - branch	4.91	lawn - tree	2.73
building - exterior	3.82	crop - bloom	2.45
crop - agriculture	4.88	snow - mountain	3.34
water flow - liquid	4.73	mountain - sunset	3.55
people - clothes	3.82	reflection - beautiful	3.00
lawn - weed	3.82	night - festival	3.55
stone - solid	4.18	active - morning	2.27
Average	4.24	Average	3.53

averaging the scores of all users, we obtain the final score for each concept pair; these final scores are listed in Tables 3.3 and 3.4. In the tables, some examples of concept pairs and their scores are listed, and scores at the bottom are the average scores of the 55 concept pairs. Based on these results, it can be said that the proposed method is better than the conventional method [8] at extracting the relationships that are always true. Specifically, the pairs extracted by the proposed method include hyponymy/hypernymy relationships such as “stone”-“solid” and meronymy relations such as “tree”-“branch”. This is because of the effective use of visual features and concept co-occurrence.

In the following subsection, we apply the concept relationships to automatic image annotation for quantitatively evaluating the concept relationship extraction methods.

3.4.3 Application to automatic image annotation

In this subsection, we apply the concept relationships to automatic image annotation in order to evaluate quantitatively the effectiveness of the proposed method. For introducing the semantic relationships between concepts into image annotation, the dual cross-media relevance model (DCMRM) [62] has been proposed. The DCMRM calculates an annotation score for each concept, and concepts that have the highest m values, as determined using Eq. (2.16), are provided as annotations on the tested image. Figures 3.5 and 3.6 show the tested images’ annotation results obtained using the proposed method and comparative method A; here, $m = 7$. The concepts provided by the proposed method are semantically correlated because concepts that have visual feature-based relevance enhance each other. However, we can find a weakness of the proposed method from Fig. 4. In the right-hand result of Fig. 4, the provision of concepts such as “square” and “plane” is incorrect. In fact, the proposed method incorrectly extracted the relationships of the shape-based concept “square,” which was connected with the concepts “cobblestone” and “dune.” As mentioned in the previous subsection, the proposed method cannot describe the relationships around concepts such as “round” and “square.” This weakness leads to poor annotation results, and we should therefore use more effective features to describe each concept.

Furthermore, the quality of automatic image annotation is measured through the process of retrieving testing images with a single concept. For each concept C_k , the number of correctly annotated images is denoted as $|A_k|$, number of retrieved images is denoted as $|S_k|$, and number of truly related images in the testing set is denoted as $|R_k|$. The

Testing images.				
				
tee-shirt, man, court, player, tennis, woman, stadium	flower, lamp, people, bench, tree, palm, building	gate, corridor, wall, brick, flower, pot, window	port, sky, cloud, boat, building, ship, view	cobblestone, dune, square, range, plane, lookout, bell, highway
The top seven concepts provided by Comparative method A.				
woman, court, fence, people, meadow, tennis, flag	people, man, woman, tree, flower, lamp, sky	window, gate, sea, building, flag, fence, rock	airport, sky, cloud, house, man, car, sea	cobblestone, dune, rock, hill, mountain, road, car

Fig. 3.5 Image annotation results of proposed method and comparative method A for the IAPR-TC12 dataset. The top seven concepts are provided.

precision, recall, and F-measure are computed as follows:

$$\text{precision}(C_k) = \frac{|A_k|}{|S_k|}, \quad (3.8)$$

$$\text{recall}(C_k) = \frac{|A_k|}{|R_k|}, \quad (3.9)$$

$$\text{F-measure}(C_k) = \frac{2 \times \text{precision}(C_k) \times \text{recall}(C_k)}{\text{precision}(C_k) + \text{recall}(C_k)}, \quad (3.10)$$

We calculate the average precision, recall, and F-measure over the concepts that exist in the testing image for performance evaluation. Here, in order to show the effectiveness of the concept relationship, we also compare the performance with the following methods:

- **Comparative method D:** image annotation using non-linear SVM with the RBF kernel.
- **Comparative method E:** the cross-media relevance model (CMRM) [7], which annotates images using labels of training images that are visually similar to a given

Table 3.5 Precision, recall, and F-measure of the conventional methods [3, 4, 7, 8] and the proposed method for the IAPR-TC12 dataset, in which $m = 5$.

Method	Precision	Recall	F-measure
Proposed method	0.206	0.183	0.194
Comparative method A	0.180	0.159	0.169
Comparative method B	0.140	0.158	0.148
Comparative method C	0.160	0.162	0.161
Comparative method D	0.180	0.177	0.179
Comparative method E	0.147	0.146	0.146

testing image.

Performance comparisons are summarized in Tables 3.5 and 3.6, in which $m = 5$ and $m = 12$, respectively. From these tables, it can be inferred that our concept relationship extraction method performs better than comparative methods A, B, and C in terms of image annotation. Compared with comparative method C, which extracts context information separately from concept co-occurrences and visual similarities, and combines the extracted information [8], our method offers better performance owing to the simultaneous use of context information and visual similarities. Moreover, when comparing the proposed method with comparative method D, it becomes clear that concept relationships are effective for image annotation. The precisions with different values of m are shown in Figs. 3.7 and 3.8. These figures show that for each value of m , the image annotation precision achieved using the proposed method is better than that achieved using comparative methods. From these experimental results, we can conclude that the proposed method can extract suitable concept relationships for image annotation by focusing on visual feature-based relevance. Even if we use LR, which is a simple discriminative classifier, the proposed framework can effectively extract concept relationships. It is expected that if non-linear classifiers including SVM are used in the proposed framework, we can improve its concept-relationship-extraction performance. We should discuss the possibility of improvement by comparing classifiers, which is our future work.

Table 3.6 Precision, recall, and F-measure of the conventional methods [3, 4, 7, 8] and the proposed method for the Hokkaido landscape database, in which $m = 12$.

Method	Precision	Recall	F-measure
Proposed method	0.289	0.258	0.273
Comparative method A	0.206	0.179	0.192
Comparative method B	0.169	0.107	0.131
Comparative method C	0.245	0.242	0.243
Comparative method D	0.151	0.215	0.177
Comparative method E	0.112	0.143	0.126

3.5 Conclusion and Future Work

Conclusion

This chapter presents a classifier-based approach that uses concept co-occurrence and visual features for extracting inter-concept relationships from an image database. In the proposed approach, we first detect visual features related to each concept through classifier training. Then, to measure the semantic relationship between concepts, we remove the detected features from the training images and monitor the performance of the other concepts. This is the biggest departure of our method from the conventional approaches.

The proposed framework needs a classifier that can (i) detect visual features relevant to each concept and (ii) monitor performance after training. To satisfy these requirements, we use LR in this work. Results of the experiments conducted using image databases show that our method can analyze concept relationships by using visual features. We applied the extracted concept relationships to automatic image annotation in the experiments. The obtained results show that our method has better image annotation performance than the conventional methods.

Future work

We find from the experimental results that more effective features are needed for detecting the features related to each concept. We will investigate various types of feature extraction methods for the proposed approach to improve its performance. In addition, two types of verification protocols are required in future works. First, we should consider the relationship between the performance of the proposed method and the number of training images. Apart from experiments in Section 3.4, we experimentally reduced the

number of training samples and evaluated image annotation performance. In the IAPR-TC12 dataset, when reducing the number of training samples from 17665 to 10000, the F-measure at $m = 5$ decreased to 0.181. In the Hokkaido landscape database, when reducing the number of training samples from 20000 to 10000, the F-measure at $m = 12$ decreased to 0.258. It is considered that such poor performance is caused by insufficient training of the LR. These results suggest that we should investigate the influence of the number of training images on the performance of the method. In relation to this, as mentioned in Section 3.3.4, in future works, we should investigate various classifiers for determining their suitability for the proposed framework.

Second, although the parameters of LR show the degree of relevance of the corresponding features, the proposed method does not effectively use their coefficients: features that correspond to negative parameters are simply discarded. If the absolute values of the parameters are used effectively, the performance of the proposed method will improve. The effective use of classifier parameters will be studied in addition to an investigation of suitable classifiers.

Third, we should investigate those concepts that the proposed method does not work well with owing to their visual diversity. In the IAPR-TC12 dataset, the most visually diverse concepts were “area,” “bit” and “lot,” and the most visually representative concepts were “sky,” “tree,” and “mountain.” In the Hokkaido landscape dataset, the most visually diverse concepts were “picnic,” “curve,” and “transparency,” and the most visually representative concepts were “cloud,” “mountain,” and “forest.” In fact, the annotation performance with the visually diverse concepts tends to be lower than that with the visually representative concepts, even when the proposed method is applied. Therefore, we believe that the visually diverse concepts should be separated from the other concepts, and some special procedures should be developed for extracting relationships for those concepts. We will conduct these experiments in detail and improve the proposed framework in future works. Motivated by this fact, in Chapter 4, we present an approach for quantifying the abstraction levels of concepts in a target image database.

Limitations of proposed method

In this chapter, we assume the use of an image database that is accurately labeled, and using this database, we train a classifier for each concept. In experiments conducted under this condition, we showed the effectiveness of using both the visual features and concept co-occurrence. This is an important finding obtained by our approach. However, if we use a dataset collected from social media websites, many labels (tags) might be irrel-

evant to a corresponding image, and the performance of the supervised approach will deteriorate. In such a situation, we need to apply techniques that remove noisy tags from the images to improve the tag quality [93]. In the following chapter, we use Web images for extracting inter-concept relationships, which models the probabilistic distributions of concepts considering the noisy tags as contextual information.

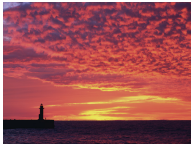
Testing images.				
				
The top seven keywords provided by the proposed method.				
farm, animal, work, pasture, herbivore, grazing, crowd	people, art, cloth, dance, team, building, festival	water, flow, liquid, river, people, animal, waterside	morning, sun, morning glow, silhouette, romantic, back-lighting, cloud	sea, illumination, night, mountain, city, romantic, building
The top seven keywords provided by Comparative method A.				
mammal, animal, fence, work, herbivore, wall, emotional	life, evening, people, reflection, dark, summer, shade	river, water, liquid, flow, solid, waterside, rock	evening, silhouette, reflection, bright, dazzling, morning, bright	romantic, night, reflection, tower, building, harbor, dark
Testing images.				
				
The top seven keywords provided by the proposed method.				
food, seafood, leaf, maple, fresh, waterside, animal	fallen leaves, maple, sun, leaf, autumn, street, evening	farm, flower forest, agriculture, full bloom, flower, gradation, forest	lighting, art, night, people, building, flickering, street	food, fruit, forest, animal, fresh, bloom, wild
The top seven keywords provided by Comparative method A.				
morning, people, autumn, solid, gradation, picnic, maple	evening, reflection, leaf, gradation, shadow, dark, tree	blue sky, agriculture, cloud, mountain, hill, forest, tree	evening, people, reflection, tree, bright, building, dark	land, summer, weed, tree, grass, agriculture, leaf

Fig. 3.6 Image annotation results of proposed method and comparative method A for the Hokkaido landscape database. The top seven keywords are provided.

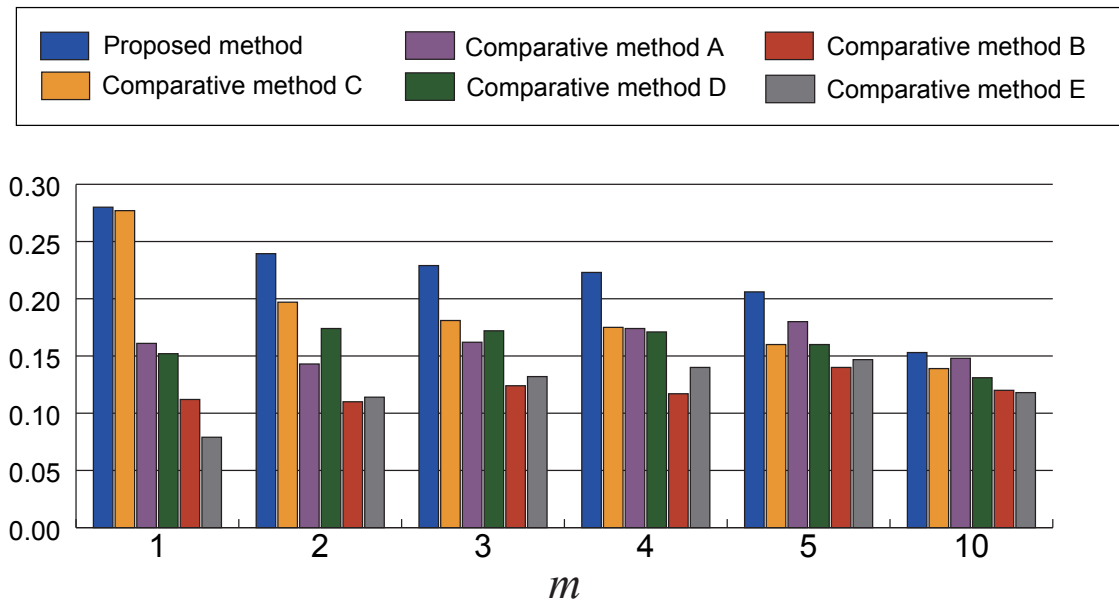


Fig. 3.7 Average precision ($m = 1, 2, 3, 4, 5$, and 10) on testing images of the IAPR-TC12 dataset.

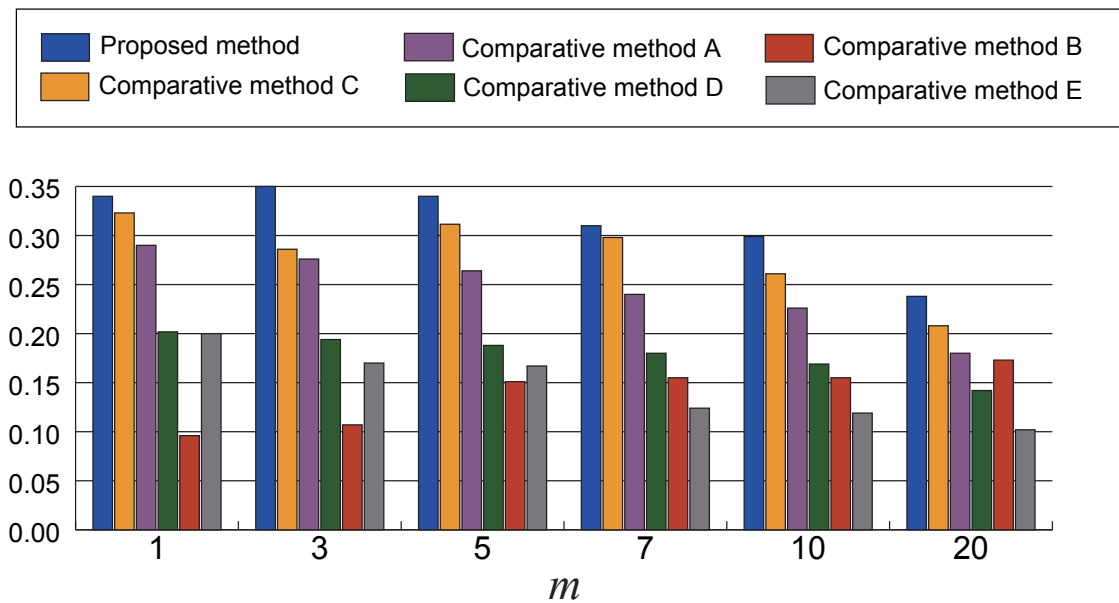


Fig. 3.8 Average precision ($m = 1, 3, 5, 7, 10$, and 20) on testing images of the Hokkaido landscape dataset.

Chapter 4

A Cross-Modal Approach for Extracting Inter-Concept Relationships From Tagged Images

4.1 Introduction

Extraction of semantic relationships between concepts has attracted much attention in several research fields, including the fields of natural language processing, information retrieval, and multimedia processing. Much interest has been shown in the numerical determination of the following two measures: the semantic relatedness between concepts and the abstraction level of each concept. These measures have been used to improve the performance of the computational techniques, such as query expansion in information retrieval [94], document clustering [95], and multimedia annotation and retrieval [32, 33, 64, 67, 68, 96, 97].

We have presented an approach for determining semantic relations between concepts using tagged images in Chapter 3, and showed its effectiveness compared with conventional measures. However, our previous approach has two limitations: (i) the previous approach cannot extract semantic relations beyond hyponymy/hypernymy relations, i.e., it is difficult to calculate the semantic relatedness or abstraction levels of concepts; (ii) the previous approach cannot deal with images that are not accurately tagged because it is based on supervised training of classifiers. To solve this problem and extend the framework to extract other semantic relations, we should consider the probabilistic distribution that encompass the statistics of tagging tendency.

The measures for extracting the semantic relatedness or abstraction levels of concepts, which has been described in Chapter 2, use only a single modality. That is, they use either text or visual features. As a result, there could be a lack of information: methods that use only text features cannot capture visual similarity [4], while methods that use only visual features cannot effectively show the properties of abstract concepts due to the well-known semantic gap [17]. Thus, we argue that both of these modalities are essential to capturing the semantics of the concepts and that they act in a complementary way. If not single but multiple types of modalities are used effectively, we can extract concept relationships that are coherent with human cognition.

In this chapter, a cross-modal approach for extracting semantic relationships between concepts using tagged images on Flickr is presented. In the proposed method, we first project two feature modalities to a common subspace using canonical correlation analysis (CCA) [98]. CCA is a well-known technique for finding correlations between two sets of multidimensional variables. Specifically, under a probabilistic interpretation of CCA [2], low-dimensional representations of a sample can be obtained in terms of probabilistic distributions of latent variables. The latent variables can be considered “semantic topics” that connect text features and visual features. Our approach aims to discover the underlying semantic topics from the feature pairs and use them to extract semantic relationships of concepts. By considering the probabilistic interpretation of CCA, the semantic context of each tagged image can be represented as a distribution of the latent variables.

The rest of this chapter is as follows: In Section 4.2, an overview of canonical correlation analysis (CCA) and its probabilistic interpretation are given. In Section 4.3, a cross-modal approach to extract semantic relations between concepts using tagged images on Flickr is presented. In Section 4.4, the results of the experiments on verifying the effectiveness of the proposed method are presented. Finally, this chapter is summarized, and some possible directions for future work are suggested in Section 4.5.

4.2 Canonical Correlation Analysis and Its Probabilistic Interpretation

This section first describes an overview of canonical correlation analysis (CCA) [98], and then introduce its probabilistic interpretation [2].

4.2.1 An overview of canonical correlation analysis

CCA is a well-known technique for finding correlations between two sets of multidimensional variables [98]. Given a set of feature pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are the d_x - and d_y -dimensional feature vectors obtained from the i -th sample of the set, CCA searches for the linear transformations such that This essentially reduce the dimensionality of the features $\mathbf{x}_i$ and $\mathbf{y}_i$ to q ($1 \leq q \leq \min\{d_x, d_y\}$). Let $\mathbf{C} = \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xy} \\ \mathbf{C}_{yx} & \mathbf{C}_{yy} \end{pmatrix}$ be the $d \times d$ (where $d = d_x + d_y$) sample covariance matrix obtained from the dataset. Each block is defined as:

$$\mathbf{C}_{xx} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu}_x)(\mathbf{x}_i - \boldsymbol{\mu}_x)^T, \quad (4.1)$$

$$\mathbf{C}_{yy} = \frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i - \boldsymbol{\mu}_y)(\mathbf{y}_i - \boldsymbol{\mu}_y)^T, \quad (4.2)$$

$$\mathbf{C}_{xy} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu}_x)(\mathbf{y}_i - \boldsymbol{\mu}_y)^T, \quad (4.3)$$

$$\mathbf{C}_{yx} = \mathbf{C}_{xy}^T. \quad (4.4)$$

In the above equations, $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$ are the sample means of features. Using the covariance matrix, the projection matrices $\mathbf{U}_x$ and $\mathbf{U}_y$ can be computed as solutions of the following generalized eigenvalue problems:

$$\mathbf{C}_{xy}\mathbf{C}_{yy}^{-1}\mathbf{C}_{yx}\mathbf{U}_x = \mathbf{C}_{xx}\mathbf{U}_x\Lambda^2, \quad (4.5)$$

$$\mathbf{C}_{yx}\mathbf{C}_{xx}^{-1}\mathbf{C}_{xy}\mathbf{U}_y = \mathbf{C}_{yy}\mathbf{U}_y\Lambda^2, \quad (4.6)$$

where Λ is the diagonal matrix of the first q canonical correlations. CCA has been effectively used to find correlations between visual features and text features of images [58], and its performance for image annotation is better than the separate use of these features.

4.2.2 Probabilistic interpretation of canonical correlation analysis

The authors in [2] presents a probabilistic interpretation of CCA, which is derived from latent probabilistic modeling. The probabilistic interpretation of CCA is analogous to the probabilistic interpretation of PCA [99]. In [2], a graphical model is considered as shown

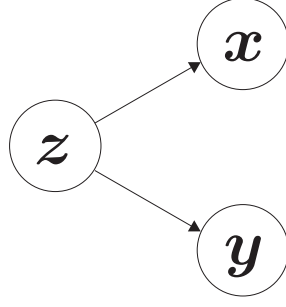


Fig. 4.1 A graphical model for canonical correlation analysis [2].

in Fig. 4.1, which can be written as:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q), \quad 1 \leq q \leq \min\{d_x, d_y\}, \quad (4.7)$$

$$\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}_x \mathbf{z} + \boldsymbol{\mu}_x, \Psi_x), \quad \mathbf{W}_x \in \mathbb{R}^{d_x \times q}, \Psi_x \geq 0, \quad (4.8)$$

$$\mathbf{y}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}_y \mathbf{z} + \boldsymbol{\mu}_y, \Psi_y), \quad \mathbf{W}_y \in \mathbb{R}^{d_y \times q}, \Psi_y \geq 0, \quad (4.9)$$

where $\mathcal{N}$ is a Gaussian, $\Psi_x \geq 0$ and $\Psi_y \geq 0$ indicate that Ψ_x and Ψ_y are positive semi-definite matrices, $\mathbf{z}$ is an unobserved latent variable. In the above equations, Ψ_x , Ψ_y , $\mathbf{W}_x$, $\mathbf{W}_y$ are defined as follows:

$$\mathbf{W}_x = \mathbf{C}_{xx} \mathbf{U}_x \mathbf{M}_x, \quad (4.10)$$

$$\mathbf{W}_y = \mathbf{C}_{yy} \mathbf{U}_y \mathbf{M}_y, \quad (4.11)$$

$$\Psi_x = \mathbf{C}_{xx} - \mathbf{W}_x \mathbf{W}_x^T, \quad (4.12)$$

$$\Psi_y = \mathbf{C}_{yy} - \mathbf{W}_y \mathbf{W}_y^T, \quad (4.13)$$

where $\mathbf{M}_x$ and $\mathbf{M}_y$ are arbitrary $(q \times q)$ matrices with spectral norms that are smaller than one, such that $\mathbf{M}_x \mathbf{M}_y = \Lambda$. These equations have a direct link to the CCA solution.

The probabilistic model can provide a better understanding of the quantities that are shared by the two features. In particular, it is shown in [2] that the posterior probability of the latent variables $\mathbf{z}$ given $\mathbf{x}_i$ and $\mathbf{y}_i$ follows a normal distribution whose mean and

variance are calculated using the following equations:

$$\begin{aligned} \mathbf{m}_i &= E(\mathbf{z}|\mathbf{x}_i, \mathbf{y}_i) \\ &= \begin{pmatrix} \mathbf{M}_x \\ \mathbf{M}_y \end{pmatrix}^T \begin{pmatrix} (\mathbf{I}_q - \Lambda^2)^{-1} & -(\mathbf{I}_q - \Lambda^2)^{-1} \Lambda \\ -(\mathbf{I}_q - \Lambda^2)^{-1} \Lambda & (\mathbf{I}_q - \Lambda^2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{U}_x^T(\mathbf{x}_i - \bar{\mathbf{x}}) \\ \mathbf{U}_y^T(\mathbf{y}_i - \bar{\mathbf{y}}) \end{pmatrix}, \end{aligned} \quad (4.14)$$

$$\begin{aligned} \mathbf{S}_i &= var(\mathbf{z}|\mathbf{x}_i, \mathbf{y}_i) \\ &= \mathbf{I}_q - \begin{pmatrix} \mathbf{M}_x \\ \mathbf{M}_y \end{pmatrix}^T \begin{pmatrix} (\mathbf{I}_q - \Lambda^2)^{-1} & -(\mathbf{I}_q - \Lambda^2)^{-1} \Lambda \\ -(\mathbf{I}_q - \Lambda^2)^{-1} \Lambda & (\mathbf{I}_q - \Lambda^2)^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{M}_x \\ \mathbf{M}_y \end{pmatrix}, \end{aligned} \quad (4.15)$$

where $\mathbf{I}_q$ indicates a q -dimensional identity matrix, $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are sample means of $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{y}_i\}_{i=1}^N$, respectively. In our implementation, we set $\mathbf{M}_x = \mathbf{M}_y = \Lambda^{1/2}$.

The probabilistic structure of CCA has been used in many applications because it can provide posterior expectations of the latent variables that lie in the subspace found by CCA [57, 100]. For example, in [100], the probabilistic interpretation of CCA is used to introduce the correlations of two features into a probabilistic visual tracking model. Their experiments show that a low-dimensional representation derived from multiple features can improve the application rather than a low-dimensional representation of single features. In [57], the posterior distributions of latent variables are compared between two given images for image content recognition. If we focus on a set of samples of a concept under a probabilistic interpretation of CCA, then the underlying contexts of the concept in two modalities can be represented as a distribution of the latent variables.

4.3 A CCA-based Approach for Extracting Semantic Relations Between Concepts

In this section, we propose a cross-modal approach for extracting semantic relations between concepts using tagged images on Flickr. An overview of the proposed method is shown in Fig. 4.2. First, given a tagged image I_i ($i = 1, 2, \dots, N$, where N is the number of images), we extract its d_x -dimensional visual features $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,d_x}]^T$ and d_y -dimensional text features $\mathbf{y}_i = [y_{i,1}, y_{i,2}, \dots, y_{i,d_y}]^T$, as shown in Fig. 4.2 (a). A pair $\{\mathbf{x}_i, \mathbf{y}_i\}$ is considered a sample, and the set of samples is represented by $R = \{\mathbf{x}_i, \mathbf{y}_i | i = 1, 2, \dots, N\}$. To account for the implied correlations between the text features and the visual features, we apply CCA to the set R and obtain the projection matrices, as described in the previous section. Specifically, we calculate the covariance matrices using the text features

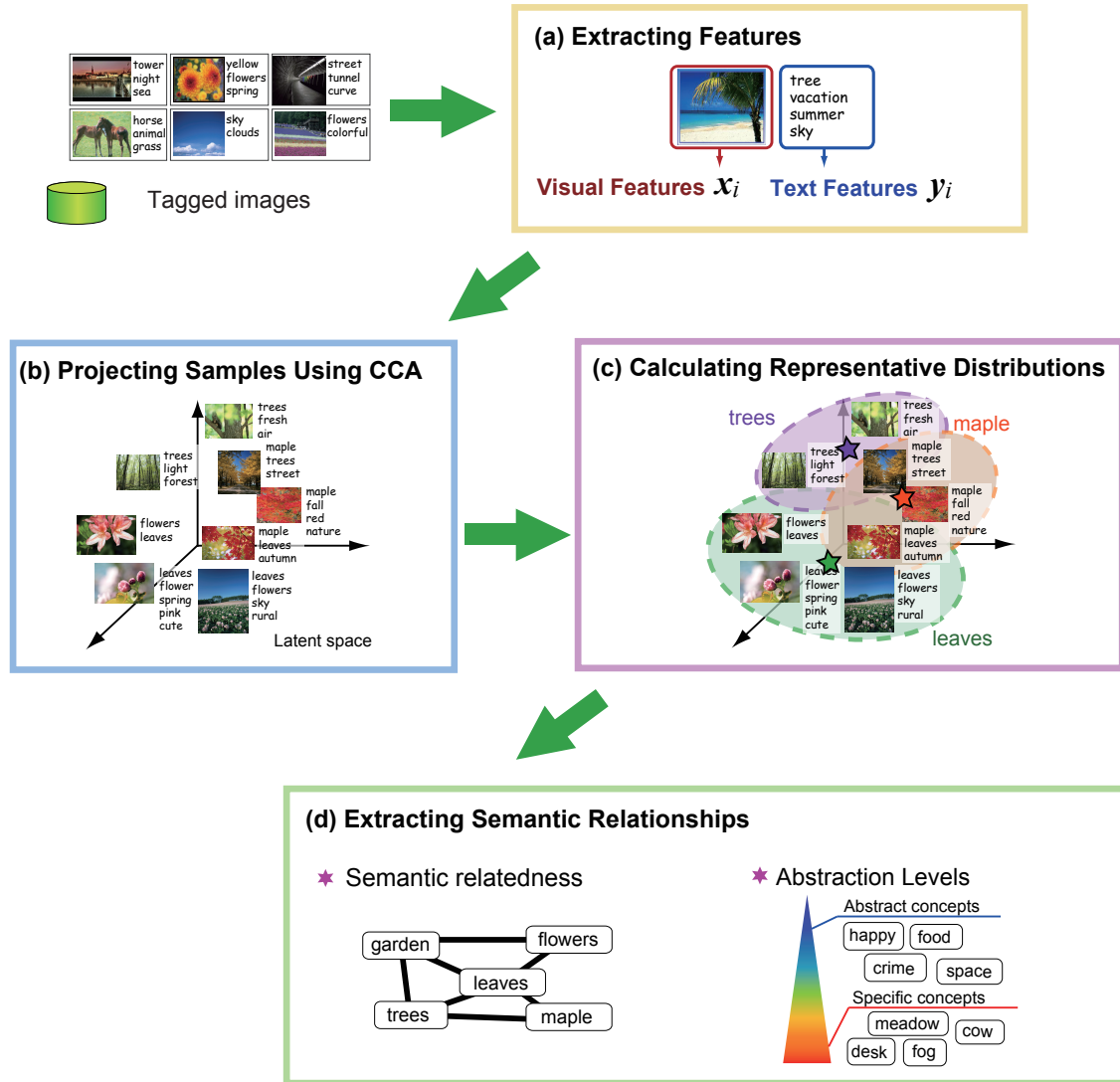


Fig. 4.2 The proposed steps that are taken for measuring concept relationships: (a) feature extraction from tagged images, (b) sample projection using CCA, (c) calculation of representative distributions in the latent space, and (d) semantic relationship extraction.

and the visual features that are extracted from all samples, in which the computational complexity is $O(N(d_x^2 + d_y^2 + d_x d_y))$. Second, we solved the eigenvalue problems of Eq. (4.5) and Eq. (4.6), in which the computational complexity is $O(d_x^3 + d_y^3 + d_x^2 d_y + d_x d_y^2)$. All feature pairs are projected to a latent space using CCA, which corresponds to Fig. 4.2 (b). To summarize the semantic topics for each concept, we obtain a representative distribution of the latent variables for each concept, as shown in Fig. 4.2 (c). Last, based on the representative distributions of the concepts, we present two types of measures, as shown in Fig. 4.2 (d): the semantic relatedness between the concepts and the abstraction level of each concept. In the remainder of this section, the two steps that correspond to Figs. 4.2 (c) and 4.2 (d) are detailed in 4.3.1 and 4.3.2, respectively.

4.3.1 Calculation of the representative distribution of a concept

In this subsection, we describe how to calculate the representative distribution of a concept in the latent space found by CCA.

Under the probabilistic interpretation of CCA, the i -th sample $\{\mathbf{x}_i, \mathbf{y}_i\}$ of R can be represented as a normal distribution, in which the parameters are calculated using Eqs. (3) and (4). To consider how concept C_k is distributed in the latent space, we approximate a set of normal distributions $\{p(\mathbf{z}|\mathbf{m}_i, \mathbf{S}_i), i \in \tau_k\}$ with one normal distribution that is parameterized by $\{\boldsymbol{\mu}_k, \Sigma_k\}$. We consider this task as the problem of finding a representative distribution of a set of normal distributions. The problem of finding the representative distribution of concept C_k can be written as:

$$p(\mathbf{z}|\boldsymbol{\mu}_k^*, \Sigma_k^*) = \arg \min_{p(\mathbf{z}|\boldsymbol{\mu}_k, \Sigma_k)} = \frac{1}{|\tau_k|} \sum_{i \in \tau_k} D(p(\mathbf{z}|\mathbf{m}_i, \mathbf{S}_i) || p(\mathbf{z}|\boldsymbol{\mu}_k, \Sigma_k)), \quad (4.16)$$

where $|\tau_k|$ is the cardinality of τ_k and $D(\cdot || \cdot)$ is the distance measure between two normal distributions. In this paper, we use the Kullback-Leibler (KL) divergence (also referred to as the relative entropy) [101] as the distance measure D . The KL divergence between two normal distributions $p(\mathbf{z}|\mathbf{m}, \mathbf{S})$ and $p(\mathbf{z}|\boldsymbol{\mu}, \Sigma)$ can be written as [102]:

$$\begin{aligned} & D_{KL}(p(\mathbf{z}|\mathbf{m}, \mathbf{S}) || p(\mathbf{z}|\boldsymbol{\mu}, \Sigma)) \\ &= \frac{1}{2} \left\{ -d + \text{tr}(\mathbf{S}\Sigma^{-1}) - \log|\mathbf{S}\Sigma^{-1}| + (\mathbf{m} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{m} - \boldsymbol{\mu}) \right\} \\ &= \frac{1}{2} B(\mathbf{S}, \Sigma) + \frac{1}{2} M_{\Sigma^{-1}}(\mathbf{m}, \boldsymbol{\mu}) \end{aligned} \quad (4.17)$$

where

$$B(\mathbf{S}, \Sigma) = \text{tr}(\mathbf{S}\Sigma^{-1}) - \log|\mathbf{S}\Sigma^{-1}| - d$$

is the Burg matrix divergence and

$$M_{\Sigma^{-1}}(\mathbf{m}, \boldsymbol{\mu}) = (\mathbf{m} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{m} - \boldsymbol{\mu})$$

is the Mahalanobis distance, which is parameterized by the covariance matrix Σ . Using this equation, the problem of finding the optimal representative distribution in set $\{p(\mathbf{z}|\mathbf{m}_i, \mathbf{S}_i), i \in \tau_k\}$ can be rewritten as follows [102]:

$$p(\mathbf{z}|\boldsymbol{\mu}_k^*, \Sigma_k^*) = \arg \min_{p(\mathbf{z}|\boldsymbol{\mu}_k, \Sigma_k)} \left\{ \frac{1}{|\tau_k|} \sum_{i \in \tau_k} \frac{1}{2} B(\mathbf{S}_i, \Sigma_k) + \frac{1}{|\tau_k|} \sum_{i \in \tau_k} \frac{1}{2} M_{\Sigma_k^{-1}}(\mathbf{m}_i, \boldsymbol{\mu}_k) \right\}. \quad (4.18)$$

Minimizing the second term of Eq. (4.18) can be viewed as the minimization of the Bregman information [103] with respect to a fixed Mahalanobis distance. According to [103], this computation has a unique minimizer, which is expressed as follows:

$$\boldsymbol{\mu}_k^* = \frac{1}{|\tau_k|} \sum_{i \in \tau_k} \mathbf{m}_i. \quad (4.19)$$

After substituting Eq. (4.19) in Eq. (4.18), the equation is strictly convex in Σ_k^{-1} . By setting the gradient of Eq. (4.18) with respect to Σ_k^{-1} to zero, we can derive the optimal Σ_k^* as follows [102]:

$$\Sigma_k^* = \frac{1}{|\tau_k|} \sum_{i \in \tau_k} \left\{ \mathbf{S}_i + (\mathbf{m}_i - \boldsymbol{\mu}_k^*)(\mathbf{m}_i - \boldsymbol{\mu}_k^*)^T \right\}. \quad (4.20)$$

Thus, the representative distribution for concept C_k can be represented by a normal distribution that is parameterized by $\boldsymbol{\mu}_k^*$ and Σ_k^* . We store all representative distributions as prototypes of the concepts and use them to measure the semantic relationships of the concepts. By finding the representative distribution for each concept, we can capture the semantic meanings of the concept. Furthermore, this modeling of a concept has the advantage of being able to focus not only on the mean but also on the covariance information of the latent variables. If the covariance information is also considered, then the performance of measuring the distance between the sample sets of two concepts will be improved [104, 105]. This computation enables us to capture the similarities in the

diversities of the semantic contexts between two concepts.

4.3.2 Measures of semantic relationships

In this subsection, we present two types of measures of semantic relationships, i.e., the semantic relatedness between two concepts and the abstraction level of each concept, which are based on the representative distributions of the concepts. First, we present a measure of the semantic relatedness between two concepts using the following hypothesis:

Hypothesis 1. (semantic relatedness)

If two concepts are semantically correlated, then their corresponding samples are closely distributed in the latent space.

An overview of Hypothesis 1 is shown in Fig. 4.3. Using Fig. 4.3 as an example, the sample distributions of semantically correlated concepts are intuitively similar, whereas the sample distributions of unrelated concepts differ from one another. To evaluate this distributional overlap, the proposed method computes the distance between the representative distributions of the two target concepts $p(\mathbf{z}|\boldsymbol{\mu}_k^*, \Sigma_k^*)$ and $p(\mathbf{z}|\boldsymbol{\mu}_j^*, \Sigma_j^*)$. Several methods can be used to calculate the distance between two probabilistic distributions. In this paper, we investigate the characteristics of the following three distance measures: the Euclidean distance, J-divergence and the Hellinger distance. As the first measure, we calculate the Euclidean distance between the means of two distributions, as follows [106]:

$$dist_E(C_k, C_j) = \|\boldsymbol{\mu}_k^* - \boldsymbol{\mu}_j^*\|^2. \quad (4.21)$$

This equation does not employ the covariances of the distributions. The second choice of measure is J-divergence, which is also referred to as the symmetric KL divergence. Using J-divergence, an overlap between two representative distributions can be measured as follows:

$$dist_J(C_k, C_j) = \frac{1}{2} \left\{ D_{KL}(p(\mathbf{z}|\boldsymbol{\mu}_k^*, \Sigma_k^*) || p(\mathbf{z}|\boldsymbol{\mu}_j^*, \Sigma_j^*)) + D_{KL}(p(\mathbf{z}|\boldsymbol{\mu}_j^*, \Sigma_j^*) || p(\mathbf{z}|\boldsymbol{\mu}_k^*, \Sigma_k^*)) \right\}. \quad (4.22)$$

J-divergence does not satisfy the triangle inequality of distance, and therefore, it is not a true metric. It has been noted that measures that do not have metric properties, especially metrics that do not satisfy the triangle inequality, cannot accurately calculate the

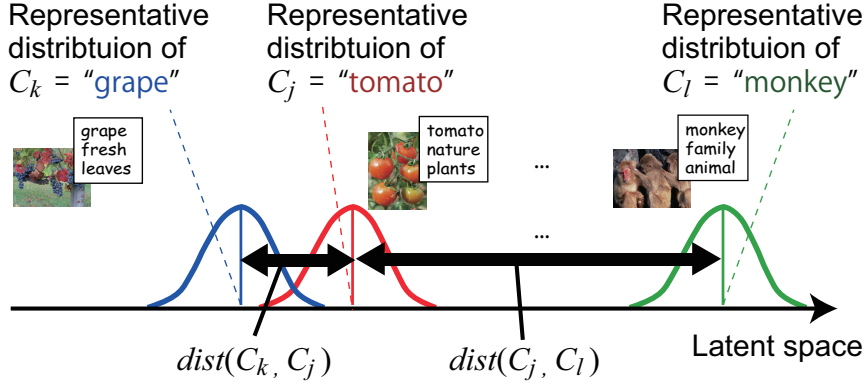


Fig. 4.3 An illustration of Hypothesis 1 (semantic relatedness), which represents how semantically similar two concepts are. This measure is based on the distance between their representative distributions.

relative dissimilarity among all members in a feature space; thus, they may degrade the performance of a system compared with the true metrics [107]. Therefore, as the third choice of distance measure, this paper newly exploits the Hellinger distance, which satisfies the triangle inequality of distance [108]. The Hellinger distance is also a symmetric measure between two probabilistic distributions and can be derived from the Bhattacharyya coefficient. The closed form of the Bhattacharyya coefficient $\rho(C_k, C_j)$ between the representative distributions $p(\mathbf{z}|\boldsymbol{\mu}_k^*, \Sigma_k^*)$ and $p(\mathbf{z}|\boldsymbol{\mu}_j^*, \Sigma_j^*)$ can be computed as:

$$\rho(C_k, C_j) = \frac{|\Sigma_k^*|^{\frac{1}{4}} |\Sigma_j^*|^{\frac{1}{4}}}{|\frac{1}{2}\Sigma_k^* + \frac{1}{2}\Sigma_j^*|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{8} \mathbf{u}_{k,j}^T \left(\frac{1}{2}\Sigma_k^* + \frac{1}{2}\Sigma_j^* \right)^{-1} \mathbf{u}_{k,j} \right\}, \quad (4.23)$$

where $\mathbf{u}_{k,j} = \boldsymbol{\mu}_k^* - \boldsymbol{\mu}_j^*$. Using the Bhattacharyya coefficient and Eq. (4.24), we compute the Hellinger distance between $p(\mathbf{z}|\boldsymbol{\mu}_k^*, \Sigma_k^*)$ and $p(\mathbf{z}|\boldsymbol{\mu}_j^*, \Sigma_j^*)$ as the semantic distance between the two concepts C_k and C_j .

$$dist_H(C_k, C_j) = \sqrt{1 - \rho(C_k, C_j)}. \quad (4.24)$$

where $0 \leq dist_H(C_k, C_j) \leq 1$. If these two concepts exhibit strong relatedness, then $dist_H(C_k, C_j)$ approaches 0. Using the Hellinger distance, we can simultaneously consider the differences in the mean vectors and the covariance matrices for two representative distributions [109]. We investigate whether the metric characteristics of the Hellinger distance can improve our framework compared with the other measures by conducting a subjec-

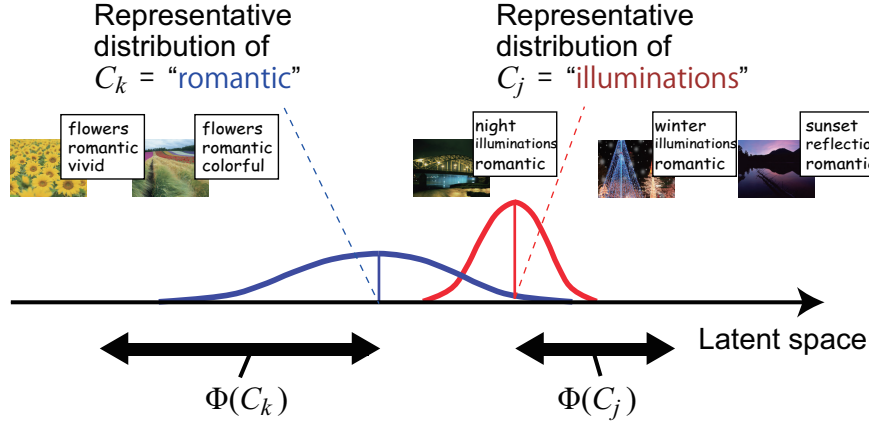


Fig. 4.4 An illustration of Hypothesis 2 (abstraction levels of concepts), which represents whether a target concept subsumes several semantics or appears in a certain context. This measure is based on the entropy of the concept's representative distribution.

tive evaluation and using real-world applications in Section 4.4.

Next, to derive a measure for the abstraction level, we provide the following definition:

Hypothesis 2. (abstraction levels of concepts)

A concept C_k has a specific meaning if its samples are distributed close to each other.

In other words, samples that correspond to concepts that represent specific meanings possess similar semantic topics. An illustration of Hypothesis 2 is shown in Fig. 4.4. Abstract concepts can be found in diverse semantic contexts because they subsume several objects and scenes. Conversely, specific concepts should appear in certain contexts to ensure that their samples are semantically similar. To determine whether the samples are distributed closely to each other, we calculate the entropy of the representative distribution of concept C_k as follows:

$$\Phi(C_k) = \frac{1}{2} \{q + \log(2\pi)^q |\Sigma_k^*|\}, \quad (4.25)$$

where q denotes the dimensionality of the latent variable $\mathbf{z}$ in the probabilistic interpretation of CCA. A small $\Phi(C_k)$ signifies that the samples that correspond to concept C_k are close to the centroid of those samples, which signifies that they are semantically similar to one another. Conversely, a large $\Phi(C_k)$ indicates that concept C_k is more abstract because it encompasses several lower-level meanings. By monitoring the obtained scores, we can compare two concepts in terms of the abstraction levels and distinguish highly

Table 4.1 The computational complexities of the proposed method (Hellinger distance) and the conventional methods [4, 5] for each procedure.

Procedure	FD [4]	GMM [5]	Proposed method
Calculation of the distribution of each concept	$O(TKd_x^3n)$	$O(TKd_xn)$	$O(nq(q + d_x + d_y)) + O(nq^2)$
Calculation of the semantic relatedness between the concepts	$O(K^2d_x^3)$	$O(K^2d_x)$	$O(q^3)$

abstract concepts and highly specific concepts.

4.3.3 Computational complexity

If a new concept is provided, we can calculate its representative distribution in a latent space and measure a pairwise semantic relatedness between the new concept and each of the existing concepts. This methodology is similar to the conventional measures that exploit visual features [4, 5]. The computational complexities of the steps for each method are summarized in Table 4.1, in which d_x and d_y represents the dimensionalities of the visual features and the dimensionalities of the text features, respectively; q is the dimensionality after performing a CCA; n is the number of samples used for modeling each concept; K is the number of components of a model; and T is the number of iterations in the Expectation-Maximization (EM) algorithm. As shown in Table 4.1, the computational complexities of FD and GMM are dependent on the dimensionality of the visual features d_x , whereas the computational complexity of our method is dependent on the dimensionality q . It should be noted that the implementation of GMM was based on diagonal covariance matrix; for GMM with full covariance matrix, the computational complexities for (i) and (ii) are $O(TKd_x^3n)$ and $O(K^3d_x^3)$, respectively. Compared with the conventional measures that exploit visual features [4, 5], our method can be computationally efficient because the dimensionality of the latent space q is substantially smaller than the dimensionality of the visual features d_x . Furthermore, given a new sample of a target concept, our method can recursively update the parameters of the representative distribution for the concept, in which the computational complexity is $O(q(q + d_x + d_y)) + O(q^2)$.

Table 4.2 Statistics of the MIRFLICKR-1M dataset [9].

Number of images	1,000,000
Number of unique tags	862,115
Average number of tags per image	11
Maximum/minimum number of tags per image	75 / 0
Number of owners	43,164
Average number of images per owner	23
Maximum number of images per owner	4,723

4.4 Experiments

In this section, we present the experimental results to verify the effectiveness of the proposed method.

4.4.1 Experimental settings

Our experiments were conducted using the MIRFLICKR-1M dataset [9], which contains 1,000,000 tagged images collected from Flickr. The original MIR Flickr collection consists of 25,000 fully annotated images [110], and it was extended to the MIRFLICKR-1M dataset. The images in the dataset were selected according to their Flickr interestingness score¹, which was calculated using several factors such as where clickthroughs on the images are coming from, who comments on it and when, or who marks the image as a favorite. All of images are under the Creative Commons attribution licenses². Specifically, for each date in the period from March 21, 2007 to June 30, 2008, the most interesting images uploaded at that date were requested, from which then only the images with an attribution license were downloaded. Table 4.2 summarizes the statistics of the MIRFLICKR-1M dataset. Figure 4.5 illustrates the tag frequency distribution in the MIRFLICKR-1M dataset, from which we observed that the tag frequency follows a power law distribution. Figure 4.6 shows the distribution of the number of tags per image.

Because a ground truth set of concepts for extracting semantic relationships between concepts of an image collection is unavailable, several conventional studies have created a list of concepts that can be used in experiments [4, 5, 50, 111]. Similarly, a list of concepts used in our experiments was created as follows: We sorted tags in the MIRFLICKR-1M dataset in descending order by the number of tag frequencies. After removing stop

¹<https://www.flickr.com/explore/interesting/>

²<http://creativecommons.org/>

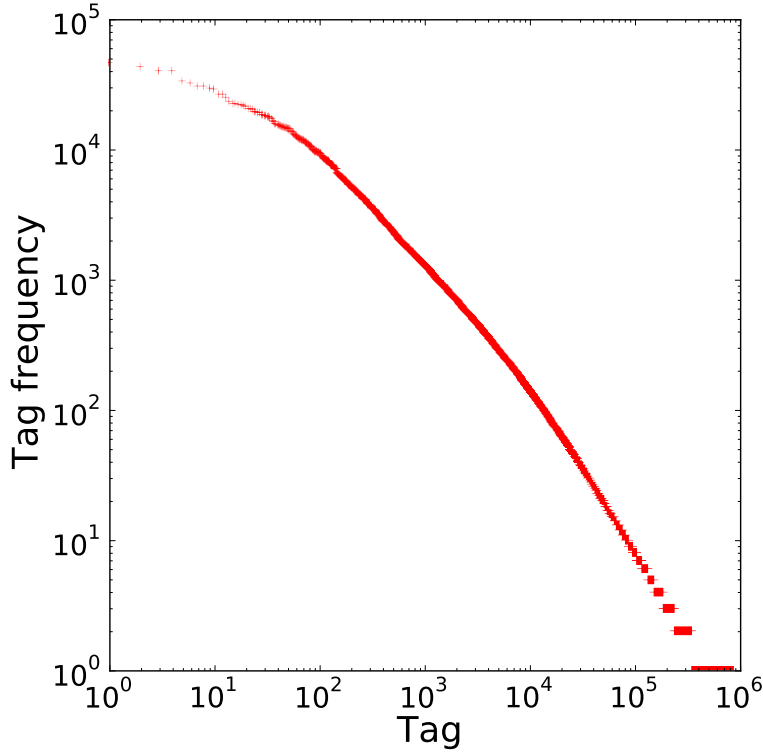


Fig. 4.5 The frequency distribution of tags in the MIRFLICKR-1M dataset.

words that are defined in the NLTK Stopwords Corpus [112] and applying Porter stemming [113], we filtered certain tags such as specific company names, tags including numbers or non-English words, tags used by less than 50 distinct users, and tags with lengths shorter than three. The top 850 ranked tags were selected as the concepts for assessing the effectiveness of our measures on semantic relationships. To obtain a representative distribution of each concept, we employed a set of images that were tagged with the concept. In the experiments, we used RGB-SIFT [114] to describe the visual features of the tagged images. Specifically, the SIFT features [115] that were independently extracted from each RGB channel were combined into 384-dimensional features. The bag-of-words (BoW) model was used for the vector quantization of the obtained features in an image, which produced a 2,986-dimensional vector per image. To characterize the text features of the tagged images, the tags in the dataset were pre-processed to remove stop words and were stemmed with the Porter algorithm. We used 3,575-dimensional binary features that represent whether the corresponding tag is provided to the image as a tag. We experimentally set the parameter q of CCA to 500, which corresponds to an approximate reduction of about 83.3% of the original visual feature dimensions.

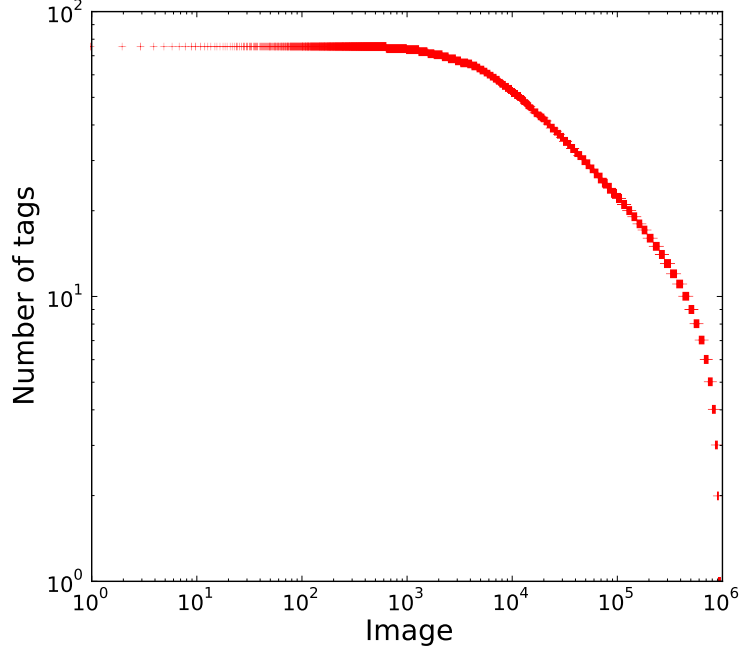


Fig. 4.6 The number of tags per image in the MIRFLICKR-1M dataset.

4.4.2 Visualizing Concept Relations

In this subsection, we present a novel visualization of inter-concept relations. The two types of semantic measures calculated by our method are represented as a graph $G = (V, E)$, where node $v_j \in V$ ($j = 1, 2, \dots, K$) corresponds to concept C_j , and the edge $e(v_j, v_l) \in E$ ($j, l = 1, 2, \dots, K$) indicate the semantic relatedness between two concepts C_j and C_l . In graph G , we provide the inverse of our measure $dist(C_j, C_l)$ as the weight of edge $e(v_j, v_l)$, and the abstraction level $\Phi(C_j)$ as the weight of node v_j . We depict the network by using NetDraw [92]. Since it is difficult to show the whole network, we show only the top 1% of strongest relatedness as edges for easier viewing. The depicted network is shown in Fig. 4.7, where node colors are set according to node weights by using the color bar shown in the right side. Red nodes indicate the corresponding concepts are at high levels of abstraction (i.e., abstract concepts), whereas blue nodes indicate the corresponding concepts are at lower levels of abstraction (i.e., specific concepts). In the following subsections, we evaluate the performance of the measures of the semantic relatedness and the abstraction level in 4.4.3 and 4.4.4, respectively.

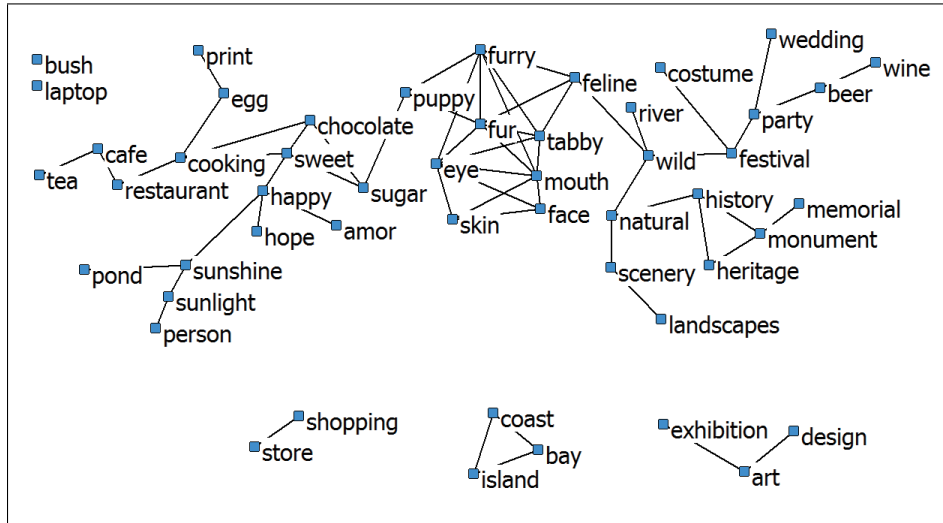
In the experiments, we first visualized the semantic relatedness that was measured by different methods as concept networks, and we then quantitatively evaluated their performances in several settings.

Comparison among different concept networks

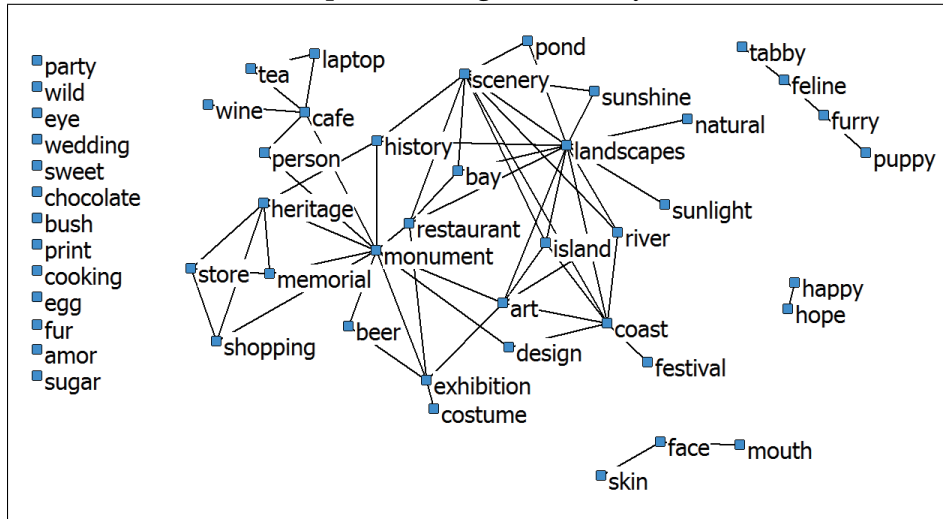
In this experiment, we show the subsets of different concept networks that were constructed by the proposed method (P-H) and conventional methods [3, 4]. The procedures of this experiment are shown below: we first randomly selected 50 concepts from a concept list. Using each of the different measures, we chose 60 concept pairs that have the strongest relatedness in the concept list. A concept network was generated using NetDraw [92] based on the 60 concept pairs, where each node corresponds to a concept and each edge indicates the strong semantic relatedness extracted by each method. The generated networks are shown in Fig. 4.8. As shown in Fig. 4.8, although the three networks were constructed from the same 50 concepts, their edges are totally different. In the following experiments, we quantitatively evaluated the properties of these measures based on several experiments: user evaluation, concept clustering, image annotation, and tag recommendation.

Comparison with human judgment of the most relevant pairs

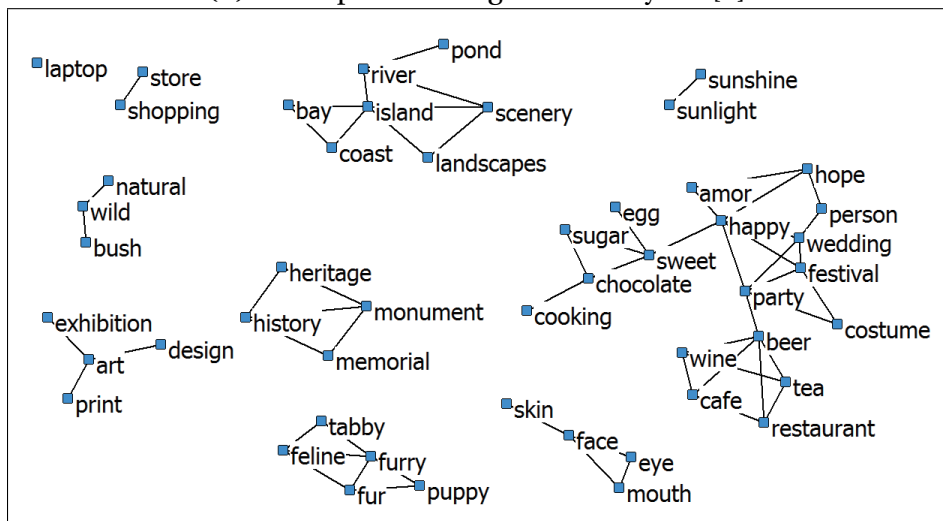
A distinct metric for evaluating a measure of semantic relatedness is its correspondence to human judgments of relatedness. Several studies [27, 31, 116] have shown that the use of human ratings can be a reliable reference when comparing computational methods. In this experiment, we investigated whether each measure that extracts relationships from tagged images conforms to human judgment for the relevant concept pairs. We extracted the 50 most relevant concept pairs using each of the different eight measures and removed duplicate pairs, which resulted in 252 relevant concept pairs. We presented 35 users with a list of the 252 concept pairs as detailed in conventional studies [4, 116]. For each concept pair, the user was required to provide a score that ranged from 1 (“semantically unrelated”) to 5 (“highly correlated”) according to their knowledge. The mean of the scores for each pair given by all users was considered to be the “user score” for that pair. The subjective evaluation results for the concept pairs are shown in Table 4.3, in which all measures are normalized to the range [1,5] as correlation scores. In this table, only a few pairs are shown due to space limitations and the concept pairs are sorted in descending order according to the user scores. The correlation scores that are closest to



(a) Concept network generated by TCD [3].



(b) Concept network generated by FD [4].



(c) Concept network generated by P-H.

Fig. 4.8 Concept networks whose edges are calculated using the proposed method (P-H) and the conventional methods [3, 4]. Each node represents a concept, and the edges between the concepts represent the strongest 60 relations.

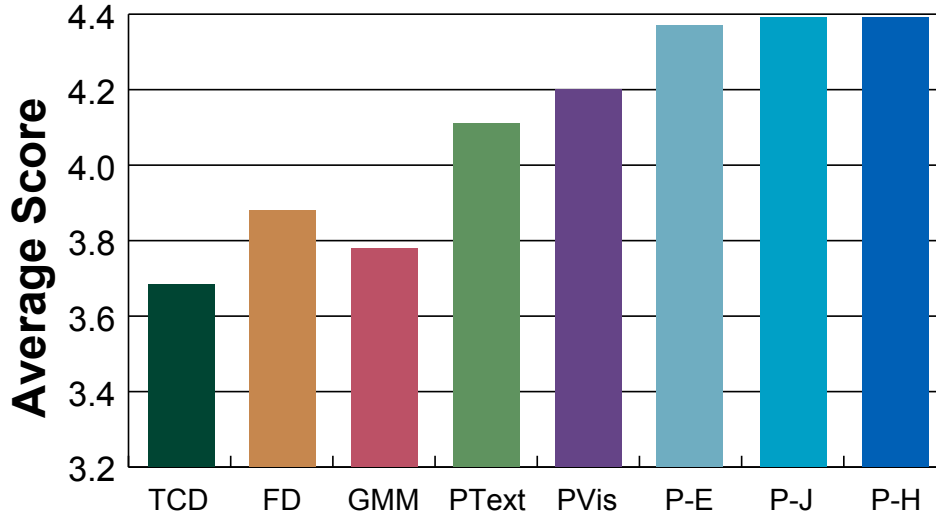


Fig. 4.9 Average user scores for the most relevant concept pairs extracted using different methods. A high score means that the method can accurately extract the top relevant concept pairs.

the user scores are highlighted in bold for reference. To evaluate whether each method accurately extracted relevant concept pairs, we averaged the user scores that were provided for the top 50 concept pairs extracted using each measure. Because the 50 concept pairs that were extracted using each measure differ, a measure that obtains a high average user score indicates that it accurately extracted relevant concept pairs compared with the remaining measures. The calculated average user scores for different measures are shown in Fig. 4.9. Our measure obtained better user scores compared with the other measures, which indicates the ability of our measure to detect highly correlated concept pairs. Conversely, the score for TCD was lower than the scores for other methods; this lower score may be attributed to the use of tag co-occurrence, which reflects the tendency of the dataset without capturing the semantic contexts of the concepts [117, 118]. For example, the most relevant pairs extracted by TCD included "shine-garage" and "fast-chrome"; they tend to co-occur in the same images (e.g., car show images). However, the two concepts in these pairs do not represent similar meanings, and therefore, they did

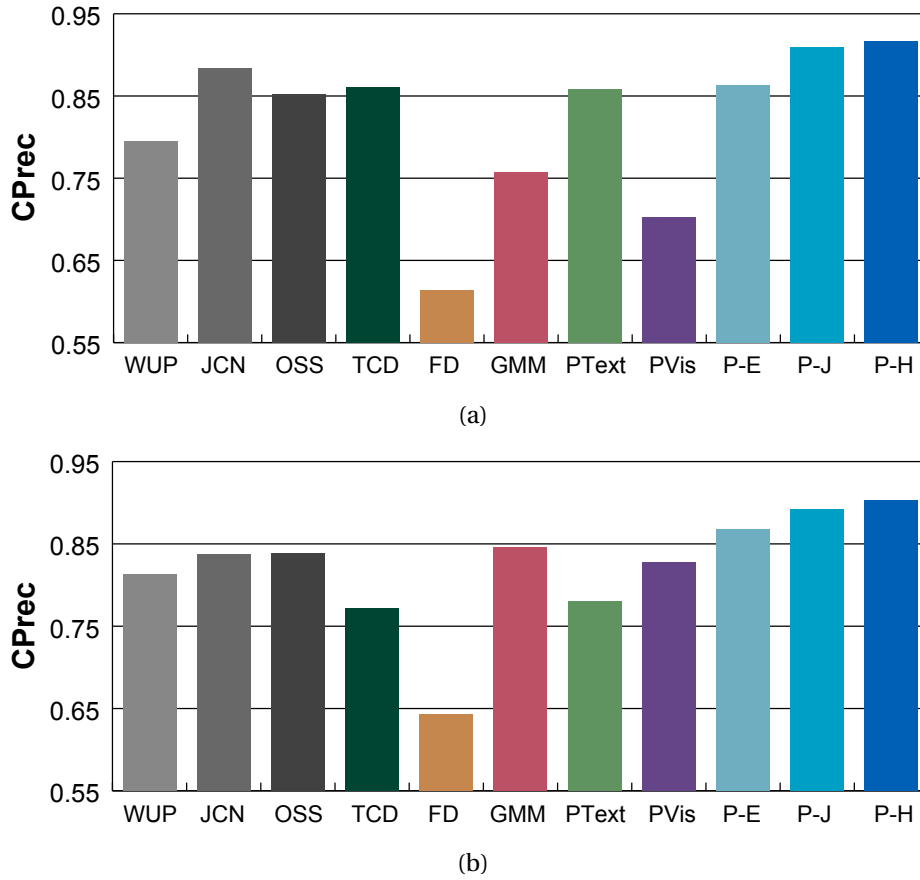


Fig. 4.10 Precision of clustering concepts using different measures of semantic relatedness: (a) results for the first clustering experiment (strongly correlated concepts), where text-based measures (TCD and PText) obtained high precisions, (b) results for the second clustering experiment (weakly correlated concepts), where visual-based measures (GMM and PVis) performed better than text-based measures.

not obtain high user scores. The results indicate that our method outperforms conventional methods with regard to obtaining the top relevant concept pairs.

An application to concept clustering

Here, we compare the measures of semantic relatedness based on the performance of an application in concept clustering. To construct a dataset for this experiment, we manually assigned category labels to all 850 concepts. In this experiment, we randomly sampled 180 concepts from 13 categories, in which several subsumption relations (hypernymy/hyponymy) existed. The categories and the numbers of concepts are as follows: vegetation (14), person (22), seascape (12), cityscape (24), animal (23), transportation

(16), food (17), jungle (14), music (8), bird (6), insect (6), electronic device (7), and furniture (11). Examples of the categories and their concepts are shown in Table 4.4. We performed spectral clustering based on each measure; the obtained set of N_θ concept clusters was denoted by $\Theta = \{\theta_1, \theta_2, \dots, \theta_{N_\theta}\}$. For performance evaluation, we calculated the following precision [119, 120] as

$$CPrec = \frac{1}{N_\theta} \sum_{\theta \in \Theta} \frac{\max_{\psi \in \Psi} |\theta \cap \psi|}{|\theta|}, \quad (4.26)$$

where $\Psi = \{\psi_1, \psi_2, \dots, \psi_{N_\psi}\}$ is the set of N_ψ ground truth concept clusters, $|\theta|$ represents the number of concepts that belong to the cluster $\theta \in \Theta$, and $|\theta \cap \psi|$ represents the number of concepts that belong to the clusters $\theta \in \Theta$ and $\psi \in \Psi$. The results are shown in Fig. 4.10 (a). Based on this figure, we determine that the precision of our cross-modal approach is highest among all methods. In this setting, the text-based methods (TCD and PText) provide better results than the visual-based methods (FD, GMM and PVis). Compared with the results in Fig. 4.9, TCD yielded better results, as shown in Fig. 4.10 (a), because the concept categories defined in this experiment can be moderately discriminated in terms of co-occurrences. To produce more extensive experiments, we also randomly sampled 138 concepts from 12 categories as the second clustering experiment. Specifically, we discarded the hypernymy member of the pairs in any subsumption relations. The categories and the numbers of concepts are as follows: person (17), food (17), animal (15), weather (10), urban city name (10), scenery (9), vivid color (12), architecture (5), water (6), vegetation (6), winter (6), and event (25). Examples of the categories and their concepts used in this experiment are shown in Table 4.5. The concepts within each category do not always co-occur in an image but instead are correlated compared to the concepts of other categories. For example, although concepts that represent urban city names (e.g., Tokyo and Shanghai) are not simultaneously provided to the same image, they are semantically correlated compared with the concepts that belong to the other category. The clustering results based on this second set are shown in Fig. 4.10 (b). In this setting, GMM and PVis exhibit better results than TCD and PText. Thus, we consider that the visual features are effective for grouping weakly correlated concepts. Our cross-modal approach outperforms other methods in both settings. The results indicate that our measure is also effective for the semantic clustering of concepts. In our approach, P-J and P-H performed better than P-E. We can consider that the use of covariance information of distributions produces a better estimation of semantic relatedness. We discover that WordNet-based measures (WUP, JCN and OSS) also show moderately high results

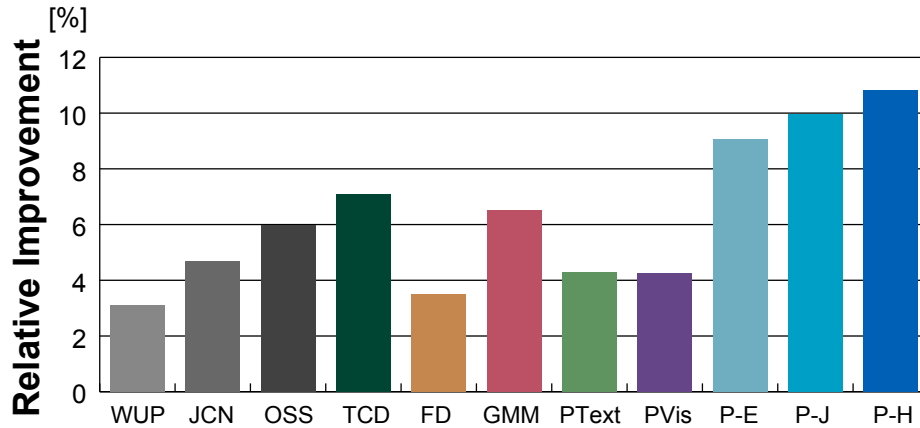


Fig. 4.11 Relative improvement of image annotations with different semantic measures. The baseline MAP was computed based on initial annotations. A high MAP gain means that the method refined annotations better.

in both settings; however, they do not capture some relationships within the same category. For example, “squirrel” was considered to be more correlated with “person” than with other animals, such as “cat” and “horse”; as a result, the clustering sometimes failed.

FD performs better than GMM, as shown in Fig. 4.9, but FD yields worse results than GMM, as shown in Fig. 4.10. This difference is attributed to the feature space of FD: the model of the FD introduces spatial information to visual features and the resulting feature distribution of each concept is highly sparse [4]. The similarity between two sparse distributions cannot be effectively characterized using KL divergence in the high-dimensional feature space of the FD [52], which affects the relative order of the relatedness between the concepts in the dataset. FD can successfully capture the relatedness between two concepts if their image layouts are very similar: as shown in Table 4.3, the top relevant concept pairs extracted by FD included “sunset-sunrise” and “seascape-lake”. The concepts in these pairs were successfully grouped to the same clusters; however, the relatedness between other concepts cannot be adequately characterized due to the previously mentioned sparseness. Thus, the total clustering failed as shown in Fig. 4.10.

An application to image annotation

Recent image annotation techniques introduce measures of semantic relatedness to improve the annotation results. In [67, 68, 97], Random Walk with Restarts (RWR) [121] is used to refine the initial annotations of an image. This subsection evaluates the perfor-

mance of each measure in an application of image annotation using RWR. We denote V as a vocabulary of concepts, and $|V|$ is the vocabulary size. Let $\mathbf{A}$ be a row-normalized adjacency matrix whose (i, j) -th entry denotes semantic relatedness between the i -th concept and the j -th concept of V . Given a testing image, a $|V|$ -dimensional initial annotation score vector $\mathbf{v}$ is first provided using its visual features by assuming that each concept is independent from the others, where the j -th entry of $\mathbf{v}$ denotes a detection score of the j -th concept of V . Then, the annotation confidence scores of the concepts are updated at the t -th iteration using Eq. (2.19). In this paper, we tuned the parameter α to achieve the best performance for each measure. By sorting the scores, the top concepts are provided to the testing image as annotation words. To conduct this experiment, we used the NUS-WIDE dataset [122], which contains 269,648 Flickr images. These images are divided into a training set (161,789 images) and a testing set (107,859 images). All of the images are manually labeled with 81 semantic concepts, and the labels were regarded as ground truth. We used detection scores that were provided by the authors of [123] as an initial annotation score vector, which were calculated using SVMs trained on BoW-based visual features. To evaluate the performance of the image annotation with each measure, we calculated the relative improvement in the Mean Average Precision (MAP) compared with the initial annotations. The obtained results are shown in Fig. 4.11. This figure shows that our measure using the Hellinger distance obtained the best relative improvement among all of the measures. Some examples of testing images and their top seven annotations are shown in Table 4.6. As shown in this table, P-H refined the annotation results to be semantically consistent. We note that annotation refinement using semantic relatedness between concepts can depend on the quality of the used vocabulary. If the vocabulary has a bias such that the number of concepts that belong to each category differs widely from other categories, some of the concepts that are in the minority might not benefit from the semantic measure. Thus, we must discuss and tackle this problem in future work.

An application to tag recommendation

Tag recommendation is also an important application of semantic relatedness when providing related tags to users who wish to tag an image. We also evaluate different measures by comparing their performance in a tag recommendation task. To conduct this experiment, we exploited a *sum* strategy, which is presented in [80]. A brief outline of the overall procedures is as follows: given a testing image with initial tags, each measure first created a set of indexes of tags that are related to the j -th initial tag based on semantic related-

ness. The set is denoted as Ω_j . Then, a recommendation score of the k -th tag of the vocabulary, $score(k)$, is calculated using Eq. (2.20). Tags that have high scores are finally recommended to the image. We randomly sampled 6,000 images from each category of the NUS-WIDE dataset. For each of the images, the top five recommendations from each of the semantic measures were used to construct a tag pool. Each tag in the pool was evaluated by 12 users to construct the ground truth in the same way as in [80]. To evaluate the results, we used standard information retrieval metrics: the success at rank k (Success@ k) and the precision at rank k (Prec@ k). Success@ k indicates the mean probability that a relevant tag occurs within the top k of the recommendations, and Prec@ k denotes the average proportion of the relevant tags within the top k . The results are shown in Table 4.7. We can find from this figure that our measure with the Hellinger distance outperformed the other measures. Compared with the WordNet-based measures (WUP, JCN and OSS), the advantage of our method includes not only its performance but also the coverage of the tags, as explained in Section 2.1; methods that extract semantic relationships from data on the Web can be applied to various tags, which increase in number daily.

Parameter effect analysis

The variable q , which was experimentally established in the experiments, is a parameter that determines a trade-off between the computational efficiency and performance of the proposed method. To investigate these two properties of the proposed method when changing q , we performed additional experiments, in which q was set to 1,000 and 50. First, to demonstrate the influence on efficiency, we measured the computation time for the following two computations: (a) the calculation of representative distributions of 850 concepts on one million samples and (b) the measurement of the pairwise relatedness among 850 concepts using a multi-core 3.50GHz CPU with 16G RAM. The measured time for each procedure is shown in Fig. 4.12, in which the computation time of FD and GMM are also reported. As shown in Fig. 4.12, FD requires the most computation time because the model is constructed in a high-dimensional feature space. We also observed from this figure that reducing q resulted in a faster computation of the proposed method. Note that our method shows better performance than GMM even in the case of $q = 50$. Second, to demonstrate the influence on performance, we applied two sets of the pairwise relatedness between concepts, which were obtained using $q = 1,000$ and $q = 50$, to concept clustering and image annotation, respectively. Figure 4.13 shows performances in different experiments when changing q , where the results for two concept clustering

settings are shown in Figs. 4.13 (a) and 4.13 (b) and the results for image annotation are shown in Fig. 4.13 (c). As shown in Figs. 4.13 (a) and (b), an increase in q from 500 to 1,000 produced performance improvement of the concept clustering, whereas reducing q to 50 resulted in performance degradation. Similarly, in the image annotation experiment, a reduction in q resulted in a degradation in relative improvement in MAP, as shown in Fig. 4.13 (c). Although we assume that our method is performed off-line, when considering its extension to certain online applications, the efficiency of the proposed method should be improved. Our future work includes an investigation of how to automatically determine the variable q that will ensure that reasonable performance can be maintained with high efficiency. The total computation time can also be improved by introducing rapid implementations of visual feature extraction, which is the same strategy employed in visual categorization tasks [124, 125]. We will introduce schemes to reduce the computational cost of the proposed method in future work.

Note that the effectiveness of each method that uses tagged images as an information source is dependent on the dataset. If the dataset possesses some content bias, in which many images belong to specific categories (e.g., food, animals, and illustrations), the accurate extraction of the general relationship between the concepts will become difficult. This problem is a common issue when using the proposed method and the conventional methods. We plan to tackle this bias problem when collecting tagged images in our future work.

4.4.4 Evaluation of measures on abstraction levels

In this subsection, we verify the effectiveness of the measures for the abstraction levels of the concepts. We compare the performance of our measure with the following conventional methods, which exploit either tags or images on Flickr: a tag statistics-based method [43] (denoted as **TS**), a visual feature-based method [51] (denoted as **VF**), and the single-modality versions of the proposed method (denoted as **PText** and **PVis**). By comparing the methods that exploit a single modality available on Flickr, we investigated the effectiveness of our cross-modal approach for estimating the abstraction level. As a baseline, we also show the performance of a method that uses concept frequency in the dataset (denoted as **Freq**), which assumes that the more frequently a concept occurs in the dataset, the higher is its abstraction level.

In the experiments, we focused on the properties of concepts in the top and bottom levels located by the proposed method and quantitatively evaluated the performance of

different methods by several metrics.

Comparison among top and bottom levels of concepts

First, we investigated what types of concepts are located by our method in the higher and lower levels. Table 4.8 lists the top and bottom ranked concepts based on the abstraction levels estimated by PText, PVis and the proposed method. In this table, concepts that represent objects, scenes, or activities are highlighted in bold for reference. From the results, we find that our measure tends to place specific concepts that represent objects, local city names, or activities in the lower level and abstract concepts that represent impressions, broad locations or adjectives in the higher level. In the following experiments, we examined the performance of measures of the abstraction level.

Comparison with WordNet hierarchy

We utilized the depth information of the concepts on the WordNet hierarchy as the ground truth. If an abstract concept is located on a higher level than a specific concept in each concept pair, then their abstraction levels are accurately estimated. However, it is difficult to determine the ground truth for all of the concept pairs. For example, among the concepts that belong to different categories (e.g., squirrel, sunset, and daisy), their relative order based on the abstraction level cannot be determined. Therefore, it is more reasonable to focus on a pair that is composed of a concept and its superordinate concept and determine which is the more abstract of the two concepts. For this evaluation metric, we extracted several pairs of concepts from the WordNet hierarchy that have subsumption relations as ground truth pairs. The total number of extracted hierarchical pairs was 1,082. Some of these pairs, in which the right concept is a hypernym of the left concept, i.e., the right concept should be located at a higher level than the left concept, are shown in Table 4.9. The abstraction levels between two concepts in each of the pairs were compared and the precision of all pairs was calculated as follows:

$$Precision = \frac{\# \text{accurately located pairs}}{\# \text{WordNet hierarchical pairs}}. \quad (4.27)$$

The results are shown in Fig. 4.14. This comparison illustrates that our measurement of abstraction levels outperforms the conventional measures. In this experiment, the text feature-based methods (TS and PText) obtained higher precisions than the visual feature-based methods (VF and PVis): because the concept organization in WordNet follows a psycholinguistic hierarchical structure, either or both of the concepts in each pair

are usually located at high abstraction levels. In the case in which both of the concepts are abstract, modeling the concepts with visual features is difficult [17]. Thus, visual feature-based methods failed to accurately determine the concept with a higher abstraction level.

Evaluation using concept category labels

Evaluation of the techniques for estimating the abstraction levels of the concepts is a challenging research issue because there is no extensively adopted metric for performance evaluation. According to a conventional study [51], we evaluated a partial order such that the concepts that represent objects, scenes, or activities are located in lower levels compared with the concepts that represent locations, impressions, and other entities. In this experiment, all concepts were manually labeled with one category in the same way as that in the conventional work [51]. The category labels of the concepts and the examples in each category are shown in Table 4.10. Based on these labels, we calculated the precision of the bottom m concepts as follows:

$$P@m = \frac{\text{\#concepts representing objects, scenes, or activities}}{\text{the bottom } m \text{ concepts in the sort of scores}}. \quad (4.28)$$

If $P@m$ approaches 1, then the measure can accurately locate the specific concepts in the lower levels. The results are shown in Fig. 4.15. This figure shows that the performance of PVis is better than the performance of PText, which indicates that the visual features are effective cues for detecting concepts that represent specific meanings. Because PText evaluates the compactness of the distribution in a text feature space, it can mistakenly give low scores to abstract concepts that often co-occur with particular tags in the same images. This caused the lower performance of PText compared with PVis, as shown in Fig. 4.15. Compared with the conventional methods, the proposed method can accurately place the concepts that represent objects or scenes in the lower level compared with other methods. The results indicate that our cross-modal approach is effective for detecting concepts that represent specific objects or scenes in lower levels.

The simple frequency-based method (Freq) cannot accurately estimate the abstraction levels of concepts in the two experimental settings, as shown in Fig. 4.14 and Fig. 4.15. This finding is attributed to the incompleteness of tags in social media [126, 127], i.e., users do not always label images with all true concepts.

Discussion

The experiments have shown that the proposed method can estimate the abstraction level of each concept better than conventional methods. Based on the results, we determine that the collaborative use of visual and text features is an effective solution for accurately estimating the abstraction levels of concepts. Because our method can evaluate the compactness of the distribution in terms of the semantic topics, the concepts that represent specific objects, scenes, or activities can be located in a lower level compared with the concepts that represent locations, impressions, and other topics.

As described in the discussion part of Section 4.4.3, the variable q influences the performance of the proposed method. Results of our additive experiments for Section 4.4.4 are shown in Fig. 4.16, in which q was newly set to 1,000 and 50. This figure shows that an increase in q from 500 to 1,000 resulted in a precision similar to the case of $q = 500$, whereas a reduction in q from 500 to 50 resulted in a degradation of precision. Thus, the selection of the parameters should consider the trade-off relation between efficiency and effectiveness, which is the focus of our future work.

The abstraction levels of concepts have been used in multimedia-related tasks, such as hierarchical image annotation [64, 128], in which these methods generally use a vocabulary that consists of concepts that are well defined in WordNet. Conversely, the vocabulary used in social photo-sharing websites such as Flickr include tremendous concepts that are not defined in WordNet [129]. Our method can automatically measure the abstraction levels of concepts used on the Web, which contributes to an increase in the vocabulary size. In addition, the abstraction levels of concepts, which are measured by our method, will facilitate the browsing of large-scale social annotations [130].

Based on the experiments described in Section 4.4.3 and Section 4.4.4, the cross-modal approach that uses both text and visual features from tagged images is shown to be more effective for extracting semantic relationships between the concepts than conventional methods that use either of the two features.

4.5 Conclusion

In this chapter, we present a cross-modal approach for extracting semantic relationships between concepts using tagged images on Flickr. In the proposed method, we first project the text and visual features of the tagged images to a latent space using canonical correlation analysis (CCA). Then, under the probabilistic interpretation of CCA, we

calculate a representative distribution of latent variables for each concept. Based on the representative distributions of the concepts, we derive two types of measures: the semantic relatedness between the concepts and the abstraction level of each concept. The experimental results show that our measures are more coherent with human cognition than conventional methods that use either text or visual features. Furthermore, the applicability of our measures to multimedia-related tasks is also demonstrated by the experimental results.

Although a representative distribution for a concept is modeled as a normal distribution, we can improve the performance of the extraction of the semantic relationships between the concepts. This opportunity derives from the use of both text and visual features. However, a more complicated distribution such as GMM may be suitable for characterizing the semantic context of the concept. In future work, we will investigate approaches to modeling the representative distributions. We will also investigate how to automatically determine the dimensionality of the latent space estimated by CCA in terms of computational efficiency and performance of the proposed method. Furthermore, we will address the bias problem of collected tagged images.

Table 4.3 Examples of subjective evaluation for the most relevant pairs extracted using different methods. Each of the correlation scores are normalized from 1 (“semantically unrelated”) to 5 (“highly correlated”). The correlation scores that are closest to the user scores are highlighted in bold.

Concept pairs and their correlations scores					
Concept pair	TCD	FD	GMM	P-H	User
lady - woman	4.01	4.03	4.13	4.38	4.79
tokyo - japan	4.60	4.28	3.91	4.41	4.72
fast - speed	4.66	4.01	3.83	4.06	4.70
trip - travel	4.30	3.92	3.83	4.87	4.69
snow - winter	4.54	3.81	4.13	4.34	4.69
plants - garden	3.97	3.63	4.01	4.70	4.67
woods - forest	4.33	4.03	4.22	4.39	4.60
germany - europe	4.15	3.71	3.92	4.19	4.59
pretty - beauty	4.48	3.52	3.90	4.36	4.58
sunrise - sunset	3.61	4.57	3.95	4.41	4.50
bloom - flower	3.95	3.42	3.80	4.36	4.50
man - people	3.88	4.24	3.91	4.55	4.40
cake - chocolate	4.48	4.34	4.25	4.19	4.40
puppy - kitten	3.51	3.79	4.35	4.09	4.28
boy - girl	3.83	3.99	4.31	4.22	4.11
people - street	3.98	3.80	3.93	3.66	3.69
seascape - lake	2.78	4.76	3.71	2.58	3.55
fantastic - award	4.83	3.98	3.69	4.08	3.51
rocks - waves	4.28	3.95	3.66	3.01	3.50
monkey - pet	2.56	3.68	4.33	3.36	3.35
squirrel - puppy	2.59	3.58	4.32	3.18	3.30
fast - chrome	4.64	4.18	3.92	3.49	3.21
model - asia	4.28	4.08	3.73	3.03	2.82
islands - air	4.18	3.87	3.46	2.20	2.79
wonder - top	4.63	3.81	3.78	3.40	2.78
crowd - costume	3.36	3.22	4.32	2.58	2.58
sleeping - costume	2.70	3.35	4.32	2.21	2.25
international - engine	4.49	4.18	3.53	3.14	2.19
shine - garage	4.57	3.29	2.41	1.80	2.10

Table 4.4 Examples of the categories and concepts used for the first clustering experiment. Concepts that belong to the same category correlate strongly with one another.

Category	Concepts
music	music, concert, live, guitar, band, gig, pop, punk
person	portrait, people, girl, woman, man, child, family, friends, boy, male, female, lady, person, mother, guy, president, human, daughter, lip, face, hair, kid
seascape	beach, sea, ocean, coast, bay, shore, port, pacific, horizon, seascape, harbor, seaside

Table 4.5 Examples of the categories and concepts used for the second clustering experiment. Concepts that belong to the same category do not explain one another compared with the strongly correlated concepts used in the first clustering experiment.

Category	Concepts
animal	cat, dog, rabbit, bear, lion, sheep, squirrel, deer, bunny, cow, elephant, zebra, fox, monkey, horse
urban city name	london, paris, hongkong, manhattan, chicago, tokyo, berlin, shanghai, toronto, moscow
weather	sunset, sunrise, sunny, glow, dusk, wind, rainy, twilight, shine, cloudy

Table 4.6 Examples of testing images and automatic annotation using four different measures, where the top seven annotations are provided to each image. The correct annotations are highlighted in bold.

Image	JCN	TCD	GMM	P-H
	animal grass person plants sky elk horses	animal grass sky clouds person water elk	sky animal grass person water clouds sports	animal clouds sky grass elk dog water
	vehicle cars window road plane boats buildings	window road buildings sky clouds vehicle cars	window sky road buildings clouds vehicle grass	window sky clouds buildings road vehicle railroads
	water ocean lake buildings house grass sky	sky clouds water buildings grass ocean lake	ocean sky lake water buildings sunset grass	ocean clouds sky lake water beach sunset

Table 4.7 Evaluation results for tag recommendations with different semantic measures. Each column shows the corresponding measure's performance. The best result in each row is marked in bold.

Evaluation metrics	Prec@3	Prec@5	Success@1	Success@3	Success@5
WUP	61.91%	58.45%	61.47%	78.10%	86.01%
JCN	60.82%	59.83%	61.89%	79.35%	86.73%
OSS	61.98%	61.53%	63.02%	80.26%	88.05%
TCD	66.55%	63.63%	71.46%	85.98%	91.06%
FD	56.83%	56.59%	55.06%	74.29%	83.26%
GMM	55.83%	54.89%	57.64%	74.42%	82.39%
PText	61.17%	59.38%	64.03%	81.55%	88.15%
PVis	61.03%	59.10%	63.21%	80.41%	85.15%
P-E	72.11%	67.51%	79.74%	92.23%	94.66%
P-J	72.62%	67.49%	85.42%	92.28%	94.73%
P-H	73.55%	68.16%	87.63%	93.15%	95.26%

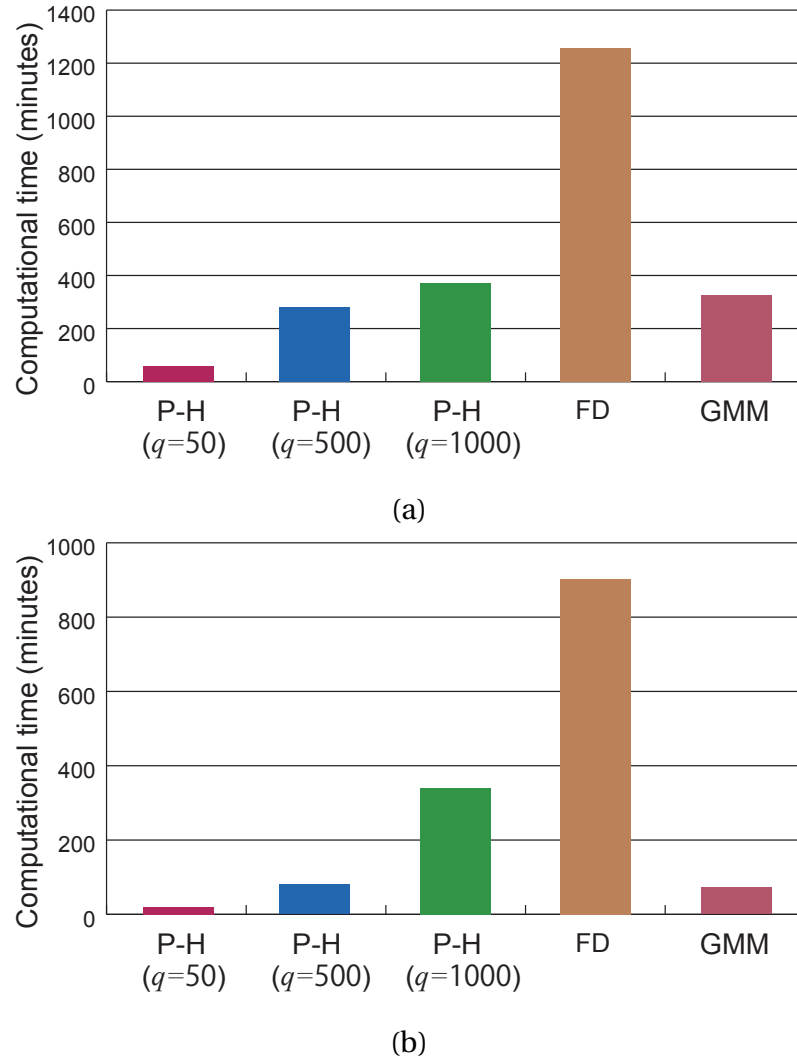


Fig. 4.12 The computation time for each procedure when changing q in the proposed method and for each procedure in the conventional methods [4, 5]: (a) the computation time for the calculation of the representative distributions of 850 concepts, (b) the computation time for the measurement of the pairwise relatedness among 850 concepts.

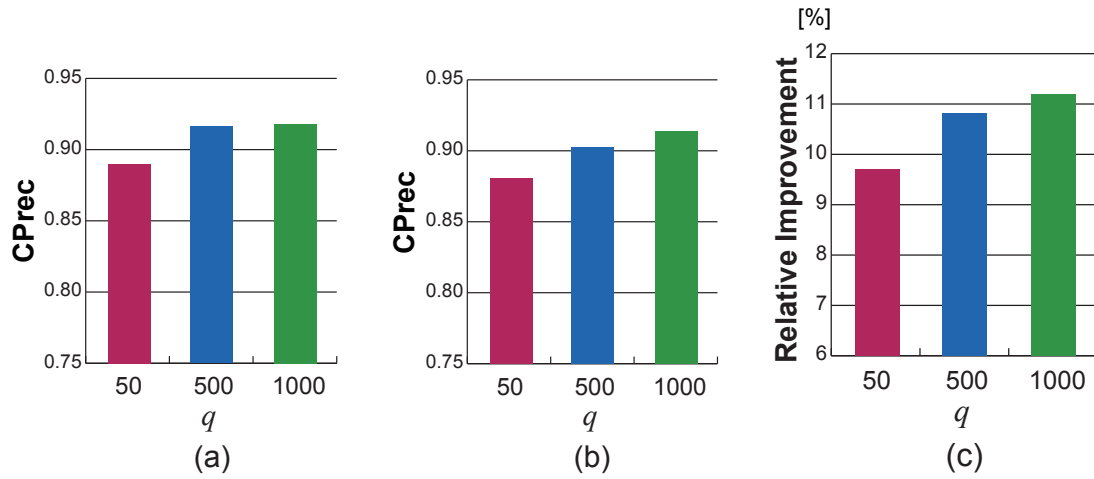


Fig. 4.13 Performance comparison in different experiments when changing q in the proposed method (P-H): (a) results for the first clustering experiment (strongly correlated concepts), (b) results for the second clustering experiment (weakly correlated concepts), and (c) results for image annotation.

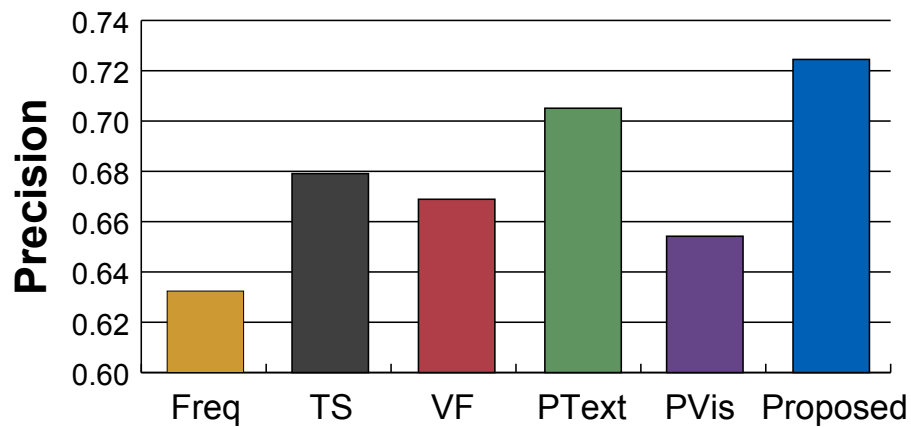


Fig. 4.14 Precision of comparing the abstraction levels between two concepts, where the ground truth is determined using the WordNet hierarchy.

Table 4.8 Results of abstraction level estimation for each of the concepts, in which the top 15 (abstract) and bottom 15 (specific) concepts are shown. Concepts that represent objects, scenes, or activities are highlighted in bold for reference.

Concepts at the top 15 and the bottom 15 abstraction levels			
Rank	PText	PVis	Proposed
t1	light	lines	sensational
t2	blue	abstract	engine
t3	color	silhouette	international
t4	white	minimalism	excellence
t5	city	blue	rac ing
t6	black	azul	classic
t7	beautiful	orange	fake
t8	travel	shadow	fabulous
t9	red	light	garage
t10	water	geometric	unique
t11	sky	sky	amazing
t12	green	color	diesel
t13	park	sun	europe
t14	dark	yellow	america
t15	summer	red	natural
b15	lego	alley	iphone
b14	livemusic	carshow	stencil
b13	soldiers	furry	cat
b12	avatar	creek	carshow
b11	anime	squirrel	deer
b10	tabby	wilderness	puppy
b9	cupcake	lion	livemusic
b8	macros	stream	kyoto
b7	alone	deer	kitten
b6	steampunk	soldiers	knitting
b5	knitting	babies	recipe
b4	carshow	era	crochet
b3	crochet	mammal	tabby
b2	origami	grave	squirrel
b1	recipe	native	origami

Table 4.9 Some of the 1,082 concept pairs used for the objective evaluation of abstraction levels. In each of the pairs, the right concept is a hypernym of the left concept.

Part of hypernymy relations in WordNet		
tabby → cat	candy → sweet	lake → water
daisy → flower	yellow → color	cupcake → cake
car → vehicle	soccer → football	bathroom → room
cruise → voyage	aircraft → vehicle	strawberry → fruit
yard → field	baby → child	street → road
wildlife → life	statue → sculpture	dandelion → plant
christmas → holiday	bug → insect	sketch → drawing
snow → weather	sidewalk → path	running → action
cycling → sport	diesel → engine	mother → woman
vinyl → plastic	airplane → aircraft	desk → table
president → person	illustration → artwork	squirrel → fur
tour → journey	sky → atmosphere	love → emotion
concert → performance	kitchen → room	cheese → food
dinner → eating	lantern → lamp	scarf → dress
origami → art	cherry → tree	drawing → art
sunlight → light	twilight → light	rose → plant
wildlife → life	daughter → female	joy → emotion
party → event	aquarium → structure	snow → weather
sidewalk → path	carnival → festival	...

Table 4.10 Category labels of concepts and examples in each category.

Label	#concepts	Examples of concepts
object	245	tabby, lantern, cellphone, gun, chocolate
scene	140	midtown, sky, sunset, park, seascape
activity	34	shopping, wedding, dance, work, hiking
location	70	america, europe, asia, japan, earth
opinion	73	excellence, favorite, beautiful, bravo, lovely
time	9	spring, summer, october, winter, era
other	279	engine, life, humor, gothic, long

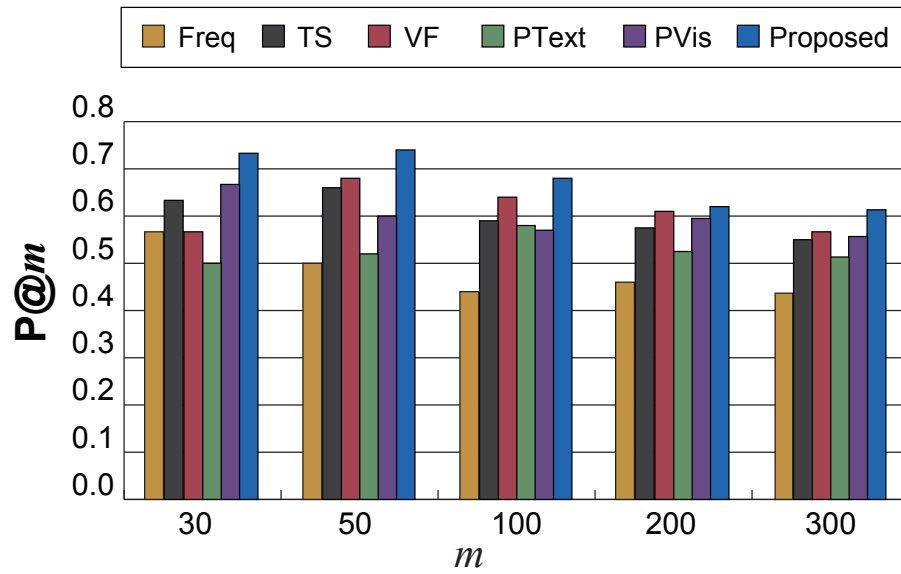


Fig. 4.15 Performance on measuring the abstraction level. A high $P@m$ means that a method accurately locates specific concepts in the bottom m concepts in the sort of the scores.

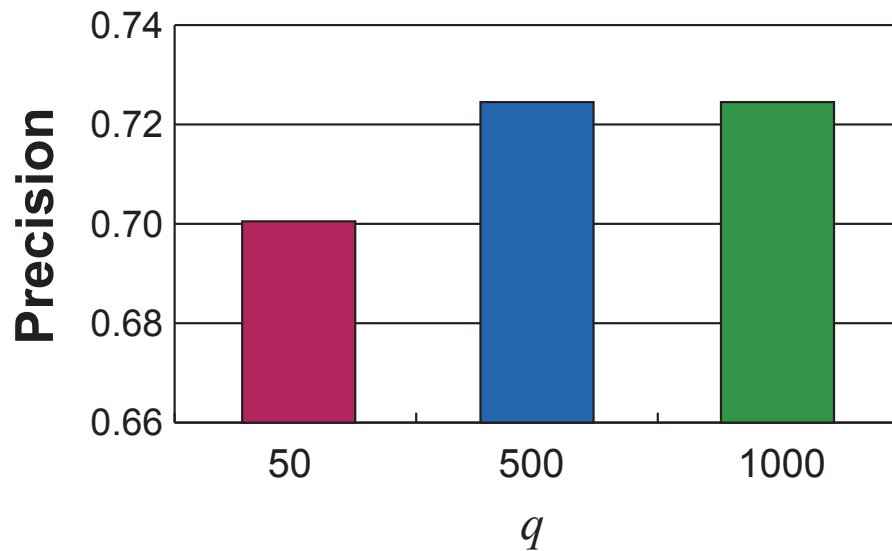


Fig. 4.16 Performance comparison in WordNet-based evaluation of the abstraction levels between two concepts when changing q in the proposed method.

Chapter 5

Exploiting Temporal Information for Finding Tag Relationship Changes

5.1 Introduction

In Chapter 3 and Chapter 4, we presented methods for extracting semantic relationships between concepts from a collection of tagged images. In particular, the method presented in Chapter 4 can be applied to tag relationship extraction in a target social media website [131]. Tag relationships extracted using the proposed method facilitate the user's understanding of tag usage in the social media website.

The previous method was performed under the general assumption that each tag always represents the same meaning over time, i.e., it did not consider *time sensitiveness* of tag relationships. On the other hand, recent data mining methods using text documents as an information source showed that considering content changes over time is effective for finding new semantic relationships between words [132, 133]. Thus, if we exploit the date when an image was taken as a new feature, we expect that novel semantic relationships between tags can be extracted from a target image collection. Owing to recent development of image capturing devices, each image on a social media website is associated with its *timestamp*, indicating when the image was taken or uploaded. Sorting images having a target tag by their timestamps, we can analyze the changes of a corresponding tag's semantics over time.

In this chapter, we exploit the temporal information of tagged images for finding tag relationship changes over time. We extend the visual feature-based tag relationship approach to capture the time characteristics of the tag. Specifically, the proposed method updates a probabilistic model that corresponds to each tag based on timestamps associ-

ated with tagged images. We first construct a Gaussian mixture model (GMM) for each tag using visual features of initial images, which is a similar approach as in [5, 134]. Then, given a new image over time, we recursively update parameters of the GMM using its visual features. The recursive update of a GMM has been used in several applications such as background subtraction [135] and anomaly detection [136, 137], and it can reflect the change of the dataset over time. Furthermore, based on the updated parameters, we present two types of measures: (i) the degree of visual change of the tag over time, and (ii) the visual similarity between tags over time. In experiments conducted using Flickr tagged images, we compare the performance of the proposed method and the conventional method that does not exploit timestamps of images.

The remainder of this chapter is organized as follows: Section 5.2 provides a brief review of the related studies that focus on temporal dynamics of Web data. Section 5.3 describes the construction of a GMM for each tag in a visual feature space. Section 5.4 presents two types of measures that represent tag characteristics over time, which are calculated based on the GMM updates. Section 5.5 presents the results of experiments for investigating the effectiveness of the proposed method. Finally, this chapter is summarized, and some possible directions for future work are suggested in Section 5.6.

5.2 Related Work

Recently, several methods that analyze temporal evolution of the data on the Web have been presented. Most of them were designed for analysis of text documents, while a few studies focus on the Web image dynamics. In this section, we provide a brief review of related work.

5.2.1 Temporal dynamics of text documents

In a research field of text retrieval, several methods try to quantify the temporal dynamics of text documents for a target concept. The temporal information of each document is exploited as a complementary source to capture the semantic relationships between words. Chien and Immorlica [138] calculated the temporal correlation between two keyword queries. Their measure is based on the correlation coefficient of queries' frequency functions, and experiments conducted using query streams from a search engine showed that it can discover semantically related keywords. Shokouhi [139] detected seasonal queries through time-series analysis of past temporal query patterns. Radinsky

et al. [132] compute the word relatedness using patterns of word usage over time in text document collections. Their approach was motivated by the observation that semantically related words do not necessarily co-occur in the same articles but they are likely to be mentioned roughly around the same time. Based on this idea, they model temporal dynamics of a concept as a series of its usage frequency over the database, and use them for constructing an ESA vector of a word. They calculated the word relatedness as the sum of pairwise concept relatedness in the corresponding two vectors using metrics for two time series, such as Dynamic Time Warping [140]. With experiments conducted on a dataset that was constructed using Amazon's Mechanical Turk service, the method showed superior performance compared to the conventional methods that do not consider temporal aspects of concepts.

The abovementioned methods used temporal patterns for individual tags. Recently, Yao et al. [133] focused on the tag correlations evolving over time to detect salient events in a social media website. In their method, a tag graph is constructed using tag co-occurrences within each time interval, and a derived graph from the original graph is used to identify the bursty events. Similar to this approach, in this chapter, we focus on the changes of a tag network over time to extract temporally-representative tag relationships.

5.2.2 Temporal dynamics of Web image collections

Compared with literature in text mining, there exist very few studies that focus on temporal dynamics of Web image collections. Rattenbury et al. [141] assumed that significant usage patterns for social event tags in a photo sharing website should appear as a burst over a period of time, and exploited signal processing techniques to identify such bursts. Their approach is related to the abovementioned text mining approaches.

Kim et al. [142] studied temporal dynamics of visual features in an image collections collected from a social media website, which aims to perform subtopic outbreak detection. Specifically, the authors in [142] presented a non-parametric approach based on Sequential Monte Carlo to create an image similarity network. The obtained results were used to classify Flickr images irrelevant to a target tag. Borth et al. [143] presented an automatic vocabulary selection method for video concept detection. They expanded concepts associated with a target video clip with relevant trending words. Kim et al. [144] revealed that considering frequencies of images corresponding to a target tag over time can improve the performance of time-sensitive image retrieval. Their method is based

on regularized multi-task regression on multivariate point processes, in which several types of visual features of an image stream is used to form a process. Uricchio et al. [145] analyzed the temporal aspects of two datasets, NUS-WIDE [122] and MIRFlickr-1M [9], using a correlation measure over two time series of tag frequencies. In this chapter, we present a novel method that focuses on the temporal visual changes of a target tag and tag relationships.

5.3 Constructing Gaussian Mixture Models for Tags

This section describes the construction of a Gaussian mixture model for each tag in a visual feature space. We first use a target tag v_i as a query and crawl a set of N_i images $\{I_1, I_2, \dots, I_{N_i}\}$ from Flickr. We then calculate a d -dimensional visual feature vector $\mathbf{x}_n$ from each image I_n ($n = 1, \dots, N_i$), and denote the set of visual features by $X_i = \{\mathbf{x}_n\}_{n=1}^{N_i}$. Let us consider a problem for construct a GMM based on the set X_i . The probability density function of a GMM can be represented using K Gaussian distributions as follows:

$$f(\mathbf{x}_n|\Theta_i) = \sum_{k=1}^K \pi_{i,k} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{i,k}, \Sigma_{i,k}), \quad (5.1)$$

where

$$\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_{i,k}, \Sigma_{i,k}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_{i,k}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_{i,k})^T \Sigma_{i,k}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_{i,k}) \right\},$$

and $\boldsymbol{\mu}_{i,k}$, $\Sigma_{i,k}$, $\pi_{i,k}$ are the mean vector, covariance matrix, and coefficient of the k -th component of the GMM for v_i , respectively; and $\Theta_i = \{\Sigma_{i,1}, \dots, \Sigma_{i,K}, \boldsymbol{\mu}_{i,1}, \dots, \boldsymbol{\mu}_{i,K}, \pi_{i,1}, \dots, \pi_{i,K}\}$ represents a set of parameters for all components. To find the parameters Θ_i using X_i , we maximize the following log-likelihood function:

$$\ln p(X|\Phi_i) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_{i,k} \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_{i,k}, \Sigma_{i,k}) \right\}. \quad (5.2)$$

In this paper, we use the Expectation-Maximization (EM) algorithm [146] to estimate the parameters. The EM algorithm consists of two steps, called the E-step and the M-step, as given below.

E step:

Using the current parameter set Σ_i , we calculate the *responsibility* for the k -th component of the GMM as follows:

$$\gamma_{i,k}^n = \frac{\pi_{i,k} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_{i,k}, \Sigma_{i,k})}{\sum_{j=1}^K \pi_{i,j} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_{i,j}, \Sigma_{i,j})} \quad (5.3)$$

M step:

Based on the responsibilities calculated in E-step, we recalculate the parameters of the model as follows:

$$\boldsymbol{\mu}_{i,k} = \frac{\sum_{n=1}^N \gamma_{i,k}^n \mathbf{x}_n}{\sum_{n=1}^N \gamma_{i,k}^n} \quad (5.4)$$

$$\Sigma_{i,k} = \frac{\sum_{n=1}^N \gamma_{i,k}^n (\mathbf{x}_n - \boldsymbol{\mu}_{i,k})(\mathbf{x}_n - \boldsymbol{\mu}_{i,k})^T}{\sum_{n=1}^N \gamma_{i,k}^n} \quad (5.5)$$

$$\pi_{i,k} = \frac{1}{N} \sum_{n=1}^N \gamma_{i,k}^n \quad (5.6)$$

The EM algorithm iterates over these two steps until convergence is reached. This gives us a set of parameters for tag v_i .

To find an optimal value of K , we can choose a model that maximizes a cost function. There have been presented several cost functions. The most commonly used measures include Akaike's information criterion (AIC), Bayesian information criteria (BIC), and minimum description length (MDL). For example, the authors in [5] exploit MDL for determining the value of K for visual tag dictionary construction. In this chapter, we experimentally set the number of components, and leave the selection of these criteria to future work.

5.4 Finding Time-varying Tags Based on Visual Features

This section presents two types of measures that represent tag characteristics over time, which are calculated based on the GMM updates. Figure 5.1 shows an overview of the proposed method. In the proposed method, we first collect a set of images associated with timestamps by querying a photo sharing website using a target tag. We use a set of images associated with each tag within a certain time period as an initial dataset, from

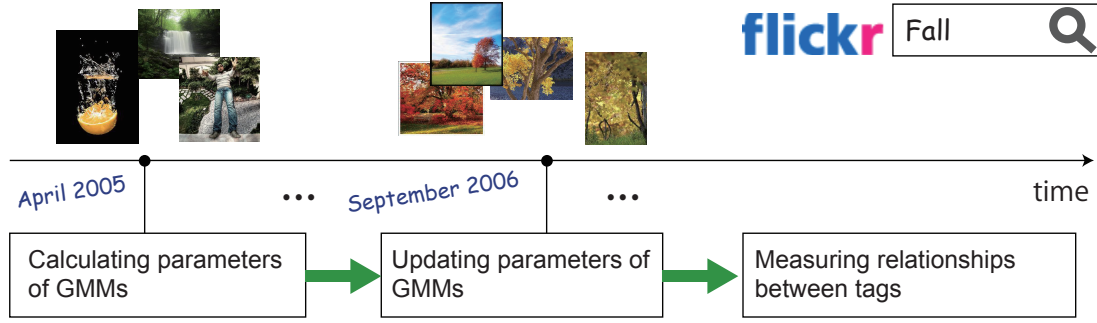


Fig. 5.1 An overview of the proposed method.

which parameters of a GMM are calculated as described in Section 5.3. Given a new image for the target tag, we recursively update the parameters of the GMM (Section 5.4.1), as shown in Fig. 5.1. Based on the updated parameters of GMM, we present measures that represent semantic relations between tags (Section 5.4.2).

5.4.1 Recursive update of GMM parameters

If the visual aspects of tag v_i are changing over time, its distribution in a visual feature space would also be different at several periods. Thus, we expect that an adaptive estimation of GMM parameters can contribute to capturing the changing semantics of the tag. Under this assumption, we exploit a recursive update of the GMM parameters [147]. Given a new image I_{new} tagged with t_i , and its timestamp t_{new} , we first calculate its visual feature vector $\mathbf{x}_{new}$. Then, for the vector $\mathbf{x}_{new}$, we calculate the responsibility for the k -th component of the current model as follows:

$$\gamma_{i,k}(\mathbf{x}_{new}) = \frac{\pi_{i,k} \mathcal{N}(\mathbf{x}_{new} | \boldsymbol{\mu}_{i,k}, \Sigma_{i,k})}{\sum_{j=1}^K \pi_{i,j} \mathcal{N}(\mathbf{x}_{new} | \boldsymbol{\mu}_{i,j}, \Sigma_{i,j})}. \quad (5.7)$$

Based on the calculated responsibility, we update the sum of the previous responsibilities sp_k for the k -th component as follows:

$$sp_{i,k} = sp_{i,k} + \gamma_{i,k}(\mathbf{x}_{new}). \quad (5.8)$$

Using the above equation, we recalculate the parameters of the k -th component as follows:

$$\boldsymbol{\mu}_{i,k} = \boldsymbol{\mu}_{i,k} + \frac{\gamma_{i,k}(\mathbf{x}_{new})}{sp_{i,k}}(\mathbf{x}_{new} - \boldsymbol{\mu}_{i,k}) \quad (5.9)$$

$$\begin{aligned} \Sigma_{i,k} = & \Sigma_{i,k} - (\boldsymbol{\mu}_{i,k} - \boldsymbol{\mu}_{i,k}^{old})(\boldsymbol{\mu}_{i,k} - \boldsymbol{\mu}_{i,k}^{old})^T \\ & + \frac{\gamma_{i,k}(\mathbf{x}_{new})}{sp_{i,k}} \{(\mathbf{x}_{new} - \boldsymbol{\mu}_{i,k})(\mathbf{x}_{new} - \boldsymbol{\mu}_{i,k})^T - \Sigma_{i,k}\} \end{aligned} \quad (5.10)$$

$$\pi_k = \frac{sp_{i,k}}{\sum_{j=1}^K sp_{i,j}} \quad (5.11)$$

The above procedures enable the recursive update of GMM parameters at time t_{new} . The adaptively updated model thus enables the extraction of tag relationships considering time information.

5.4.2 Extracting semantic relationships between tags based on the updated GMMs

This subsection presents a method for extracting semantic relationships between tags based on the updated parameters of GMMs. We first calculate a measure that represents semantic relationships between two tags v_i and v_j , which uses the similarity between distributions in a visual feature space at time t . In the proposed method, we exploit Kullback-Leibler (KL) divergence as a metric between two probabilistic distributions. Because there is no closed-form for KL divergence between GMMs, we exploit an approximation method presented in [148]. According to [148], we first find the component of a GMM of tag v_j , which is the most similar to the a -th component of a GMM of tag v_i , as follows:

$$m_i(a) = \arg \min_b \{D_{KL}(\boldsymbol{\mu}_{j,b}, \Sigma_{j,b} || \boldsymbol{\mu}_{i,a}, \Sigma_{i,a}) - \log \pi_{j,b}\}, \quad (5.12)$$

where $D_{KL}(\boldsymbol{\mu}_{j,b}, \Sigma_{j,b} || \boldsymbol{\mu}_{i,a}, \Sigma_{i,a})$ represents the KL divergence between two normal distributions $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{i,a}, \Sigma_{i,a})$ and $\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_{j,b}, \Sigma_{j,b})$, which is calculated using the following equation:

$$\begin{aligned} D_{KL}(\boldsymbol{\mu}_{j,b}, \Sigma_{j,b} || \boldsymbol{\mu}_{i,a}, \Sigma_{i,a}) = & \frac{1}{2} \left\{ \log |\Sigma_{i,a} \Sigma_{j,b}^{-1}| + tr(\Sigma_{i,a} \Sigma_{j,b}^{-1}) \right. \\ & \left. + (\boldsymbol{\mu}_{i,a} - \boldsymbol{\mu}_{j,b})^T \Sigma_{j,b}^{-1} (\boldsymbol{\mu}_{i,a} - \boldsymbol{\mu}_{j,b}) - d \right\}. \end{aligned} \quad (5.13)$$

Based on the KL divergence between two components, we approximate the KL divergence between two GMMs using the following equation:

$$D_g(\Phi_j || \Phi_i) = \pi_{i,a} \left\{ D_{KL}(\boldsymbol{\mu}_{j,m_i(a)}, \Sigma_{j,m_i(a)} || \boldsymbol{\mu}_{i,a}, \Sigma_{i,a}) + \log \frac{\pi_{i,a}}{\pi_{j,m_i(a)}} \right\}. \quad (5.14)$$

Because KL divergence is an asymmetric measure, the proposed method defines a measure that represents the semantic relatedness between tags v_i and v_j to be symmetric, as follows:

$$\begin{aligned} dist(\Phi_i, \Phi_j) \\ = \frac{1}{2} \left\{ D_g(\Phi_i || \Phi_j) + D_g(\Phi_j || \Phi_i) \right\} \end{aligned} \quad (5.15)$$

A small value of $dist(\Phi_i, \Phi_j)$ means that tags v_i, v_j are related within a time period corresponding to time t .

Next, we present a measure that represents a tag's semantic changes over time. The proposed method assumes that if visual feature vectors from a certain time period significantly change the model of a target tag, semantics of the tag are changed at the time period. Based on this assumption, we calculate the KL divergence between two GMMs that are before and after the parameter update, respectively, as a degree of model changes. Let Δt be a time span, and denote parameter sets of GMMs for tag v at time t and at time $t + \Delta t$ by Φ^t and $\Phi^{t+\Delta t}$, respectively. Then, the KL divergence between the two GMMs is calculated as follows:

$$\Omega(t, \Delta) = \frac{1}{2} \left\{ D_g(\Phi^t || \Phi^{t+\Delta t}) + D_g(\Phi^{t+\Delta t} || \Phi^t) \right\} \quad (5.16)$$

A large $\Omega(t, \Delta)$ means that the visual aspects of tag v significantly change within a time period $(t, t + \Delta t)$. Plotting the values for all time t lets us understand the visual changes of tags over time.

5.5 Experiments

This section provides the results of experiments conducted on Web images associated with timestamps to show the effectiveness of our two types of measures. Section 5.5.1 describes details of the dataset used. Section 5.5.2 shows examples of analysis using the

Table 5.1 Tags and their numbers of images in the dataset.

Tag	Number of images
apple	121,436
jaguar	159,729
river	132,079
sunset	144,428
blackberry	39,508

first measure for tag characteristics over time, and Section 5.5.3 shows examples of results obtained using the second measure.

5.5.1 Dataset

We used the Flickr API¹ to construct a dataset for our experiments. The query tags used for dataset construction and their numbers of collected images are listed in Table 5.1. All images were captured between January 2006 and April 2012.

In the experiments, we used RGB-SIFT [114] to describe the visual features of the tagged images. Specifically, the SIFT features [115] that were independently extracted from each RGB channel were combined into 384-dimensional features. The bag-of-words (BoW) model was used for the vector quantization of the obtained features in an image, which produced a 500-dimensional vector per image. We used image collections that were captured from January 2006 to July 2006 as initial datasets for GMMs. The number of components of GMMs were experimentally set to $K = 20$.

5.5.2 Analyzing a tag's visual changes over time

First, we show results obtained using the measure that represents the content changes of tags over time. In this experiment, we segment the timestamps into time periods to let the time span Δt be 14 days (2 weeks), and calculated Eq. 5.15. As an example of our results, Fig. 5.2 shows the changes of degrees for the tags “apple” and “sunset” and their example images at each time period. As shown in Fig. 5.2, the measure calculated for “apple” has peaks at several time periods. From these peaks and their corresponding images, we can find multiple events that change the visual aspects of the tag. For example, within time period A, a new device from Apple Inc.² was announced, and images of the device significantly increased. Within time period B, many images that represent fruits

¹<http://www.flickr.com/services/api/>

²<https://www.apple.com/>

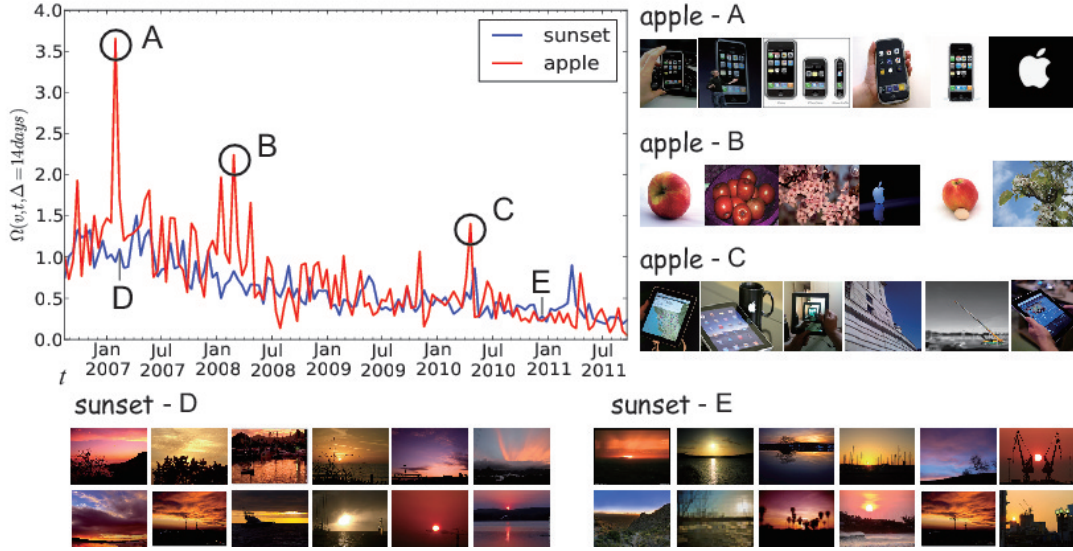


Fig. 5.2 Results obtained using the first measure for tags “apple” and “sunset”, in which their example images at each time period are shown.

or trees were taken, which resulted in increasing visual diversity of the tag. Furthermore, within time period C, in addition to a new device’s announcement, images taken using the new device were also tagged with “apple”: this makes semantics of the tag more ambiguous. In contrast, the tag “sunset” shows few changes of its visual aspects over time, compared with “apple”. We observed that images taken within time period D do not have a large difference in visual aspects from images taken within time period E. From these results, it can be considered that visual changes of tags over time are different from one another.

5.5.3 Finding tag relationship changes

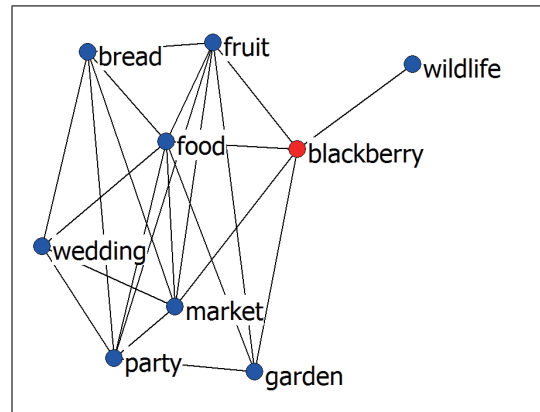
Second, we show the results obtained using the proposed measure that represents similarity between tags at each time period. In this experiment, we segment the timestamps into time periods to let the time span Δt be 28 days (4 weeks), and calculated Eq. 5.15. We compare the proposed method with the following comparative method: for each tag, a set of images is randomly sampled from an entire image collection, from which a GMM was constructed. This comparative method does not use temporal information for measuring relatedness between tags and is the same approach as in the conventional method [5]. Extracted tag relationships using each method were visualized using NetDraw [92], where each node represents a tag, and each edge represents the strong relationship be-

tween the tags. Figure 5.3 shows tag networks around “blackberry,” which were constructed using the proposed and comparative methods. In this figure, only the highest 15 relationships around the tag are depicted as edges for easier viewing. In Fig. 5.3 (a), the extracted relationships around “blackberry” shows food-related concepts. On the other hand, in Fig. 5.3 (b), the extracted relationships are different from those of Fig. 5.3 (a): new concepts related to the mobile phone appear. Compared with results shown in Fig. 5.3 (c), our method can find such tag relationship changes using temporal information of tagged images.

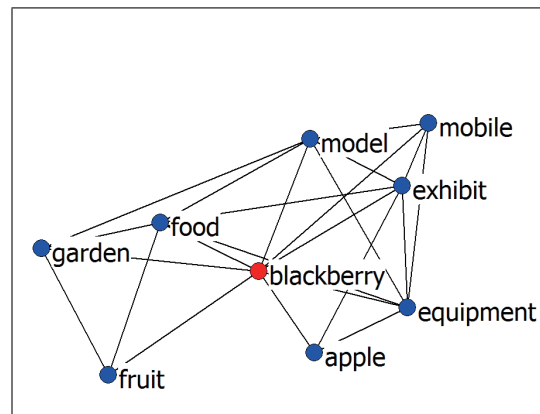
Figure 5.4 shows tag networks around “jaguar,” which were constructed using the proposed and comparative methods. Similar to the results in Fig. 5.3, our method extracts different two tag relationships at two different time periods. For example, in Fig. 5.4 (a), the extracted tag relationships describe some car-related concepts. It is highly likely that such relationships were derived from car show events. In Fig. 5.4 (b), concepts related to animals are shown. Tag relationships shown in Fig. 5.4 (c) consist of concepts similar to those of Fig. 5.4 (a). Considering such tag relationship changes over time, to improve performance of multimedia-related applications such as image retrieval and annotation, we propose that knowledge about tags should be extracted adaptively from social media websites.

In contrast to the abovementioned results, we observed that there might exist tags that do not significantly change relationships with other tags over time. Figure 5.5 shows tag networks around “river,” which were constructed using the proposed method and comparative method. All networks shown in this figure are similar to one another, and the tag “river” is time-stable compared with tags “blackberry” and “jaguar.”

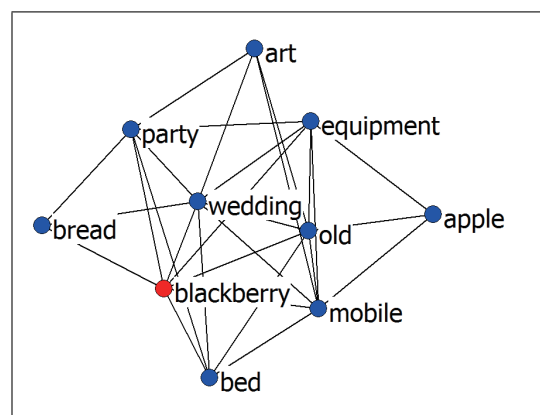
From these results, we can consider that considering temporal information enables better image retrieval and annotation. To investigate the availability of the extracted information, more in-depth experiments should be performed based on a larger dataset. Our future work includes large-scale additional experiments to investigate the characteristics of a large number of tags. In addition, quantitative evaluation using crowdsourcing tools such as the Amazon Mechanical Turk will be performed in future work.



(a) Concept network generated by the proposed method within a time period starting from 8/1/2006 to 8/29/2006.

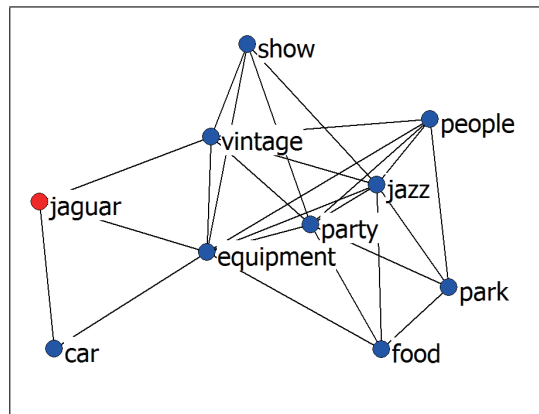


(b) Concept network generated by the proposed method within a time period starting from 9/21/2010 to 10/19/2010.

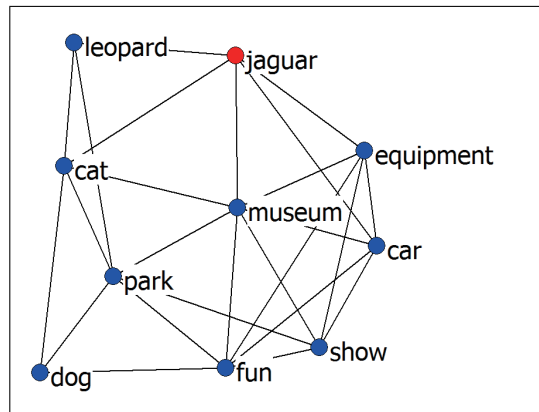


(c) Concept network generated by the comparative method.

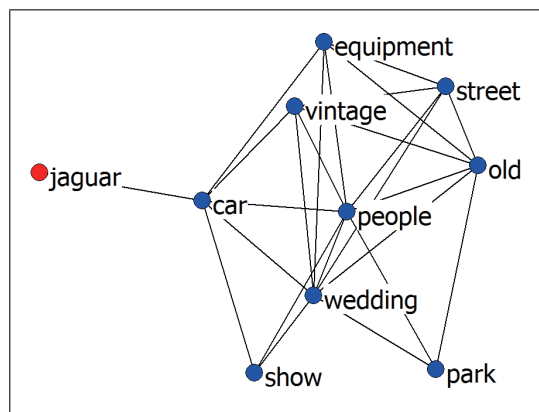
Fig. 5.3 Concept networks around “blackberry,” whose edges are calculated using the proposed method and the comparative method. Each node represents a tag, and the edges between the concepts represent the strongest relationships.



(a) Concept network generated by the proposed method from 8/29/2006 to 9/26/2006.

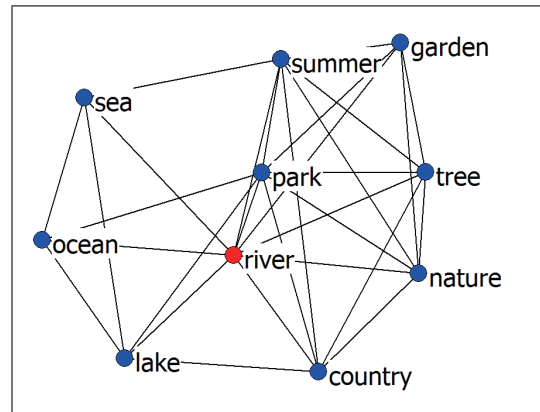


(b) Concept network generated by the proposed method from 8/21/2012 to 9/18/2012.

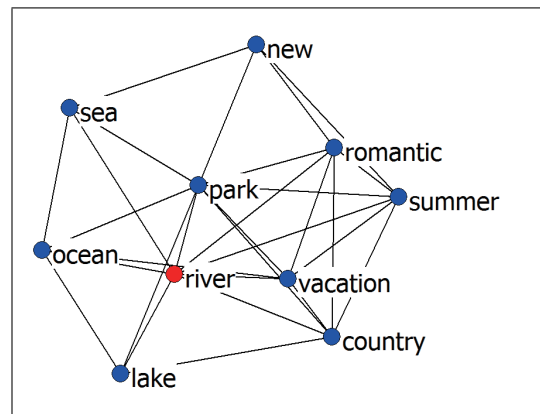


(c) Concept network generated by the comparative method.

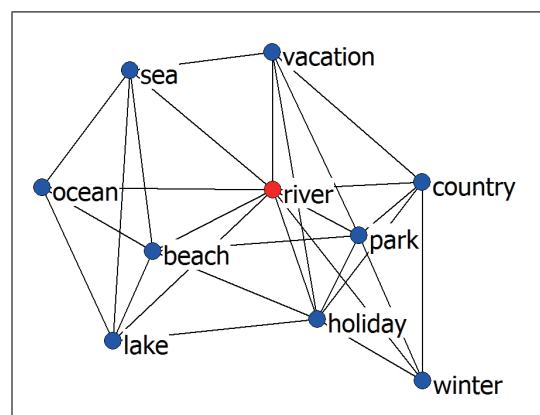
Fig. 5.4 Concept networks around “jaguar,” whose edges are calculated using the proposed method and the comparative method. Each node represents a concept, and the edges between the concepts represent the strongest relationships.



(a) Concept network generated by the proposed method within a time period starting from 8/1/2006 to 8/29/2006.



(b) Concept network generated by the proposed method within a time period starting from 8/21/2012 to 9/18/2012.



(c) Concept network generated by the comparative method.

Fig. 5.5 Concept networks around “jaguar,” whose edges are calculated using the proposed method and the comparative method. Each node represents a concept, and the edges between the concepts represent the strongest relations.

5.6 Conclusion and Future Work

Conclusion

In this chapter, we exploit the temporal information of tagged images, i.e., timestamps, for finding tag relationship changes over time. We collect a set of images for each tag from Flickr, and calculate a GMM in a visual feature space. Given a new image over time, we recursively update the parameters of the GMM. Based on the updated parameters of GMMs, we calculate two types of measures: (i) the degree of visual change of the tag over time, and (ii) the visual similarity between tags over time. In experiments conducted using Flickr tagged images, we showed results obtained using the presented measures and compared the performance with a comparative method that does not exploit timestamps of images. These results show that considering timestamps is effective for extracting new semantic relations between tags.

Future work

There is still room for improvement in the method presented in this chapter. First, the number of components of a GMM, which were set experimentally in the experiments, should be determined appropriately for each tag. In future work, we will introduce several types of visual features to improve the performance of the proposed method.

In this chapter, we used visual features extracted from images to analyze the changes of visual aspects of tags. As presented in Chapter 4, using both textual and visual features enables effectively finding changes of tag semantics over time. If a cross-modal approach is designed for finding tag relationship changes over time, we will be able to effectively extract remarkable phenomena as events.

In addition to the abovementioned issues, we should validate the effectiveness of exploiting temporal information and investigate the availability of the extracted knowledge about concepts over time on a larger image collection. Our future work includes additional experiments to investigate the characteristics of a large number of tags. In addition, quantitative evaluation using crowdsourcing such as the Amazon Mechanical Turk is required to compare the performance with conventional methods.

Chapter 6

Conclusions and Discussion

This dissertation presents novel approaches for extracting semantic relations between concepts based on textual and visual features of tagged images on the Web. Under the research background described in Chapter 1, a literature review of semantic measures and their applications was given in Chapter 2, in which I pointed out several problems the conventional methods hold and described how to solve those problems. In this chapter, I again summarize the contributions and key observations of our work, and suggest possible directions of future work that go beyond our current achievements.

6.1 Key Observations and Contributions

Our work aims to effectively use the features extracted from tagged images on the Web, which can be seen as one of studies for developing multimedia data mining technologies. The main contribution of this thesis is two-fold: (i) the collaborative use of textual and visual features, and (ii) the use of temporal information of Web images. Below, I summarize the key observations obtained from the experiments and contributions that can be seen from each of the two points of views.

The collaborative use of textual and visual features for inter-concept relationship extraction

In Chapter 3, a classifier-based approach for extracting inter-concept relationships from an image database was presented. In the proposed method, we focused on visual features shared by two concepts in the same image to capture hyponymy/hypernymy and

meronymy relations between the two concepts. This framework is achieved through classifier training for each concept. Experiments conducted on manually-labeled image datasets showed that the use of visual similarity and contextual information can improve the performance of extracting inter-concept relationships over conventional methods that exploit a single type of information.

In Chapter 4, a cross-modal approach for extracting inter-concept relationships based on both textual and visual features of tagged images on the Web was presented. In the proposed method, we first project these two types of features to a latent space, and then calculate a representative distribution of each concept in the space. Based on the representative distributions of concepts, we present two types of semantic measures: semantic relatedness between concepts and abstraction levels of concepts. In addition, a novel visualization method was presented to facilitate the intuitive understandings of inter-concept relationships in a target dataset. Experiments conducted on a large image dataset collected from Flickr showed that our approach can improve the performance of extracting semantic relationships over conventional methods that exploit single features only.

Temporal analysis of Web image collections

Beside the abovementioned methods, which work on static datasets, an analysis of temporal dynamics of visual features for each tag has been discussed in Chapter 5. Our work can be seen as one of the early studies to focus on temporal evolution of Web image collections. In the proposed method, we exploited the timestamps of images as an additional feature, and calculated representative distributions of concepts for each time period. Experiments conducted on an image dataset collected from Flickr showed that considering temporal information enables extraction of new knowledge about concepts, knowledge that cannot be found by conventional methods that work on static datasets. I believe that this direction of research will become more important for multimedia data mining in the Internet era.

6.2 Future Directions

In this dissertation, I proposed approaches that enable the extraction of semantic knowledge about inter-concept relationships from tagged images on the Web. In spite of their achievements in multimedia data mining, there are still several points to be solved along

this research. Here, I describe several ideas for improving our approach and suggest some possible directions for future research projects.

Improving the proposed frameworks

There is still room for performance improvement of the approaches presented in this thesis. First of all, I must describe the details of such insufficient points of each approach. To improve the performance of the first approach that aims to capture semantic similarity between concepts, the utilization of non-linear classifiers, including SVM, can be effective. We should investigate various classifiers for determining their suitability for the proposed framework. In addition, the effective use of classifier parameters for selecting relevant features will be studied. The presented classifier-based approach assumes the use of an image database that is accurately labeled. When using a dataset collected from social media websites, we need to apply techniques that remove noisy tags from the images to improve the tag quality.

To extend the applicability of the second approach, we should investigate how to automatically determine the variable q that will ensure that reasonable performance can be maintained with high efficiency. Another point to improve is modeling a concept in a latent space: the representative distribution was summarized by one normal distribution in the proposed method. If it can be effectively approximated by more complex distributions such as mixture models, the semantics of each concept will be described better.

We observed through experiments that analyzing millions of images requires much computation time. To further develop the third approach that aims to find concept relationship changes as salient phenomena from Web image streams, we should present a more computationally efficient method. For example, an update scheme for the distribution of each tag should be more computationally efficient. We must tackle real-time processing for analyzing Web image streams.

As a general problem for all approaches that this thesis has presented, we need to investigate the characteristics of various types of visual features in experiments. If visual descriptors can represent the image content better, an overall performance improvement can be expected. The total computation time of each method can also be improved by introducing rapid implementations of visual feature extraction, which is the same strategy employed in visual categorization tasks [124, 125]. We will introduce schemes for reducing the computational cost of the proposed method in future work.

The effectiveness of methods that use tagged images as an information source is dependent on the dataset. If the dataset possesses some content bias, e.g., many images

belong to specific categories, the accurate extraction of the general relations between the concepts will become difficult. We plan to tackle this bias problem when collecting tagged images in our future work.

Developing applications for multimedia-related tasks

For multimedia-related tasks such as annotation, retrieval, and recommendation, we should prepare a high-quality concept vocabulary. If the vocabulary has a bias such that the number of concepts that belong to each category differs widely from other categories, some of the concepts for annotation, which are in the minority, might not benefit from the semantic measure. Thus, we should discuss and tackle this problem and in future work.

The research theme of this thesis exists to introduce knowledge of the world to computational techniques as initially described in Chapter 1. The measure presented in Chapter 4, which can quantify the abstraction levels of concepts, has potential for developing hierarchical frameworks in annotation and retrieval. For example, when given a concept with a high abstraction level as a search query, we suggest more specific concepts that are relevant to the query for efficient retrieval. Following this idea, we will develop a new image retrieval scheme that effectively considers the abstraction level of a given query concept.

Effective visualization of extracted knowledge

The effective visualization of knowledge extracted using our approaches should be studied well to facilitate user understandings. Although we presented a novel visualization method of the inter-concept relationships, the visualization is still limited to static information; i.e., it cannot effectively show the temporal dynamics of the inter-concept relationships. Thus, another interesting future direction for analyzing Web image streams would be to develop ways for interactive visualization of the extracted temporal information.

In this dissertation, I first described several important characteristics of today's image sharing on the Web, which are outcomes of recent technical developments in digital image acquisition devices, high-speed Internet connection, and Web 2.0 applications. I

hope that this dissertation can contribute to future studies in the field of multimedia data mining.

References

- [1] G. A. Miller, “WordNet: a lexical database for english,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.
- [2] F. R. Bach and M. I. Jordan, “A probabilistic interpretation of canonical correlation analysis,” Department of Statistics, University of California, Berkeley, Tech. Rep. 688, 2005.
- [3] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H. J. Zhang, “Tag ranking,” in *Proc. Int. Conf. World wide web (WWW)*, 2009, pp. 351–360.
- [4] L. Wu, X.-S. Hua, N. Yu, W.-Y. Ma, and S. Li, “Flickr distance: A relationship measure for visual concepts,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 5, pp. 863–875, May 2012.
- [5] M. Wang and K. Yang, “Constructing visual tag dictionary by mining community-contributed media corpus,” *Neurocomputing*, vol. 95, pp. 3–10, Oct. 2012.
- [6] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, “The IAPR benchmark: A new evaluation resource for visual information systems,” in *Proc. Int. Conf. Language Resources and Evaluation*, 2006.
- [7] J. Jeon, V. Lavrenko, and R. Manmatha, “Automatic image annotation and retrieval using cross-media relevance models,” in *Proc. Int. ACM SIGIR Conf. Research and Development in Informaion Retrieval (SIGIR)*, 2003, pp. 119–126.
- [8] Y. G. Jiang, J. Wang, S. F. Chang, and C. W. Ngo, “Domain adaptive semantic diffusion for large scale context-based video annotation,” in *Proc. Int. Conf. Computer Vision (ICCV)*, 2009, pp. 1420–1427.
- [9] M. J. Huiskes, B. Thomee, and M. S. Lew, “New trends and ideas in visual concept detection: The MIR flickr retrieval evaluation initiative,” in *Proc. Int. Conf. Multimedia Information Retrieval (MIR)*, 2010, pp. 527–536.
- [10] J. Gantz and D. Reinsel, “The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east,” <http://idcdocserv.com/1414>, Dec 2012, last accessed: May 2014.
- [11] G. Lakoff and M. Johnson, *Philosophy in the Flesh : The Embodied Mind and Its Challenge to Western Thought*. Basic Books, Dec. 1999.

- [12] H.-H. Chen, M.-S. Lin, and Y.-C. Wei, "Novel association measures using web search with double checking," in *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006, pp. 1009–1016.
- [13] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Measuring semantic similarity between words using web search engines," in *Proc. Int. Conf. World Wide Web (WWW)*, 2007, pp. 757–766.
- [14] R. Cilibiasi and P. Vitanyi, "The google similarity distance," *IEEE Trans. Knowledge and Data Engineering*, vol. 19, no. 3, pp. 370–383, Mar. 2007.
- [15] W. Wang, P. Barnaghi, and A. Bargiela, "Probabilistic topic models for learning terminological ontologies," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 7, pp. 1028–1040, July 2010.
- [16] flickrBLOG, "6000000000," <http://blog.flickr.net/en/2011/08/04/6000000000/>, Aug 2011, last accessed: May 2014.
- [17] Y. Lu, L. Zhang, J. Liu, and Q. Tian, "Constructing concept lexica with small semantic gaps," *IEEE Trans. Multimedia*, vol. 12, no. 4, pp. 288–299, Jun. 2010.
- [18] A. Budanitsky and G. Hirst, "Evaluating WordNet-based measures of lexical semantic relatedness," *Journal Computational Linguistics*, vol. 32, no. 1, pp. 13–47, mar 2006.
- [19] S. P. Ponzetto and M. Strube, "Knowledge derived from Wikipedia for computing semantic relatedness," *Journal of Artificial Intelligence Research*, vol. 30, no. 1, pp. 181–212, Oct 2007.
- [20] J. Gracia and E. Mena, "Web-based measure of semantic relatedness," in *Proc. Int. Conf. Web Information Systems Engineering (WISE)*, 2008, pp. 136–150.
- [21] T. Zesch, "Study of semantic relatedness of words using collaboratively constructed semantic resources," Ph.D. dissertation, Universitäts-und Landesbibliothek Darmstadt, Feb. 2010.
- [22] R. El-Yaniv and D. Yanay, "Supervised learning of semantic relatedness," in *Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2012, pp. 744–759.
- [23] T. Bittner and B. Smith, "A theory of granular partitions," in *Foundations of Geographic Information Science*, M. Duckham, M. F. Goodchild, and M. F. Worboys, Eds. London: Taylor & Francis, 2002.
- [24] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Trans. Systems, Man and Cybernetics*, vol. 19, no. 1, pp. 17–30, Jan 1989.
- [25] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. the 32nd annual meeting on Association for Computational Linguistics (ACL)*, 1994, pp. 133–138.

- [26] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," in *WordNet: An electronic lexical database*, C. Fellbaum, Ed. MIT Press, 1998, pp. 265–283.
- [27] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proc. Int. Joint Conf. Artificial intelligence (IJCAI)*, 1995, pp. 448–453.
- [28] D. Lin, "An information-theoretic definition of similarity," in *Proc. Int. Conf. Machine Learning (ICML)*, 1998, pp. 296–304.
- [29] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. Int. Conf. Research on Computational Linguistics*, 1997, pp. 19–33.
- [30] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone," in *Proc. Int. Conf. Systems Documentation (SIGDOC)*, 1986, pp. 24–26.
- [31] S. Patwardhan, "Using wordnet-based context vectors to estimate the semantic relatedness of concepts," in *Proc. EACL 2006 Workshop on Making Sense of Sense: Bringing Computational Linguistics and Psycholinguistics Together*, 2006, pp. 1–8.
- [32] X. Y. Wei and C. W. Ngo, "Ontology-enriched semantic space for video search," in *Proc. Int. Conf. Multimedia (MM)*, 2007, pp. 981–990.
- [33] X. Y. Wei, C. W. Ngo, and Y. G. Jiang, "Selection of concept detectors for video search by ontology-enriched semantic spaces," *IEEE Trans. Multimedia*, vol. 10, no. 6, pp. 1085–1096, 2008.
- [34] M. Strube and S. P. Ponzetto, "WikiRelate! computing semantic relatedness using Wikipedia," in *Proc. National Conf. Artificial Intelligence (AAAI)*, 2006, pp. 1419–1424.
- [35] T. Zesch and I. Gurevych, "Wisdom of crowds versus wisdom of linguists – measuring the semantic relatedness of words," *Journal Natural Language Engineering*, vol. 16, no. 1, pp. 25–59, Jan. 2010.
- [36] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using wikipedia-based explicit semantic analysis," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2007, pp. 1606–1611.
- [37] D. Milne and I. H. Witten, "An effective, low-cost measure of semantic relatedness obtained from Wikipedia links," in *Proc. Conf. Artificial Intelligence (AAAI)*, 2008, pp. 25–30.
- [38] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. M. Petrakis, and E. E. Milios, "Semantic similarity methods in WordNet and their application to information retrieval on the web," in *Proc. ACM Int. Workshop on Web Information and Data Management (WIDM)*, 2005, pp. 10–16.

- [39] J. Hu, L. Fang, Y. Cao, H.-J. Zeng, H. Li, Q. Yang, and Z. Chen, "Enhancing text clustering by leveraging wikipedia semantics," in *Proc. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, 2008, pp. 179–186.
- [40] S. Hassan and R. Mihalcea, "Cross-lingual semantic relatedness using encyclopedic knowledge," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2009, pp. 1192–1201.
- [41] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in *Proc. Int. Conf. Information and Knowledge Management (CIKM)*, 2009, pp. 919–928.
- [42] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," *ACM Trans. Information Systems*, vol. 29, no. 2, pp. 8:1–8:34, Apr. 2011.
- [43] P. Schmitz, "Inducing ontology from flickr tags," in *Proc. Collaborative Web Tagging Workshop*, 2006, pp. 210–214.
- [44] M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitanyi, "The similarity metric," *IEEE Trans. Information Theory*, vol. 50, no. 12, pp. 3250–3264, Dec 2004.
- [45] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. Conf. Uncertainty in Artificial Intelligence (UAI)*, 1999, pp. 289–296.
- [46] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.
- [47] N. Ni, K. Liu, and Y. Li, "An automatic multi-domain thesauri construction method based on LDA," in *Proc. Int. Conf. Machine Learning and Applications and Workshops (ICMLA)*, vol. 2, Dec 2011, pp. 235–240.
- [48] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, Dec 2000.
- [49] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 5:1–5:60, May 2008.
- [50] K. Yanai and K. Barnard, "Image region entropy: a measure of "visualness" of web images associated with one concept," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2005, pp. 419–422.
- [51] A. Sun and S. S. Bhowmick, "Quantifying tag representativeness of visual content of social images," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 471–480.
- [52] J. Fan, X. He, N. Zhou, J. Peng, and R. Jain, "Quantitative characterization of semantic gaps for learning complexity estimation and inference model selection," *IEEE Trans. Multimedia*, vol. 14, no. 5, pp. 1414–1428, 2012.

- [53] Z. Xu, X.-J. Wang, and C. W. Chen, "Mining visualness," in *Proc. Int. Conf. Multimedia and Expo (ICME)*, July 2013, pp. 1–6.
- [54] J. Fan, Y. Shen, C. Yang, and N. Zhou, "Structured max-margin learning for inter-related classifier training and multilabel image annotation," *IEEE Trans. Image Processing*, vol. 20, no. 3, pp. 837–854, March 2011.
- [55] P. Dong, K. Mei, N. Zheng, H. Lei, and J. Fan, "Training inter-related classifiers for automatic image classification and annotation," *Pattern Recognition*, vol. 46, no. 5, pp. 1382–1395, 2013.
- [56] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.
- [57] H. Nakayama, T. Harada, and Y. Kuniyoshi, "Evaluation of dimensionality reduction methods for image auto-annotation," in *Proc. British Machine Vision Conference (BMVC)*, 2010, pp. 94.1–94.12.
- [58] D. R. Hardoon, C. Saunders, S. Szedmak, and J. Shawe-Taylor, "A correlation approach for automatic image annotation," in *Proc. Int. Conf. Advanced Data Mining and Applications (ADMA)*, 2006, pp. 681–692.
- [59] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE MultiMedia*, vol. 13, no. 3, pp. 86–91, July 2006.
- [60] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
- [61] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image annotations by combining multiple evidence & WordNet," in *Proc. Int. Conf. Multimedia (MM)*, 2005, pp. 706–715.
- [62] J. Liu, B. Wang, M. Li, Z. Li, W. Ma, H. Lu, and S. Ma, "Dual cross-media relevance model for image annotation," in *Proc. Int. Conf. Multimedia (MM)*, 2007, pp. 605–614.
- [63] M. Marszalek and C. Schmid, "Semantic hierarchies for visual object recognition," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2007, pp. 1–7.
- [64] J. Fan, Y. Gao, and H. Luo, "Integrating concept ontology and multitask learning to achieve more effective classifier training for multilevel image annotation," *IEEE Trans. Image Processing*, vol. 17, no. 3, pp. 407–426, Mar. 2008.
- [65] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.
- [66] R. M. T. W. L. Page, S. Brin, "The PageRank citation ranking: Bringing order to the web." Stanford University InfoLab, Tech. Rep. 688, 1999.

- [67] C. Wang, F. Jing, L. Zhang, and H. J. Zhang, "Image annotation refinement using random walk with restarts," in *Proc. ACM Int. Conf. Multimedia*, 2006, pp. 647–650.
- [68] Z. Wang, G. Guan, Y. Qiu, L. Zhuo, and D. Feng, "Semantic context based refinement for news video annotation," *Multimedia Tools and Applications*, pp. 1–21, Mar. 2012.
- [69] Y. Wang and S. Gong, "Refining image annotation using contextual relations between words," in *Proc. ACM Int. Conf. Image and Video Retrieval (CIVR)*, 2007, pp. 425–432.
- [70] G. Zhu, S. Yan, and Y. Ma, "Image tag refinement towards low-rank, content-tag prior and error sparsity," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 461–470.
- [71] R. Ji, H. Yao, X. Sun, B. Zhong, and W. Gao, "Towards semantic embedding in visual vocabulary," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2010, pp. 918–925.
- [72] D. Liu, S. Yan, X.-S. Hua, and H.-J. Zhang, "Image retagging using collaborative tag propagation," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 702–712, Aug 2011.
- [73] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 883–895, June 2012.
- [74] Y. G. Jiang, Q. Dai, J. Wang, C. W. Ngo, X. Xue, and S. F. Chang, "Fast semantic diffusion for large-scale context-based image and video annotation," *Image Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 3080–3091, June 2012.
- [75] A. Popescu, C. Millet, and P.-A. Moëllic, "Ontology driven content based image retrieval," in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, 2007, pp. 387–394.
- [76] M. Ferecatu, N. Boujemaa, and M. Crucianu, "Semantic interactive image retrieval combining visual and conceptual content description," *Multimedia Systems*, vol. 13, no. 5-6, pp. 309–322, 2008.
- [77] R. Lienhart, S. Romberg, and E. Hörster, "Multilayer pLSA for multimodal image retrieval," in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, 2009, pp. 9:1–9:8.
- [78] J. Deng, A. C. Berg, and L. Fei-Fei, "Hierarchical semantic indexing for large scale image retrieval," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 785–792.
- [79] T. Deselaers and V. Ferrari, "Visual and semantic similarity in ImageNet," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, June 2011, pp. 1777–1784.
- [80] B. Sigurbjörnsson and R. van Zwol, "Flickr tag recommendation based on collective knowledge," in *Proc. Int. Conf. World Wide Web (WWW)*, 2008, pp. 327–336.
- [81] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in *Proc. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, 2008, pp. 531–538.

- [82] N. Garg and I. Weber, "Personalized, interactive tag recommendation for flickr," in *Proc. ACM Conf. Recommender Systems (RecSys)*, 2008, pp. 67–74.
- [83] L. Wu, L. Yang, N. Yu, and X.-S. Hua, "Learning to tag," in *Proc. Int. Conf. World Wide Web (WWW)*, 2009, pp. 361–370.
- [84] F. Belém, E. Martins, T. Pontes, J. Almeida, and M. Gonçalves, "Associative tag recommendation exploiting multiple textual features," in *Proc. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, 2011, pp. 1033–1042.
- [85] S. C. H. Hoi, R. Jin, and M. R. Lyu, "Large-scale text categorization by batch mode active learning," in *Proc. Int. Conf. World Wide Web (WWW)*, 2006, pp. 633–642.
- [86] L. Kozma, A. Klami, and S. Kaski, "GaZIR: Gaze-based zooming interface for image retrieval," in *Proc. Int. Conf. Multimodal Interfaces (ICMI)*, ser. ICMI-MLMI '09, 2009.
- [87] L. Cao, J. Yu, J. Luo, and T. S. Huang, "Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression," in *Proc. Int. Conf. Multimedia (MM)*, 2009, pp. 125–134.
- [88] A. G. Hauptmann, M. G. Christel, and R. Yan, "Video retrieval based on semantic concepts," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 602–622, April 2008.
- [89] Y. S. Avrithis, A. D. Doulamis, N. D. Doulamis, and S. D. Kollias, "A stochastic framework for optimal key frame extraction from MPEG video databases," *Computer Vision and Image Understanding*, vol. 75, no. 1–2, pp. 3–24, 1999.
- [90] C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Scalable search-based image annotation of personal images," in *Proc. ACM Int. Workshop on Multimedia Information Retrieval (MIR)*, 2006, pp. 269–278.
- [91] X. Li, C. G. M. Snoek, and M. Worring, "Learning tag relevance by neighbor voting for social image retrieval," in *Proc. Int. Conf. Multimedia Information Retrieval (MIR)*, 2008, pp. 180–187.
- [92] S. P. Borgatti, "Netdraw software for network visualization," Analytic Technologies: Lexington, KY, 2002.
- [93] Z. Xia, X. Feng, J. Peng, and J. Fan, "Content-irrelevant tag cleansing via bi-layer clustering and peer cooperation," *Journal of Signal Processing Systems*, pp. 1–16, 2014.
- [94] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppín, "Placing search in context: the concept revisited," *ACM Trans. Information Systems*, vol. 20, no. 1, pp. 116–131, Jan. 2002.
- [95] A. Hotho, S. Staab, and G. Stumme, "Ontologies improve text document clustering," in *Proc. Int. Conf. Data Mining (ICDM)*, 2003, pp. 541 – 544.
- [96] C. G. M. Snoek, B. Huurnink, L. Hollink, M. de Rijke, G. Schreiber, and M. Worring, "Adding semantics to detectors for video retrieval," *IEEE Trans. Multimedia*, vol. 9, no. 5, pp. 975–986, 2007.

- [97] X. Y. Wei, Y. G. Jiang, and C. W. Ngo, "Concept-driven multi-modality fusion for video search," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 62–73, 2011.
- [98] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, 1936.
- [99] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.
- [100] M. Kim, "Correlation-based incremental visual tracking," *Pattern Recognition*, vol. 45, no. 3, pp. 1050 – 1060, 2012.
- [101] D. J. C. MacKay, *Information Theory, Inference & Learning Algorithms*. New York, NY, USA: Cambridge University Press, 2002.
- [102] J. V. Davis and I. Dhillon, "Differential entropic clustering of multi-variate Gaussians," in *Advances in Neural Information Processing Systems (NIPS)*. B. Scholkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 337–344.
- [103] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, Dec. 2005.
- [104] B. Chen, W. Lam, I. Tsang, and T. L. Wong, "Location and scatter matching for dataset shift in text mining," in *Proc. Int. Conf. Data Mining (ICDM)*, 2010, pp. 773–778.
- [105] J. Tao, F. Chung, and S. Wang, "On minimum distribution discrepancy support vector machine for domain adaptation," *Pattern Recognition*, vol. 45, no. 11, pp. 3962–3984, 2012.
- [106] X. Peng, W. Xu, and B. Wang, "Speaker clustering via novel pseudo-divergence of gaussian mixture models," in *Proc. Int. Conf. Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2005, pp. 111–114.
- [107] K. Abou-Moustafa and F. Ferrie, "A note on metric properties for some divergence measures: The gaussian case," *Journal of Machine Learning Research - Proceedings Track*, vol. 15, pp. 1–15, 2012.
- [108] V. González-Castro, R. Alaiz-Rodríguez, and E. Alegre, "Class distribution estimation based on the hellinger distance," *Information Sciences*, vol. 218, no. 0, pp. 146–164, 2013.
- [109] L. Ning, X. Jiang, and T. Georgiou, "On the geometry of covariance matrices," *IEEE Signal Processing Letters*, vol. 20, no. 8, pp. 787–790, 2013.
- [110] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. ACM Int. Conf. Multimedia Information Retrieval (MIR)*, 2008, pp. 39–43.
- [111] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, "Fusing concept detection and geo context for visual search," in *Proc. Int. Conf. Multimedia Retrieval (ICMR)*, 2012, pp. 4:1–4:8.

- [112] S. Bird, “NLTK: The natural language toolkit,” in *Proc. COLING/ACL on Interactive Presentation Sessions*, 2006, pp. 69–72.
- [113] M. F. Porter, “An algorithm for suffix stripping,” *Program*, vol. 3, no. 14, pp. 130–137, Oct. 1980.
- [114] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, Sept. 2010.
- [115] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.
- [116] G. Pirró, “A semantic similarity metric combining features and intrinsic information content,” *Data & Knowledge Engineering*, vol. 68, no. 11, pp. 1289 – 1308, 2009.
- [117] C. Biemann and S. Bordag, “Automatic acquisition of paradigmatic relations using iterated co-occurrences,” in *Proc. Int. Conf. Language Resources and Evaluation (LREC)*, 2004, pp. 967–970.
- [118] S. Chakraborti, N. Wiratunga, R. Lothian, and S. Watt, “Acquiring word similarities with higher order association mining,” in *Proc. Int. Conf. Case-Based Reasoning: Case-Based Reasoning Research and Development (ICCBR)*, 2007, pp. 61–76.
- [119] J. Radelaar, A. J. Boor, D. Vandic, J. W. Van Dam, F. Hogenboom, and F. Frasincar, “Improving the exploration of tag spaces using automated tag clustering,” in *Proc. Int. Conf. Web engineering (ICWE)*, 2011, pp. 274–288.
- [120] J. R. Wen, J. Y. Nie, and H. J. Zhang, “Query clustering using user logs,” *ACM Trans. Information Systems*, vol. 20, no. 1, pp. 59–81, Jan. 2002.
- [121] J. Y. Pan, H. J. Yang, C. Faloutsos, and P. Duygulu, “Automatic multimedia cross-modal correlation discovery,” in *Proc. Int. Conf. Knowledge discovery and data mining (KDD)*, 2004, pp. 653–658.
- [122] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “NUS-WIDE: a real-world web image database from National University of Singapore,” in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, 2009, pp. 48:1–48:9.
- [123] S. Zhu, G. Wang, C. W. Ngo, and Y. G. Jiang, “On the sampling of web images for learning visual concept classifiers,” in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, 2010, pp. 50–57.
- [124] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, “Large-scale image classification: Fast feature extraction and SVM training,” in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1689–1696.
- [125] X. Li, C. G. M. Snoek, M. Worring, and A. W. M. Smeulders, “Social negative bootstrapping for visual categorization,” in *Proc. Int. Conf. Multimedia Retrieval (ICMR)*, 2011, pp. 12:1–12:8.

- [126] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, Apr. 2006.
- [127] X. Li, C. G. M. Snoek, and M. Worring, "Annotating images by harnessing world-wide user-tagged photos," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 3717–3720.
- [128] H. Bannour and C. Hudelot, "Hierarchical image annotation using semantic hierarchies," in *Proc. Int. Conf. Information and knowledge management (CIKM)*, 2012, pp. 2431–2434.
- [129] H. Xu, X. Zhou, M. Wang, Y. Xiang, and B. Shi, "Exploring flickr's related tags for semantic annotation of web images," in *Proc. ACM Int. Conf. Image and Video Retrieval (CIVR)*, 2009, pp. 46:1–46:8.
- [130] R. Li, S. Bao, Y. Yu, B. Fei, and Z. Su, "Towards effective browsing of large scale social annotations," in *Proc. Int. Conf. World Wide Web (WWW)*, 2007, pp. 943–952.
- [131] M. Katsurai and M. Haseyama, "Exploring and visualizing tag relationships in photo sharing websites based on distributional representations," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 3617–3621.
- [132] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, "A word at a time: computing word relatedness using temporal semantic analysis," in *Proc. Int. Conf. World Wide Web (WWW)*, 2011, pp. 337–346.
- [133] J. Yao, B. Cui, Y. Huang, and Y. Zhou, "Bursty event detection from collaborative tags," *World Wide Web*, vol. 15, pp. 171–195, 2012.
- [134] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.
- [135] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," in *Proc. Int. Conf. Pattern Recognition (ICPR)*, vol. 2, 2004, pp. 28–31 Vol.2.
- [136] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 1975–1981.
- [137] X. Xie and H. Shi, "Dynamic multimode process modeling and monitoring using adaptive gaussian mixture models," *Industrial & Engineering Chemistry Research*, vol. 51, no. 15, pp. 5497–5505, 2012.
- [138] S. Chien and N. Immorlica, "Semantic similarity between search engine queries using temporal correlation," in *Proc. Int. Conf. World Wide Web (WWW)*, 2005, pp. 2–11.
- [139] M. Shokouhi, "Detecting seasonal queries by time-series analysis," in *Proc. Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR)*, 2011, pp. 1171–1172.

- [140] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proc. KDD Workshop*, 1994, pp. 359–370.
- [141] T. Rattenbury, N. Good, and M. Naaman, "Towards automatic extraction of event and place semantics from flickr tags," in *Proc. Int. Conf. Research and Development in Information Retrieval (SIGIR)*, 2007, pp. 103–110.
- [142] G. Kim, E. P. Xing, and A. Torralba, "Modeling and analysis of dynamic behaviors of web image collections," in *Proc. European Conf. Computer Vision (ECCV)*, 2010, pp. 85–98.
- [143] D. Borth, A. Ulges, and T. M. Breuel, "Dynamic vocabularies for web-based concept detection by trend discovery," in *Proc. Int. Conf. Multimedia*, 2012, pp. 977–980.
- [144] G. Kim and E. P. Xing, "Time-sensitive web image ranking and retrieval via dynamic multi-task regression," in *Proc. Int. Conf. Web Search and Data Mining (WSDM)*, 2013, pp. 163–172.
- [145] T. Uricchio, L. Ballan, M. Bertini, and A. Bimbo, "Evaluating temporal information for social image annotation and retrieval," in *Image Analysis and Processing – ICIAP 2013*, ser. Lecture Notes in Computer Science, A. Petrosino, Ed. Springer Berlin Heidelberg, 2013, vol. 8156, pp. 722–732.
- [146] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.
- [147] P. M. Engel and M. R. Heinen, "Incremental learning of multivariate gaussian mixture models," in *Advances in Artificial Intelligence – SBIA 2010*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011, vol. 6404, pp. 82–91.
- [148] J. Goldberger, S. Gordon, and H. Greenspan, "An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures," in *Proc. Int. Conf. Computer Vision (ICCV)*, 2003, pp. 487–493 vol.1.

List of Publications

Referred Journal

1. **Marie Katsurai**, Takahiro Ogawa, and Miki Haseyama, "A Cross-Modal Approach For Extracting Semantic Relationships Between Concepts Using Tagged Images," IEEE Transactions on Multimedia, vol. 16, no. 4, pp. 1059-1074, 2014.
2. **Marie Katsurai**, Takahiro Ogawa, and Miki Haseyama, "A Novel Framework For Extracting Visual Feature-Based Keyword Relationships From An Image Database," IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, vol. E95-A, no. 5, pp. 927-937, May, 2012.

Referred International Conferences

1. **Marie Katsurai** and Miki Haseyama, "Exploring and Visualizing Tag Relationships in Photo Sharing Websites Based on Distributional Representations," Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing (**ICASSP 2013**), pp. 3617-3621, May, 2013.
2. **Marie Katsurai** and Miki Haseyama, "Finding Tag Clusters for Image Retrieval in Social Media Websites," Proceedings of the International Workshop on Advanced Image Technology (**IWAIT 2013**), January, 2013.
3. **Marie Katsurai** and Miki Haseyama, "Introducing Multiple Descriptors to Estimate a Concept's Semantic Level," Proceedings of the 27th International Technical Conference on Circuits/Systems, Computers and Communications (**ITC-CSCC 2012**), F-T3-03, July, 2012.
4. **Marie Katsurai**, Takahiro Ogawa, and Miki Haseyama, "A Cross-Modal Approach For Extracting Semantic Relationships Of Concepts From An Image Database," Pro-

ceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (**ICASSP 2012**), pp. 2373-2376, March, 2012

5. **Marie Katsurai**, Takahiro Ogawa, and Miki Haseyama, "Keyword Recommendation for Image Annotation Using L1-Regularized Logistic Regression-Based Relationship Extraction," Proceedings of the Workshop on Picture Coding and Image Processing 2010 (**WPCIP 2010**), pp. 79-80, December, 2010.
6. **Marie Katsurai**, Takahiro Ogawa, and Miki Haseyama, "Effective Pedestrian Detection Using SVDD-Based Criterion For Region Integration," Proceedings of the 10th International Symposium on Communications and Information Technologies (**ISCIT 2010**), pp. 991-996, October, 2010.
7. **Marie Katsurai**, Takahiro Ogawa, and Miki Haseyama, "Visual Feature Based Semantic Keyword Recommendation For Image Annotation," Proceedings of the 25th International Technical Conference on Circuits/Systems, Computers and Communications (**ITC-CSCC 2010**), pp. 235-237, July, 2010

Domestic Conferences

1. 桂井麻里衣, 長谷山美紀, "画像集合の時間情報を考慮したタグ間の関係抽出に関する検討," 第28回信号処理シンポジウム, vol. C3-3, pp. 577-582, 下関, 2013年11月.
2. 桂井麻里衣, "ソーシャルメディアからの知識抽出と可視化," 第28回信号処理シンポジウム「若手研究者のための信号処理ワークショップ」チュートリアル講演, 下関, 2013年11月.
3. 吉崎茜, 桂井麻里衣, 小川貴弘, 長谷山美紀, "胃X線二重造影像を用いたHelicobacter pylori感染の識別に関する検討," 映像情報メディア学会技術報告, vol. 37, no. 8, pp. 25-28, 札幌, 2013年2月.
4. 桂井麻里衣, 長谷山美紀, "道路映像中のシカ検出のための画像特徴量に関する検討," 映像情報メディア学会技術報告, vol. 37, no. 8, pp. 125-128, 札幌, 2013年2月.
5. 桂井麻里衣, 小川貴弘, 長谷山美紀, "移動方向を考慮した道路映像中のシカ検出の高精度化に関する検討," 映像情報メディア学会技術報告, vol. 36, no. 9, pp. 67-71, 札幌, 2012年2月.
6. 桂井麻里衣, 小川貴弘, 長谷山美紀, "画像データベースにおけるキーワードの階層抽出に関する検討," 第26回信号処理シンポジウム, pp. 228-233, 札幌, 2011年11月.

7. 桂井麻里衣, 小川貴弘, 長谷山美紀, “確率的トピックモデルを用いた視覚的多様性に基づくキーワードの階層化手法,” 第13回DSPS教育者会議, pp. 17-18, 東京, 2011年9月.
8. 桂井麻里衣, 小川貴弘, 長谷山美紀, “画像特徴を用いたキーワードの階層化手法とその画像アノテーションへの応用,” 電子情報通信学会技術報告, vol. 110, no. 420, pp. 69-72, 札幌, 2011年2月.
9. 桂井麻里衣, 小川貴弘, 長谷山美紀, “新たな画像特徴の導入によるキーワードの関係抽出の高精度化,” 電気・情報関係学会北海道支部連合大会講演論文集, pp. 142, 札幌, 2010年10月.
10. 桂井麻里衣, 小川貴弘, 長谷山美紀, “画像特徴に基づくキーワードの関係抽出とその画像アノテーションへの応用,” 電子情報通信学会技術報告, vol. 110, no. 86, pp. 105-108, 北見, 2010年6月.
11. 桂井麻里衣, 小川貴弘, 長谷山美紀, “画像特徴量の適応的な重み付けを導入したロジスティック回帰モデルによる画像内容に基づく類似画像検索手法の検討,” 電気・情報関係学会北海道支部連合大会講演論文集, p. 156, 北見, 2009年10月.

受賞

1. 平成23年度電子情報通信学会北海道支部学生員奨励賞 受賞 (2012年3月).
2. 平成23年度 SIP学生奨励賞 受賞 (2011年11月).
3. 平成22年度電気・情報関係学会北海道支部 優秀論文発表賞 受賞 (2010年10月).

