

การวิเคราะห์ตัวแปรแฝงสำหรับการบรรยายและค้นคืนภาพ



นาย ณัฐชัย วัชรากินชัย

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต

สาขาวิชาวิศวกรรมไฟฟ้า ภาควิชาวิศวกรรมไฟฟ้า

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2553

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

LATENT VARIABLE ANALYSIS FOR IMAGE ANNOTATION AND RETRIEVAL



Mr. Nattachai Watcharapinchai

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Electrical Engineering

Department of Electrical Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2010

Copyright of Chulalongkorn University

ณัฐชัย วัชรานินชัย : การวิเคราะห์ตัวแปรแฝงสำหรับการบรรยายและค้นคืนภาพ.
(LATENT VARIABLE ANALYSIS FOR IMAGE ANNOTATION AND
RETRIEVAL) อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ศศ. ดร. สุภาวดี อ่วมวิทย์, อ. ที่ปรึกษา
วิทยานิพนธ์ร่วม : ดร. ศุภกร สิทธิไชย, 80หน้า.

ในปัจจุบันนี้ ภาพถ่ายดิจิทัลถือว่าเป็นสื่อประเภทหนึ่งที่มีความนิยมที่ใช้ในการบันทึกเหตุการณ์สำคัญที่เกิดขึ้นในชีวิตประจำวัน ด้วยจำนวนภาพถ่ายดิจิทัลที่เพิ่มมากขึ้นจึงต้องการฐานข้อมูลภาพที่รองรับทั้งการจัดเก็บและค้นหาภาพ เพื่อที่จะสามารถค้นคืนภาพได้อย่างถูกต้องตามการสอบถามของผู้ใช้จึงทำให้นักวิจัยมุ่งประเด็นในการปรับปรุงความแม่นยำของการค้นคืนภาพ และ เวลาในประมวลผลสำหรับการค้นคืนภาพ โดยเฉพาะอย่างยิ่งในระบบการค้นคืนภาพทั่วไปจะใช้ลักษณะเฉพาะระดับล่างของภาพ ซึ่งการค้นหาด้วยคำและตัวอย่างนั้นประสิทธิภาพของการค้นคืนไม่เป็นที่น่าพอใจ จุดประสงค์ของวิทยานิพนธ์เรื่องนี้ คือ การพัฒนาแบบจำลองการบรรยายภาพที่ให้ความแม่นยำสูงสำหรับการระบุและค้นคืนภาพ โดยใช้เวลาในการค้นหาที่ยอมรับได้ อีกทั้งรองรับการค้นหาหลายรูป เช่น คำ หรือ ภาพ โดยที่แบบจำลองที่นำเสนอถูกเรียกว่าแบบจำลองการวิเคราะห์ความน่าจะเป็นเชิงความหมายแบบสองตัวแปรแฝง (Two-Probabilistic Latent Semantic Analysis)

แบบจำลองที่นำเสนอใช้สองตัวแปรแฝงของแบบจำลองความน่าจะเป็นเชิงความหมายแฝง ซึ่งตัวแปรแฝงตัวแรกถูกใช้ในการจัดกลุ่มของคำในฐานข้อมูลที่มีมักจะเกิดขึ้นพร้อมกัน และ ตัวแปรแฝงที่สองถูกใช้ในการจัดกลุ่มคำภาพที่มีมักจะเกิดขึ้นด้วยกันในแต่ละคำ ด้วยเทคนิคของกระเป๋าคุณลักษณะ (Bag-of-Feature) ที่นำไปใช้กับการบรรยายภาพ ซึ่งภาพในฐานข้อมูลถูกแทนด้วยการนับจำนวนคำภาพของพจนานุกรมภาพ จากนั้นกระเป๋าคุณลักษณะของภาพร่วมกับความหมายของภาพนั้นถูกไปใช้สร้างแบบจำลองการบรรยายภาพ ได้แก่ naïve Bayes CMRM และ pLSA เปรียบเทียบกับแบบจำลองที่นำเสนอ โดยการบรรยายความหมายอย่างอัตโนมัติ ภาพที่ปราศจากการระบุความหมายจะถูกบรรยายด้วยคำจากแบบจำลองที่สร้างขึ้นแล้วคำเหล่านี้ในแต่ละภาพถูกใช้เป็นดัชนีคำ สำหรับงานการค้นคืนภาพ นอกจากนี้ประสิทธิภาพของระบบถูกประเมินด้วยความแม่นยำและความเร็ว เพื่อที่จะหาแบบจำลองที่ดีที่สุดที่รองรับการบรรยายและการค้นคืนด้วยคำ ค่าเฉลี่ยของค่าเฉลี่ยความแม่นยำ (mean Average Precision) ที่เปรียบเทียบระหว่างแบบจำลองของการบรรยายภาพ 4 แบบจำลอง จากผลที่ได้แสดงให้เห็นว่าแบบจำลองที่นำเสนอที่สามารถระบุความหมายของภาพให้ประสิทธิภาพที่เพียงพอทั้งเชิงความแม่นยำและความเร็วของการบรรยายและการค้นคืน ด้วยพารามิเตอร์ควบคุมที่เหมาะสม นอกจากนี้ แบบจำลองที่นำเสนอยังสามารถรองรับการทำงานด้วยการค้นหาภาพด้วยคำ และ ตัวอย่างของภาพ ที่ประกอบไปด้วยภาพที่ปราศจากการระบุความหมายในฐานข้อมูลที่ได้จากช่างภาพ โดยทั่วไปด้วยปราศจากเทคนิคการแยกส่วนภาพและการระบุความหมายด้วยมนุษย์

ภาควิชา...วิศวกรรมไฟฟ้า.....	ลายมือชื่อนิสิต..... <i>ณัฐชัย วัชรานินชัย</i>
สาขาวิชา..วิศวกรรมไฟฟ้า.....	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก..... <i>ศศ.ดร. สุภาวดี อ่วมวิทย์</i>
ปีการศึกษา.....2553.....	ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม..... <i>ดร. ศุภกร สิทธิไชย</i>

4971853621 : MAJOR ELECTRICAL ENGINEERING

KEYWORDS : IMAGE ANNOTATION / IMAGE RETRIEVAL / PROBABILISTICA
LATENT VARIABLE MODEL / BAG-OF-FEATURE

NATTACHAI WATCHARAPINCHAI : LATENT VARIABLE ANALYSIS FOR
IMAGE ANNOTATION AND RETRIEVAL . ADVISOR: ASST. PROF.

SUPAVADEE ARAMVITH, Ph.D., CO-ADVISOR : SUPAKORN SIDDHICHAI,
Ph.D., 80 pp.

Nowadays, digital photography is very popular type of media used to record important events in everyday's life. The increasing number of digital images requires a good image database to support image collection and image search. In order to correctly retrieve image according to the user's query, many researches focus to improve the precision of the retrieval images and the retrieval processing time especially for general image retrieval system using low-level image features where query by words and examples do not have satisfactory retrieval performance. The purpose of this dissertation is thus to develop the image annotation model aimed for higher precision in identifying and retrieving images with acceptable search time and supported query by words or images. The proposed model is called "Two-Probabilistic Latent Semantic Analysis".

The proposed model uses two latent variables of probabilistic latent semantic model, of which the first latent variable is used to group the words in an image database that often occurs, and of which the second latent variable is used to group the visual words usually appear in each word. Based on Bag-of-Feature (BoF) technique applied to image annotation, images in the database are represented by counting the number of visual word of the constructed visual vocabulary. Afterward, the BoF of images corresponding to their meaning is used to construct the annotation models namely naïve Bayes, CMRM, and pLSA, comparing with our proposed model, "Two-pLSA". Using the automatic image annotation, an unlabeled image is annotated by the words from the constructed models, and then the annotated words of that image are used for a text index for image retrieval task. Moreover, the performance has been evaluated by the precision and speed to find the best model for supporting the annotation and retrieval tasks. The performance values of both tasks are measured by mean Average Precision (mAP) to compare among 4 annotation models. The results showed that our proposed model used to identify meaningful images offers satisfactory performance both in terms of accuracy and speed of annotation and retrieval with appropriate control parameters. In addition, the proposed model also supports query by words and image examples including unlabeled images in the database taken by the photographers without object segmentation and specific meaning.

Department : Electrical Engineering.....

Student's Signature

Nattachai Watcharapinchai

Field of Study : Electrical Engineering.....

Advisor's Signature

Sud Am

Academic Year : 2010.....

Co-advisor's Signature

Supadee Aramvith

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor, Asst. Prof. Dr. Supavadee Aramvith. She has encouraged, guided, and supported me every means throughout my study. And I also would like to express my appreciation to my co-advisor, Dr. Suppakorn Siddhichai. He has provided me knowledge in the research procedure and given me some useful suggestions and provides information about the scholarship.

Secondly, I would like to show my thankfulness to all of my dissertation committees including of: the chairman, Assoc. Prof. Dr. Chedsada Chinrungrueng, and the committees, Asst. Prof. Dr. Charnchai Pluempitiwiriyawej, Asst. Prof. Dr. Suree Prumrin, and Dr. Sanparith Marukatat, which have given me some valuable comments.

Thirdly, I would like to acknowledge Thailand Graduate Institute of Science and Technology (TGIST) which has provided some grants for my research.

Finally, I would like to express my deepest gratefulness to my family who have entirely supported encouraged and believed in me without any doubts, and also to thank all of all colleagues and friends at the Center of Excellent in Telecommunication Engineering. They assist and support me in many ways during my study.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

Contents

	Page
Abstract in Thai.....	iv
Abstract in English.....	v
Acknowledgements	vi
Contents.....	vii
List of Tables.....	x
List of Figures	xi
Chapter I Introduction.....	1
1.1 Background and Motivation	1
1.2 Research Objective.....	3
1.3 Scope.....	3
1.4 Expected Prospects.....	4
1.5 Research Procedure	4
Chapter II Literature Review and Related works.....	6
2.1 Image Representation.....	6
2.1.1. Bag-of-Features from points of interest	6
2.2 Literature Reviews and Related Works	11
2.2.1. Generative Model	12
2.2.2. Discriminative Model	15
2.3 Existing Annotation models.....	17

	Page
2.3.1. Naive Bayes Model.....	17
2.3.2. Cross Media Relevance Model (CMRM)	18
Chapter III Probabilistic Graphical Model.....	20
3.1 Introduction to Probabilistic Graphical Model	20
3.1.1. Maximum Likelihood Learning	22
3.2 Probabilistic Latent Semantic Analysis	24
3.2.1. pLSA Learning using EM-Algorithm.....	26
3.2.2. Inference: pLSA of a new document	28
3.2.3. Convergence Condition.....	28
3.2.4. Image Annotation with pLSA Model.....	29
Chapter IV The Proposed Technique.....	31
4.1 Learning parameters using EM Algorithm.....	34
4.2 Annotating an image with two-pLSA	38
4.3 Image Retrieval by Text Query.....	38
4.4 Image Retrieval by Image Query	39
Chapter V Experimental Results	40
5.1 Experiment 1: Dimensionality Reduction of SIFT using PCA for object Categorization.....	41
5.1.1. Dataset and simulation	41
5.1.2. Performance of PCA-SIFT on dimension reduction and average precision	42
5.1.3. Performance of PCA-SIFT on the number of visual word.....	46
5.2 Experiment 2: Selecting Visual word with criterion.....	48

	Page
5.2.1. Dataset and Simulation Condition	48
5.2.2. Results using scale parameter by threshold	48
5.2.3. Maximum Probability and Entropy Criterion for choosing visual words	51
5.2.4. Discriminative Criterion for choosing visual words	53
5.3 Experiment 3: Performance in PASCAL 2008 Dataset	55
5.3.1. Dataset and Simulation Condition	55
5.3.2. Image Ranking with latent variable Z	56
5.3.3. Image Annotation: mean Average Precision Performance and Processing Time	58
5.3.4. Text-based Image Retrieval in unlabeled images	60
5.3.5. Unlabeled Image Retrieval based on automatic image annotation	63
5.4 Experiment 4: Performance in MIRFLICKR25000 dataset.....	66
5.4.1. Dataset and Simulation Condition	67
5.4.2. Image Annotation: mean Average Precision Performance and Processing time	67
5.4.3. Unlabeled Image Retrieval based on automatic image annotation	70
5.5 Experiment 5: Performance in MIRFLICKR25000 dataset when the constructed visual vocabulary by PASCAL 2008 dataset.....	74
Chapter VI Conclusions	78
References.....	81
Vitae.....	88

List of Tables

	Page
Table 5.1 The number dimension Of PCA	42
Table 5.2 Average Precision	43
Table 5.3 Number of Visual Words and Mean AP	47
Table 5.4 Statistics of PASCAL 2008 Image dataset.....	55
Table 5.5 Performance of image retrieval with a word query where the size of visual word as 1000 visual words	62
Table 5.6 The performance in the crossed models where a model as indexing model and another as querying model.....	65
Table 5.7 The number of image in each word on MIRFlickr25000 dataset	66
Table 5.8 mean Average Precision of image annotation with existing model on MIRFLICKR25000 dataset.....	67
Table 5.9 mean Average Precision of image annotation when the number of vocabulary as 1000 visual words	68
Table 5.10 mean Average Precision of image annotation when the number of vocabulary as 5000 visual words	68
Table 5.11 processing time (seconds) of annotation with existing model of MIRFlickr25000 dataset.....	69
Table 5.12 mean Average Precision of image retrieval using a query image with existing model on MIRFlickr25000 dataset.....	70
Table 5.13 mean Average Precision of image retrieval using a query image with two- pLSA where the size of visual vocabulary as 1000 visual words.....	71
Table 5.14 mean Average Precision of image retrieval using a query image with two- pLSA where the size of visual vocabulary as 5000 visual words.....	71
Table 5.15 The performance in the crossed models of 1000 visual words where a model as indexing model and another as querying model in MIRFlirkc25000 dataset.....	73

Table 5.16 The performance in the crossed models of 5000 visual words where a model as indexing model and another as querying model in MIRFlirkc25000 dataset.....	73
Table 5.17 Performance of image annotation compared between two visual vocabularies	75
Table 5.18 Performance of image retrieval by a text query compared between two visual vocabularies	75
Table 5.19 Performance of image retrieval by an image query compared between two visual vocabularies	76



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

List of Figures

	Page
Figure 2.1 For each octave of scale space	7
Figure 2.2 maxima and minima of the difference-of-Gaussian images	8
Figure 2.3 A local feature	9
Figure 3.1 Example of directed graphical model	21
Figure 3.2 Graphical Model of pLSA	26
Figure 4.1 our proposed model, two latent aspects pLSA	32
Figure 4.2 Diagram of Image retrieval by Text Query.....	38
Figure 4.3 Diagram of Image retrieval by Image Query	39
Figure 5.1 Precision-recall curve on baseline technique.....	44
Figure 5.2 precision-recall curve of PCA-SIFT for $\epsilon = 0.5$	44
Figure 5.3 Precision-recall curve PCA-SIFT for $\epsilon = 0.25$	45
Figure 5.4 Precision-recall curve PCA-SIFT for $\epsilon = 0.1$	45
Figure 5.5 The results of eliminated on an airplane image	49
Figure 5.6 The results of elimination on an airplane image with noises.....	50
Figure 5.7 The results of elimination on a motorbike and human image	51
Figure 5.8 The maximum probability and entropy criterion result	53
Figure 5.9 The discriminative criterion result	54
Figure 5.10 The 11 most probable images in 7 latent variables of grouping in 20 latent variables from PASCAL 2008 Dataset variable.....	57
Figure 5.11 mean Average Precision of image annotation in PASCAL 2008 dataset.....	58
Figure 5.12 An example image of improved annotation performance.....	59
Figure 5.13 Processing time of image annotation per an image of the PASCAL dataset.....	60
Figure 5.14 mAP of image retrieval by text query in PASCAL 2008 dataset	61

Figure 5.15 mAP Performance of image retrieval using image query for unlabeled image in PASCAL dataset	63
Figure 5.16 Example of image retrieval in the unlabeled image where a ground truth word from PASCAL dataset.....	64
Figure 5.17 Example of image retrieval in the unlabeled image where two ground truth word from PASCAL dataset.....	64
Figure 5.18 Example of image retrieval in the unlabeled image where three ground truth word from PASCAL dataset.....	64
Figure 5.19 The processing time of two-pLSA model when increasing the number of latent variable Z and l for image annotation process.....	70
Figure 5.20 The processing time when increasing the size of visual vocabulary compared between the estimated texts search and BoF from a query image.....	72

CHAPTER I

INTRODUCTION

1.1 Background and Motivation

Digital multimedia recording and storage devices become common thus causes that the number of digital multimedia, such as digital image and digital video, is also increase considerably. Therefore, to efficiently access the image/video collection requires a system to handle search and organization of this information. Such system is called image/video retrieval system that organize and index database of a photos. The ideal retrieval system should be designed to support intuitively search for the user, and require minimal amount of human interaction and to be applicable to large collections.

The prevalent approach to image retrieval falls into two main categories, namely text-based and content-based image retrieval. In text-based image retrieval system, images are first annotated with text, and the traditional text retrieval techniques can be used to perform image retrieval. The main advantages of text-based retrieval are its simplicity and convenient direct adoption of mature textual information retrieval techniques. In addition, it is easy to use text to express textual characteristic related to images. Some commercial image retrieval systems, such as Google and Yahoo!, have adopted this technique. In general, there are two strategies to associate image with text. One strategy is to annotate image by human. These annotations provided by people usually are close to the semantic of images. However, this strategy suffers two drawbacks. First, it is very tedious and time-consuming to annotate image manually, especially when the size of image collection is huge. Although manual annotation involves a substantial amount of work, and often results in heavy cost, there is a system that utilizes a collaborative system approach called LabelMe [1] which takes advantage

of its member as annotators. Second, these annotations are usually subjective because different people may give different descriptions to the same image. Another strategy automatically annotates image with terms or words. The main advantage of this strategy is the complete automatic process of image annotation without human interference. However, the use of text involves some problems. For example, it is possible that all texts are irrelevant to image contents, and some images are presented or annotated without text information.

Because of the emergence of large scale image collection, the difficulties faced by the manual annotation approach became more and more accurate. To overcome these difficulties, content-based image retrieval (CBIR) [29-33] is proposed. CBIR automatically indexes images by using the extracted low-level visual features such as color and texture, and then the retrieval of images is based solely upon these indexed image features. Since then, many techniques in this research direction have been developed. Although CBIR is a promising methodology for image retrieval, it still suffers from some issues. The computational cost of extracting image features for a large collection might be unacceptable. In addition, the user query must be provided in the form of an example of the desired image, which is not simple to common users. Moreover, the images with similar low-level features may not represent the similar meanings. Finally, the reliance on visual similarity for judging semantic similarity in all current approaches is not reliable due to the so-called semantic gap between low-level features and high-level concepts or semantic meanings.

Recently, automatic image annotation techniques are proposed to address the semantic gap problem. Automatic image annotation is the process by which a computer system automatically assigns keywords to a digital image. The primary purpose of a practical content-based image retrieval system is to discover images relating to a given concept in the absence of reliable metadata. In contrary to CBIR, annotations can facilitate image search through the use of semantic meaning such as text. This methodology assures the good performance of image retrieval that if the results of

mapping between images and words are reasonable, text-based image retrieval system can be semantically more meaningful than search with CBIR. However, this technique is still in its infancy and is not sophisticated enough to extract satisfactory semantic concepts. Moreover, many experiments show that current image annotation techniques still have poor performance in the context of image retrieval because of that irrelevant keywords associated with images often lessen image retrieval performance.

1.2 Research Objective

A novel latent variable modeling technique for image annotation and retrieval is proposed. This model is useful in annotating the images with relevant semantic meanings as well as in retrieving images which satisfy the user's query with specific text or image. The framework of two-step latent variable is proposed to support multi-functionality of the retrieval and annotation system.

Furthermore, the existing and the proposed image annotation models are compared in terms of their annotating performance. Images from standard databases are used in the comparison in order to identify the best model for automatic image annotation, using precision-recall measurement. Local features, or visual words, of each image in the database are extracted using SIFT and clustering techniques. Each image is then represented by Bag-of-Features (BoF) which is a histogram of visual words. Semantic meanings can then be related to each BoF using latent variable for annotation purposes. Subsequently, for image retrieval, each image query is also related to semantic meanings. Finally, image retrieval is achieved by matching semantic meanings of the query with those of the images in the database using a second latent variable.

1.3 Scope

1. Study the performance and the limitation of traditional image annotation and retrieval models namely co-occurrence, cross-media relevance, and probabilistic latent semantic analysis.

2. Develop a novel two-step latent variable model that can be applied to image annotation and image retrieval systems on the standard image databases.
3. Study the performance of the proposed model and provide a comparative study of proposed techniques with other traditional models proposed by other researchers.

1.4 Expected Prospects

1. Acquire a basic knowledge of image database modeling techniques for applying to image/video retrieval system.
2. Obtain the new model that is used for a new baseline in the novel image retrieval system.
3. Publish in an international journal or conference papers.
4. Obtain the empirical knowledge of advantages and disadvantages in using the proposed model techniques in image retrieval.
5. Understand the necessity of the modeling techniques for image annotation and retrieval system.

1.5 Research Procedure

1. Study previous research paper relevant to the research works of the dissertation.
2. Develop the new image annotation and retrieval model.
3. Develop the high performance of image annotation technique and high ranking of image retrieval technique when using our model.
4. Develop simulation programs.
5. Test the proposal algorithms using standard generic image databases such as PASCAL 2008, MIRFLICKR25000.
6. Perform the proposed modeling technique on image retrieval system including image annotation and image ranking.
7. Collect and analyze computational results obtained from simulation programs.

8. Summarize the major finding as we found in Step 7 and conclude the performance of the proposed model in all concerned aspects.
9. Publish in an international journal or conference papers.
10. Check whether the conclusion meet all the objectives of the research work of the dissertation.
11. Write up.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER II

LITERATURE REVIEW AND RELATED WORKS

The details of the visual vocabulary are described on this chapter comprising two sections such as image representation technique in Section 2.1, the literature reviews in Section 2.2 and the existing annotation model in Section 2.3. For image representation, the Bag-of-Feature of Scale Invariant Feature Transform and the constructed visual vocabulary are described in Section 2.1. Afterward the review of image annotation and image retrieval based annotation model is described in Section 2.2. Finally, the existing annotation models using in this dissertation are described in Section 2.3 namely naïve Bayes model, cross media relevance model and probabilistic latent analysis model.

2.1 Image Representation

The image representation comprises of two importance procedures including of extracting local features from points of interest and constructing a visual vocabulary from the local features.

2.1.1. *Bag-of-Features from points of interest*

The Bag-of-Feature (BoF) construction consists of the following steps: (i) automatic detection of regions or points of interest, (ii) extraction of local features from regions of detected points, (iii) these descriptors are then quantized and formed a visual vocabulary, (iv) BoF uses the number of occurrences of each visual word in an image to form a visual word histogram that represents a given image.

2.1.1.1 Points of interest detection

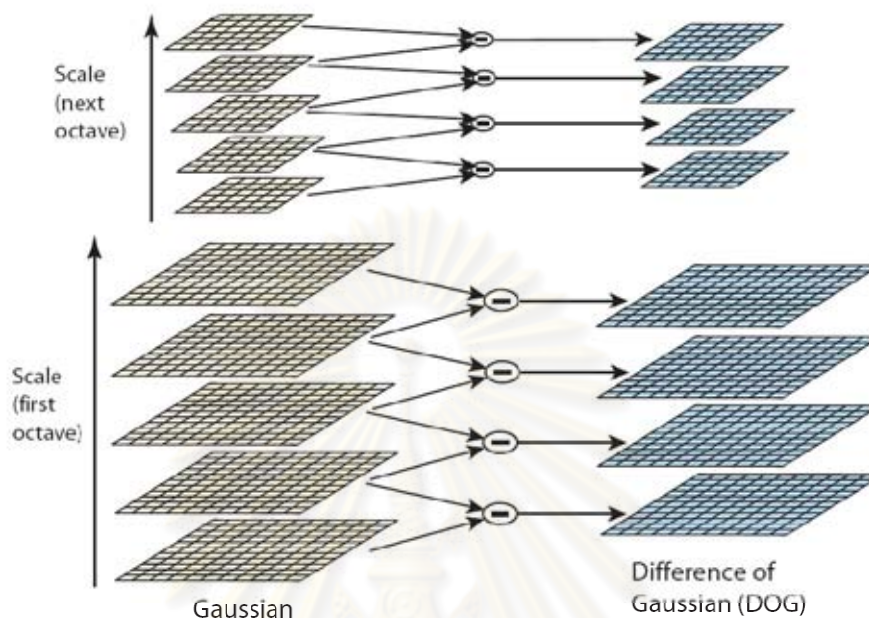


Figure 2.1 For each octave of scale space, the given image is repeatedly convolved with to build a set of blurring images shown on the left. Adjacent Gaussian Images are subtracted to construct the difference-of-Gaussian images on the right. After each octave, the Gaussian image is down-sampled and the process repeated [17].

The points of interest detection step aims to automatically find points which are invariant to some geometric and photometric transformations. In order to ensure that given an image and its transformed version, the same points will be detected from the image when that image is transformed by rotation and translation matrix. Several points of interest detectors exist in [16-17]. In this work, the difference of Gaussian filter is used at various scales. Following are the major stages of computation used to detect points of interests.

- Convoluting the image with Gaussian filter at various scales.
- Constructing the different of Gaussian images from adjacent blurred images as shown in Figure 2.1. An efficient approach to construction of different of Gaussian image has been referred in [16-17]. The initial image is incrementally

convolved with Gaussian to produce image separated by a constant factor k in scale space, shown stacked in the left column. We choose to divide each octave of scale space in an integer number s of intervals, so the intervals of each are produced with $s + 3$ images in the stack of blurred image for each octave, which means the scale intervals of scale parameter equals to 3 scale images per an octave. And the variance parameter of Gaussian in the first image equals to 1.6. and the number of octave is used in this dissertation as 4 octaves.

- Detecting the candidate keypoints by Scale-space extrema detection, shown in Figure 2.2.
- Removing keypoints with low contrast and responding along edges
- Assigning keypoints with their orientation and their scale.

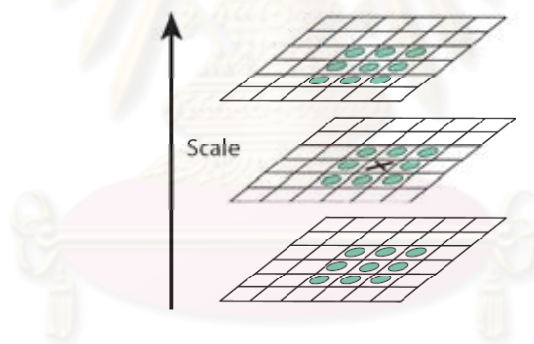


Figure 2.2 maxima and minima of the difference-of-Gaussian images are detected by comparing a pixel, marked with X, to its neighbors in 3x3 regions at the current and adjacent scale, marked with circles [17]. These maxima or minima pixels are candidate keypoints.

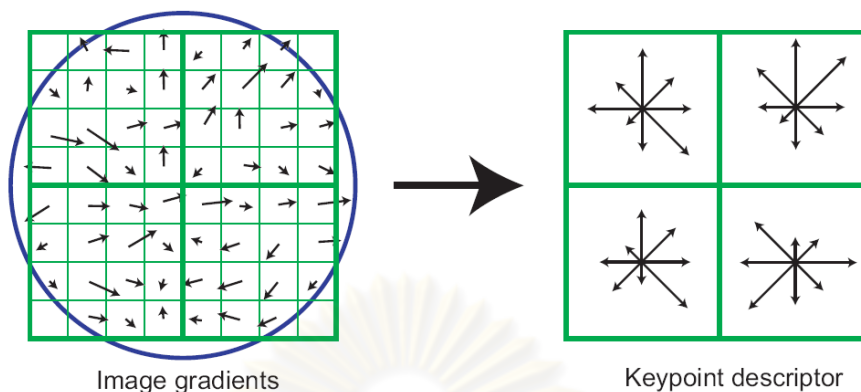


Figure 2.3 A local feature is computed by first determining the gradient magnitude and orientation of each neighbors point in a 8x8 pixels region around location of an points of interest as shown on the left, and then computed histogram of 8 orientation over 4x4 sub-regions as shown on the right. So in this case, the dimensionality of a local feature is equaled to 32 dimensions.

2.1.1.2 Local Features

Local features are extracted from the regions around each the points of interests assigned by points of interests detector. In this work we use the SIFT, Scale invariant Feature Transform, as local features [17], because in [16] it was shown to perform best in terms of specificity of region representation and robustness to image transformation. It can compute by local histogram of edge directions compared with given the orientation of the image region estimated from points of interest detector in Figure 2.3. In this work we use neighbors point in 16x 16 regions around a point of interest location. So, when each sub-region is computed 8 orientations over 4x4, the dimensionality of our local feature is equal to $4 \times 4 \times 8 = 128$ dimensions.

2.1.1.3 Quantization of Local Feature into Visual Vocabulary

From the two previous point detection and extracting local features step, we obtain a set of visual feature of images. To construct a visual vocabulary, we use K-mean clustering technique, an unsupervised learning which estimates mean vectors of each cluster. K-

mean. Using the clusters model in this case, a testing feature \mathbf{y} is assigned to the cluster i^* having the smallest distance, as shown in Eq. (2.1)

$$i^* = \underset{k}{\operatorname{argmin}} \|x_n - \mu_k\|^2 \quad (2.1)$$

Generally, in order to achieve a simple, fixed size of image representation, we can rewrite each cluster of K-mean from local feature according to the smallest distance in Eq. (2.1) by Eq. (2.2). This means that local features of an image are quantized into a set of cluster, called visual vocabulary, and each cluster is called visual word.

$$x \rightarrow Q(x) = v_k \Leftrightarrow \operatorname{dist}(x, v_k) \leq \operatorname{dist}(x, v_j) \quad \forall j \in \{1, \dots, N\} \quad (2.2)$$

,where N denotes the size of visual vocabulary or the number of cluster set.

Algorithm 2.1 EM algorithm for K-mean, where N is the number of the unclassified local features and n_k is the number of the classified local feature at each cluster k .

Initial the mean vectors μ_k , using random the local features
 $\mu_k \leftarrow \mu_{k_init} \leftarrow \text{random}$
Repeat
 UE-StepU: classify a local feature x_n into a local cluster, fixed the mean vectors μ_k

$$z_k \leftarrow \begin{cases} 1, & \text{if } k = \underset{j}{\operatorname{argmin}} \|x_n - \mu_j\|^2 \\ 0, & \text{otherwise} \end{cases}$$

 UM-StepU: update parameters of mean vectors, which estimates given

$$n_k \leftarrow \sum_{n=1}^N z_{nk}$$
 if $n_k > 0$

$$\mu_k \leftarrow \sum_{n=1}^N z_{nk} x_n$$
 else

$$\mu_k \leftarrow \mu_{k_init}$$

Until classify not change or more than maximum iteration
Return μ_k

Technically, the K-mean method aims to group of similar local features into a specific visual word. So when we could be considering as similar to the stemming

preprocessing step in textual retrieval system. Instance of stemming technique in textual task, the term *man*, *men*, *human* are mapped to same stem *man*, which becomes a word for bag-of-words representation in textual task. In same motivation, the bag-of-features is to map local feature into a visual word that is similarity by Eq. (2.1)

2.1.1.4 Bag-of-Features

The first image representation, we will consider the bag-of-features (BoF) that can be constructed from the local feature according to

$$s(d_i) = \{n(d_i, v_1), n(d_i, v_2), \dots, n(d_i, v_N)\}, \quad (2.3)$$

, where $n(d_i, v_j)$ denotes the number of occurrences of visual word v_j in an image d_i . This representation has not information of spatial relation between each visual word. Because, in textual retrieval task, the bag-of-feature places importance in a significant amount of information in documents and images than ordering of information. So, the spatial relation is completely removed from bag-of-feature representation.

2.2 Literature Reviews and Related Works

Several approaches have been proposed in the literature on automatic image annotation [2-15, 37-52]. Different models and machine learning techniques are developed to learn about the correlation between low-level features and textual words from the examples of annotated image and then apply such correlation to predict words for new images. In this section we review some pioneer works about automatic image annotation which divided into two major categories namely, generative model and discriminative model. The basic idea of generative model [1-11] is to construct a model from joint probability of image features and words, and then use Bayes' rule and marginalization of probability to estimate the conditional probability of words given by image features for automatically image annotation. But in the discriminative model, the model is directly being constructed the conditional probability of words given image features.

2.2.1. Generative Model

In the first category, generative mode, there are three major models which are considered for image annotation namely co-occurrence model [53], relevance model [39-45], and latent semantic analysis model [3-7]. The first generative model is the co-occurrence model [53], which collects the co-occurrence, is to count words and image features matrix, and use its matrix to predict annotated image words for images. In [37], Duygulu et al. proposed the improved co-occurrence model by utilizing machine translation models. In this method, it considers image annotation as a process of translation from visual feature to texts and collects the co-occurrence information by estimation of translation probability. They proposed to describe image using a vocabulary of blobs. Each image is generated by using a certain number of these blobs. Their Translation Model – a substantial improvement on the co-occurrence model – assumes the image annotation can be viewed as the task of translation from a vocabulary of blobs to a vocabulary of words.

2.2.1.1 Relevance Model

Second model, relevance model, some researcher used relevance language model which has been successfully applied to automatic image annotation. The essential idea is to first find annotated images that are similar to a text image and then use the words shared by the annotations of the similar images to annotate to the test image. There are two subcategories in relevance model, namely discrete variable, and continuous variable. In discrete variable, it is the basic idea of cross-media relevance model. This model aims to improve the co-occurrence model described in [39], Jeon et al. assume a one-to-one corresponded between blobs and words in images. Images are considered as a set of words and blobs, which are assumed independent given the image. The conditional probability of word given a training image is estimated by the count of word in this image smoothed by the average count of this word in training set. These posterior distributions allow the estimation of the probability of a potential caption (set of words) and unseen blobs as an expectation over all training images. Another

approach to improve co-occurrence model is Multiple Bernoulli relevance models [41]. Multiple-Bernoulli relevance model is based on Cross-media relevance model but it is different in the word distribution hypothesis. Cross-media relevance assumes that annotation words for any given image follow a multinomial distribution, while this model uses Bernoulli process to generate words.

For the continuous variable model, while the cross-media relevance model is to counting word in given training set that is discrete random variable techniques, but same authors considered blobs correspond to word which called Continuous-space relevance Model described in [40]. There are two significant differences between Cross-media relevance model and Continuous-space relevance model. First cross-media relevance model is a discrete model and cannot take advantage of continuous features. In order to use cross-media relevance model for image annotation we have to quantize continues feature vectors into a discrete vocabulary. Secondly, the difference of the cross-media relevance model relies on clustering of the feature into blobs. Annotation quality of the cross-media relevance model is very sensitive to clustering errors, and it depends on being able to priori select the correct cluster granularity: too many clusters will results in extreme sparse of the space, while too few will lead us to confuse different objects in the images. Continuous relevance model does not rely on clustering and consequently does not suffer from the granularity issues.

Currently P. Huang et al [42] proposed combining three co-occurrence models including translation model, cross-media relevance model and multiple Bernoulli relevance model. They showed the comparison performance between individual model and combining models. The combining model gives the better performance for image annotation. In addition, they compare among individual models that the result of Multi-Bernoulli relevance model gives better performance than others. In addition, in [45], Jin et al. proposed a coherent language model for automatic image annotation that takes into account the word-to-word correlation to estimate a coherent language model for an image. Because a common problem shared by most approaches for automatic image

annotation is that each annotated word for image is predict independently, the word-to-word correlation is important particularly when image features are insufficient in determining an appropriate word annotation.

Another approach in relevance model, in [43], Marukatat proposed a semi-supervised technique. This work estimated a conditional probability of words given by un-annotation image and estimated the parameter model by semi-supervised learning. Three qualities were computed, namely, posterior probability on the image in dataset, similarity between un-annotated images and annotated image, and the probability of annotated words to estimate the conditional probability of word given images.

2.2.1.2 Latent Semantic Analysis Model

Another way of capturing co-occurrence information is to introduce latent variables to associate visual feature with words. Standard latent semantic analysis (LSA) and probabilistic latent semantic analysis (pLSA), are applied to automatic image annotation [3-6]. A significant step forward in this approach was proposed by Hofmann [3], who presented the probability LSA model, also known as the aspect model, as an alternative to LSI, which has used for text retrieval research. The pLSA approach models each word in document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representation of topics. Thus each word is generated from single topic, and different words in a document may be generated from different topics. Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topic. This distribution sometime is called the reduce descriptor associated with document. By this approach, Money et al [4, 5] have extended the experiments to bigger collection of 8000 images and shown encouraging results to using pLSA model. And In [6], Bosch et al investigated whether dimensionality reduction using a latent model is beneficial for the task of weakly supervised scene classification which used k-nearest neighbor or Support Vector Machines (SVM) for their experiment. This technique is a combination of pLSA and supervised learning.

Some works were extended pLSA model to annotation images. Blei and Jordan [7] extended the aspect model as the latent Dirichlet Allocation (LDA) Model and proposed a correlation LDA (CORR-LDA) model. This model assumes that a Dirichlet distribution can be used to generate a mixture of latent factors. This mixture of latent factors is then used to generate words and regions. Fei-Fei and Perona [8] modified LDA model and have extended the experiments to learn natural scene categories. Their algorithm provides a principled approach to learn relevant intermediate representation of scenes automatically and without supervision to what humans would do. In another extension of the aspect model approach, Zhang et al [38] proposed the aspect model as Gaussian mixture distribution. Instead of assuming artificial distribution of annotation word and the unreliable association evidence used in many existing approaches, they assume a latent aspect as the connection between the visual feature and the annotation words to explicitly exploit the synergy among the modalities. In [44], Pham et al. study the effect of Latent Semantic Analysis on two different tasks namely image retrieval task and automatic annotation task. This result ensures that LSA model when combining to image retrieval system can improve automatic annotation image.

2.2.2. Discriminative Model

The second category, discriminative model, the method is closely resembling to generic object detection task. The problem of detecting is often posed as a binary classification task; namely, distinguish between an object, which means that a word describes its object, and background class. Such a classifier can turn into a detector by sliding it across the image and classifying each local window. Alternatively, in [12,13] one can extract local window at locations and scales returned by an interesting point detector and classifier these, either as a whole object or as a part of an object. In [14], Torralba et al proposed the training multiple binary classifiers at the same time needs less training data since many objects share similar image features. In this assumption, they show that classifier is fast since the computation of many features can be shared among different objects. In [46], Chang et al. learned an ensemble of binary, each

specific label. In [47], Li et al proposed a Confidence-based Dynamic Ensemble model, which is a two-stage classifier. In [48], Liu et al proposed SVM-based active feedback using clustering and unlabeled image to address the small size problem. In [49], Zhou et al. proposed biased discriminative analysis (BDA) and its kernel form to find the transformed space where the positive examples cluster while the negative ones scatters away. To handle the singularity problem caused by small sample size, Tao [50] designed direct kernel BDA scheme where direct discriminative analysis is used to replace the regularization method used in BDA. Recently active learning studies the strategy for the learner to actively select samples to query the user for labels, in order to achieve the maximum information gain in decision making. In [51] Cox et al. used entropy minimization in search of the set of images that, once labels, will minimize the expected number of future iteration. In [52], Tong and Chang proposed SVM-active algorithm for applications in text classification and image retrieval. In this category, research aim to determine a function that can classify, detect the interest object and then annotate and ranking with result which obtain from classification. However, in our idea, the retrieval system should be designed with multifunction to support many requirements from user. So the generative approach is interesting to design the system than discriminative model. However the discriminate model approach is designed for object recognition which is a specific object detected on images, thus it is not suitable for image retrieval task.

Because the previous models are designed for individual task, for instance, the co-occurrence model is useful for automatic annotation but it is not suitable for image retrieval task. While the latent semantic analysis is designed for image retrieval task but it is not suitable for automatic image annotation. Although, the discriminative model archives the best performance for image annotation task, but each semantic meaning word is modeled by stand alone model because we have to design each object to detect it. So the model is not suitable for image retrieval. Therefore, to design a novel image retrieval system should be the expert system based on generative model that is a joint probability among low-level image feature, semantic word feature vectors and

image document term. Our model can automatically annotate new images which will be added into that system and can also retrieve images by user's query with text or image. So, in this dissertation, we propose a new model, called the two latent aspects pLSA model, which can work in multifunction, including of image retrieval function and automatically image annotation for image retrieval system.

2.3 Existing Annotation models

The existing automatic image annotation techniques in this dissertation are compared namely naïve Bayes model, Cross Media Relevance Model (CMRM) and Probabilistic Latent Analysis Model (pLSA Model). The efficiency of these models are measured by mean Average Precision and processing time, so that the accuracy in each rank of which each model predict words are considered to find which model is suitable for image annotation and retrieval problem. In this section, we would like introduce the existing model, except pLSA model will be explained in next chapter.

2.3.1. Naïve Bayes Model

The naïve Bayes model [18] is a simple classifier used for often in text categorization. It can be viewed as the maximum a posterior probability classifier for the generative model in which: (1) a image category is selected according to class prior probability; (2) each word in the image is chosen independently from multinomial distribution over specific words.

Considering image annotation problem, a set of labeled image $I = \{I_i\}$ and the set of visual vocabulary $V = \{v_t\}$ at visual word t . The naïve Bayes model is constructed by counting the number visual word t belonged word w . To annotate a new image, the Bayes' rule is applied and ranked of posterior score as:

$$P(w_j|I_i) \approx P(w_j)P(I_i|w_j) = P(w_j) \prod_{t=1}^V P(v_t|w_j)^{N(t,i)} \quad (2.4)$$

Where $N(t, i)$ is the count of the visual word v_t in an unlabeled image I_i , and in Eq. (2.4), the naïve Bayes model requires the estimating of the class conditional probabilities of visual word v_t given word w_j . In order to avoid probabilities of zero, the parameters $P(v_t|w_j)$ are computed with Laplace smoothing:

$$P(v_t|w_j) = \frac{1 + \sum_{\{I_i \in w_j\}} N(t, i)}{|V| + \sum_{s=1}^{|V|} \sum_{\{I_i \in w_j\}} N(s, t)} \quad (2.5)$$

2.3.2. Cross Media Relevance Model (CMRM)

Cross Media Relevance Model (CMRM) [40] is improved from relevance-based learning model, which the problem of automatically annotating images and the ranked retrieved images are considered the joint probability distribution of visual word and words. Images are considered as sets of words and visual words, The CMRM is assumed independently given the images. In training process, the conditional probability of word and visual word given a training image dataset is estimated by the counting of visual word and words, which are smoothed by the average count of these words. In testing process, the posterior distribution allows the estimation of the probability of set of word and an unseen image as expectation over all training image.

To annotating an unseen image d_{new} is based on the joint probability of all its t constituting visual word v_t and word w_j . This joint probabilities are estimated by its expectation over the M training images,

$$P(w_j, v_1, \dots, v_t) = \sum_{i=1}^M P(d_i) P(w_j, v_1, \dots, v_t | d_i). \quad (2.6)$$

The visual words are considered independently given an image d_i , which gives:

$$P(w_j, v_1, \dots, v_t) = \sum_{i=1}^M P(d_i) P(w_j | d_i) \prod_{t=1}^{N_v} P(v_t | d_i)^{N(d_i, v_t)}. \quad (2.7)$$

where $N(d_i, v_t)$ is the count of the visual word v_t in image d_i . The probability of word w_j in a training image d_i is the likelihood of word in image combined with the likelihood of word in all the training dataset. A fusion parameter α controls the importance of the image and the likelihood of training set:

$$P(w_j|d_i) = (1 - \alpha) \frac{N(w_j, d_i)}{\sum_{j=1}^{N_w} N(w_j, d_i) + \sum_{t=1}^{N_v} N(v_t, d_i)} + \alpha \frac{N(w_j, d)}{M}. \quad (2.8)$$

where $N(w_j, d_i)$ denotes the count of word w_j in the image d_i , $N(v_t, d_i)$ is the count of visual word v_t in image d_i , M is the number of training images, and $N(w_j, d)$ is the number of images in which the word w_j appears. Similarly, the probability of a visual word given an image d_i is estimated by its likelihood in the image smoothed by its likelihood in training set, controlled by a parameter β :

$$P(v_t|d_i) = (1 - \beta) \frac{N(v_t, d_i)}{\sum_{j=1}^{N_w} N(w_j, d_i) + \sum_{t=1}^{N_v} N(v_t, d_i)} + \beta \frac{N(v_t, d)}{M}. \quad (2.9)$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER III

PROBABILISTIC GRAPHICAL MODEL

In order to study the probabilistic model for image annotation and retrieval, in this dissertation, we firstly introduce a basic idea via probabilistic graphical model, which is used to effectively design a generative model for several applications. The probabilistic model will be explained in Section 3.1. In Section 3.2, we briefly discuss Probabilistic Latent Semantic Analysis (pLSA), which can be expressed as graphical model.

3.1 Introduction to Probabilistic Graphical Model

Probabilistic Graphical Model provides a general framework for representing models in which a number of random variables interact. It has used in many field such as expert system, image/video understanding, pattern recognition/classification etc. Figure 3.1 shows a basic example of graphical model in which each node represents a random variable or a set of random variables. Each edge in the graph represents the qualitative dependencies among the random variables in the graph. The absence of an edge between two nodes means that those two nodes are independent. The qualitative dependencies between both of variables connected via edges are specified via parameterized conditional distributions, or non-negative called “potential functions”. The structure of graphical model is specified as a joint probability distribution over all the random variables.

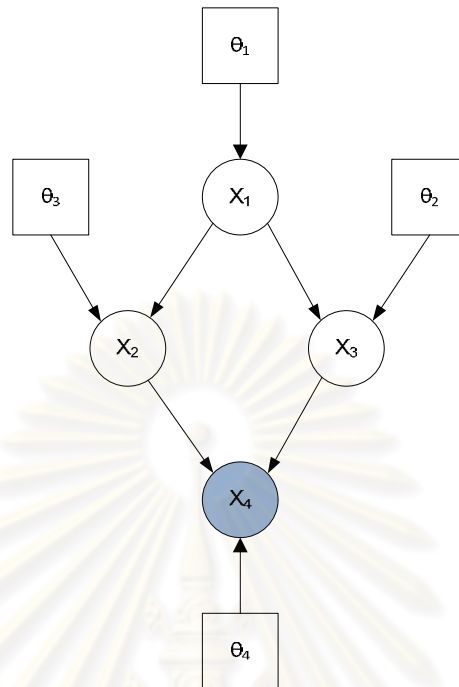


Figure 3.1 Example of directed graphical model. Circle nodes denote random variables, while squared nodes denote parameters of the model. The shaded circle denotes observed random variables, while non-shaded circled nodes represent hidden random variables.

Generally there are two main types of graphical model. The first model is called directed graphical model, also known as Bayesian networks, where all the edges are considered to have a direction from parent to child nodes denoting the conditional dependency among the corresponding random variables. In addition we assume that the directed graph is acyclic, i.e., contain no cycles. In contrast, the second model is called undirected graphical model, also known as Markov random field. In the following paragraph, we will explain high level explanation of the basic idea of directed graphical model.

Directed Graphical Model or Bayesian Network presents the joint probability distribution of d random variables, $x = (X_1, X_2, \dots, X_d)$, by a directed acyclic graph in which each node, i , representing variable X_i , depends on directed edges from its set of parent node π_i . So the joint distribution of X can be factored into the product of

conditional probability distribution of each variable given its parents, we can write general form of joint probability distribution in Eq. (3.1), for each setting \mathbf{x} of the variable \mathbf{x} ,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^d p(x_i|\mathbf{x}_{\pi_i}, \boldsymbol{\theta}_i) \quad (3.1)$$

Where, given its parents, X_i is statistical independent of all other variables expecting itself, $\boldsymbol{\theta}_i$ is the set of parameter being the conditional probability which relates \mathbf{X}_{π_i} to X_i . And the set of all parameter denoted by $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots, \boldsymbol{\theta}_d)$.

After constructing the joint probability distribution, all parameters are then estimated using EM algorithm or other techniques. In the next section, we explain the problem of learning maximum likelihood (ML) parameter when all the variables are observed, and introduce the EM Algorithm. Finally, we explain the Bayesian approach in which a posterior distribution over parameter is inferred from data.

3.1.1. Maximum Likelihood Learning

Given a data set \mathbf{D} of N independent and identically distribution observations of all variables in the graphical model $\mathbf{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$, where $\mathbf{x}^{(n)} = (x_1^{(n)}, \dots, x_d^{(n)})$, the likelihood can be written in Eq. (3.2) as a function of the all parameters which are proportional to the probability of the observed data.

$$p(\mathbf{D}|\boldsymbol{\theta}) = \prod_{n=1}^N p(\mathbf{x}^{(n)}|\boldsymbol{\theta}) \quad (3.2)$$

Estimating parameters are previously unknown, so we can maximize the log likelihood as shown in Eq. (3.3).

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \log p(\mathbf{D}|\boldsymbol{\theta}) = \sum_{n=1}^N \log p(\mathbf{x}^{(n)}|\boldsymbol{\theta}) \\ \mathcal{L}(\boldsymbol{\theta}) &= \sum_{n=1}^N \sum_{i=1}^d \log p(x_i^{(n)}|\mathbf{x}_{\pi_i}, \boldsymbol{\theta}_i)\end{aligned}\tag{3.3}$$

If the parameters $\boldsymbol{\theta}_i$ given its parents are assumed that are distinct and functionally independent of the parameters governing the conditional probability distribution of others in the graphical model, then the log likelihood can be decomposed to the sum of local terms involving each node and its parents:

$$\mathcal{L}_{ML}(\boldsymbol{\theta}) = \sum_{i=1}^d \mathcal{L}_i(\boldsymbol{\theta}_i)\tag{3.4}$$

where, $\mathcal{L}_i(\boldsymbol{\theta}_i) = \sum_{n=1}^N \log(x_i^{(n)}|\mathbf{x}_{\pi_i}, \boldsymbol{\theta}_i)$. While each $\mathcal{L}_i(\boldsymbol{\theta}_i)$ can be maximized independently as a function of $\boldsymbol{\theta}_i$.

In contrast, Maximum a posterior (MAP), its parameters are estimated incorporating prior knowledge assumed in the form of a probability distribution $p(\boldsymbol{\theta})$. The objective of MAP estimation is to find the parameter setting that maximizes the posterior probability, $p(\boldsymbol{\theta}|\mathbf{D})$, being proportional to the prior times the likelihood. If the posterior probability are independently on observed data, and each parameter, $p(\boldsymbol{\theta}) = \prod_{i=1}^d p(\theta_i)$, then MAP estimation can be found by maximizing

$$\mathcal{L}_{MAP}(\boldsymbol{\theta}) = \sum_{i=1}^d \mathcal{L}_i(\boldsymbol{\theta}_i) + \log p(\boldsymbol{\theta}).\tag{3.5}$$

The final term in Eq. (3.5), log prior function, can be seen as a regularize function, which can help reduce the overfitting in situations where there is insufficient data for the parameter to be well-determined.

3.2 Probabilistic Latent Semantic Analysis

The Probabilistic Latent Semantic Analysis model (pLSA) [3-6] the initial element of the aspect model family, was proposed by Hofmann [3] as a probabilistic alternative to the linear algebra-based Latent Semantic Analysis (LSA) method. It proposes an interesting probabilistic formulation of the concept of topics in text collections, decomposing a document into a mixture of latent aspects defined by a multinomial distribution over the words in the vocabulary. In the following description of the pLSA in [4, 5] the term document defines sets of discrete elements, referring to either text or image documents. This aspect model is a latent variable or hidden variable model for co-occurrence data which associates an unobserved class variable with each observation, an observation being the occurrence of the word in a particular document.

For pLSA, the probabilities of variables are defined in this following. Documents or images are represented by a discrete random variable D , that can take the values d_i , where $i \in \{1, \dots, M\}$, and M is the number of documents. Each document or image is represented by a set of elements regarded as the observation of a discrete random variable X , that can take the value x_j , where $j \in \{1, \dots, N\}$, x_j ranges over the vocabulary words in the text case, over the different visual words for image case, So N is the number of elements. Under the pLSA assumptions, the observation of elements X is conditionally independent with the observation of document D given a hidden variable Z , called a latent aspect. This discrete variable is not observed, and can take the positive values z_k , where $k \in \{1, \dots, K\}$, and K is the number of aspects. The joint probability of observing d_i and x_j is thus given by the marginalization over all the possible values z_k , and can form $P(d_i, x_j)$ in Eq. (3.6)

$$P(d_i, x_j) = \sum_{k=1}^K P(d_i, z_k, x_j) \quad (3.6)$$

The conditional independence between d_i and x_j given z_k translates into factorization of the joint probability of d_i , x_j , and z_k in Eq. (3.7) and being substituted into Eq. (3.6) to obtain a joint probability in Eq. (3.9)

$$P(d_i, z_k, x_j) = P(d_i)P(x_j|z_k)P(z_k|d_i) \quad (3.7)$$

$$P(d_i, x_j) = \sum_{k=1}^K P(d_i)P(x_j|z_k)P(z_k|d_i) \quad (3.8)$$

$$P(d_i, x_j) = P(d_i) \sum_{k=1}^K P(x_j|z_k)P(z_k|d_i) \quad (3.9)$$

The conditional independence probability in Eq. (3.7) makes each document d_i a mixture of latent aspects, defined by the multinomial distribution $P(z_k|d_i)$. Each latent aspect z_k is defined by the multinomial distribution $P(x_j|z_k)$, which gives the probability of each element x_j , given an latent aspect z_k . And $P(d_i)$ denote the probability that a word occurrence will be observed in a particular document d_i . We also represent this latent aspect model in term of graphical model depicted in Figure 3.2. Using these definitions, the generative model is defined for elements and documents co-occurrences by the following scheme:

1. Select a document d_i with probability $P(d_i)$,
2. Pick a latent aspect z_k with probability $P(z_k|d_i)$,
3. Generate a element x_j with probability $P(x_j|z_k)$.

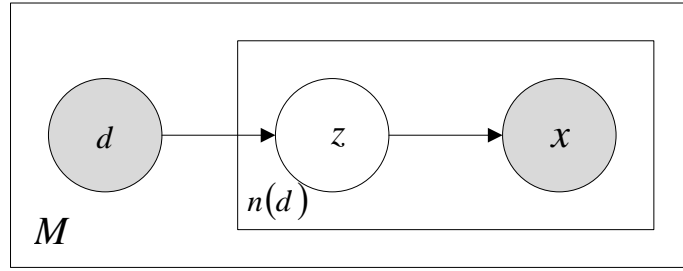


Figure 3.2 Graphical Model of pLSA where M is number of training document and $n(d_i) = \sum_j^N n(d_i, x_j)$ is the number of elements in document d_i .

The conditional probability distributions, $P(x|z)$, and $P(z|d)$, are considered as multinomial distribution, given that both z and d are discrete random variables. The parameters of these distributions are estimated by the EM algorithm [3, 34]. For a vocabulary of N different elements, $P(x|z)$ is a N -by- K table that stores the parameter of the K multinomial distributions $P(x|z_k)$. And a K -by- M table $P(z|d)$ stores the parameters of the M multinomial distribution $P(z|d_i)$ that describes the training document d_i .

3.2.1. pLSA Learning using EM-Algorithm

In order to learn two parameter tables, one can use a maximum likelihood formulation of the learning problem. The standard procedure for maximum likelihood estimation in latent variable model is EM algorithm for incomplete data [34]. EM alternates two steps: (i) an Expectation (E) step where posterior probabilities are completed for the latent variables, based on the current estimates of the parameters, (ii) a Maximization (M) step where parameters are updated based on the so-called expected complete data log-likelihood which depends on the posterior probabilities computed in the E-step.

For the E-step to infer the latent aspect z_k given the observation pair d_i and x_j from the previous estimate of the model parameter, that simply applies Bayes' rule in Eq. (3.7) to obtain

$$P(z_k | d_i, x_j) = \frac{P(x_j | z_k)P(z_k | d_i)}{\sum_{k=1}^K P(x_j | z_k)P(z_k | d_i)}. \quad (3.10)$$

For the M-step has to maximize the expected complete data log-likelihood $\mathbf{E}(\mathcal{L}^c)$ being given by Eq. (3.15)

$$\mathbf{E}(\mathcal{L}^c) = \mathbf{E}_{z_k}(\ln P(D, X, Z) | d_i, x_j) \quad (3.11)$$

$$\mathbf{E}(\mathcal{L}^c) = \mathbf{E}_{z_k} \left(\ln \prod_{i=1}^M \prod_{j=1}^N P(d_i, x_j, z_k)^{n(d_i, x_j)} \middle| d_i, x_j \right) \quad (3.12)$$

$$\mathbf{E}(\mathcal{L}^c) = \sum_{k=1}^K \left(\sum_{i=1}^M \sum_{j=1}^N n(d_i, x_j) \ln P(d_i, x_j, z_k) P(z_k | d_i, x_j) \right) \quad (3.13)$$

$$\mathbf{E}(\mathcal{L}^c) = \sum_{k=1}^K \left(\sum_{i=1}^M \sum_{j=1}^N n(d_i, x_j) (\ln P(d_i) + \ln P(z_k | d_i) P(x_j | z_k)) P(z_k | d_i, x_j) \right) \quad (3.14)$$

$$\mathbf{E}(\mathcal{L}^c) \cong \sum_{i=1}^M \sum_{j=1}^N n(d_i, x_j) \left[\sum_{k=1}^K P(z_k | d_i, x_j) \ln \{P(z_k | d_i) P(x_j | z_k)\} \right] \quad (3.15)$$

In order to take care of the normalization constraints in Eq. (3.15) has to be augmented by appropriate Lagrange multiplier τ_k and ρ_i , is given by Eq. (3.16).

$$\mathcal{H} = \mathbf{E}(\mathcal{L}^c) + \sum_{k=1}^K \tau_k \left(1 - \sum_{j=1}^N P(x_j | z_k) \right) + \sum_{i=1}^M \rho_i \left(1 - \sum_{k=1}^K P(z_k | d_i) \right) \quad (3.16)$$

Maximization of \mathcal{H} with respect to the probability mass function leads to the following set of stationary equations

$$\sum_{i=1}^M n(d_i, x_j) P(z_k | d_i, x_j) - \tau_k P(x_j | z_k) = 0 \quad (3.17)$$

$$\sum_{j=1}^N n(d_i, x_j) P(z_k | d_i, x_j) - \rho_i P(z_k | d_i) = 0 \quad (3.18)$$

After eliminating the Lagrange multiplier, we can obtain the M-step re-estimation equations

$$P(x_j | z_k) = \frac{\sum_{i=1}^M n(d_i, x_j) P(z_k | d_i, x_j)}{\sum_{j=1}^N \sum_{i=1}^M n(d_i, x_j) P(z_k | d_i, x_j)} \quad (3.19)$$

$$P(z_k | d_i) = \frac{\sum_{j=1}^N n(d_i, x_j) P(z_k | d_i, x_j)}{n(d_i)} \quad (3.20)$$

The E-step and the M-step equations are alternated until a termination condition is met. This can be a convergence condition, but it may also use a technique known as early stopping.

3.2.2. Inference: pLSA of a new document

The conditional probability distribution over aspects $P(z | d_{new})$ can be inferred for an unseen document d_{new} . The folding-in method proposed in [5] maximizes the likelihood of the document d_{new} with a partial version of the EM algorithm described where $P(x | z)$ is obtained from training and kept fixed meaning that is not updated at M-step. In doing so $P(z | d_{new})$ maximizes the likelihood of the document d_{new} with respect to the previously learned $P(x | z)$ parameters.

3.2.3. Convergence Condition

The convergence condition [5] is used to control overfitting of the pLSA model using early stop. The probability of aspects given each validation document $P(z | d_{valid})$ is first estimated using the folding-in method, describe in previous section. The fold-in likelihood of the validation set given by the current parameters is then computed by

$$\mathcal{L}_{valid} = \prod_{i=1}^{M_{valid}} \prod_{j=1}^N P(x_j | d_i) = \prod_{i=1}^{M_{valid}} \prod_{j=1}^N \sum_{k=1}^K P(x_j | z_k) P(z_k | d_i). \quad (3.21)$$

And the model parameters corresponding to the highest fold-in likelihood value is kept.

3.2.4. Image Annotation with pLSA Model

In [5], they proposed three alternatives to learn a pLSA model for the co-occurrence of visual and textual feature in annotated images. The first approach is pLSA-mixed using the early integration of visual and textual modalities. The two others based on a variation of the pLSA EM algorithm can constrain the estimation of the conditional distributions of latent aspect given the training documents from one of the two modalities only. This allows to choose between textual and the visual modality to estimate the mixture of aspects in a given document. They are called asymmetric pLSA.

3.2.4.1 pLSA-Mixed

This model learns a standard pLSA model from a concatenated representation of the textual and the visual features $x(d) = \{w(d), v(d)\}$. Both of textual and visual co-occurrences are learned simultaneously to estimate $P(x|z)$. The distribution of word given an aspect and the distribution of visual feature given an aspect $P(v|z)$ are then extracted from $P(x|z)$, and normalized such that $\sum_{j=1}^{N_w} P(w_j|z_k) = 1$ and $\sum_{j=1}^{N_v} P(v_j|z_k) = 1$. Intrinsically, pLSA-mixed assumes that two modalities have an equivalent importance in estimating the latent aspect space.

3.2.4.2 Asymmetric pLSA

Because of the imbalance of pLSA-mixed between the number of words and the number of visual features in the images, the parameters of the mixture aspects are not freely controlled in practical. Therefore, Asymmetric pLSA was proposed by choosing between the textual and the visual modality to estimate the mixture aspects in a given document. Moreover, an image is modeled by a mixture of latent aspects that is either defined by its text captions or by its visual features, so we obtain the model in different aspect mixture weight. In learning with this model, the aspect distributions $P(z|d_i)$ are learned for all training documents from one modality only, visual or textual modality, and then kept for the other modality, textual or visual modality respectively.

3.2.4.3 Annotating an image with pLSA Model

Given new visual features $v(d)$ and the previously estimated $P(v|z)$ parameters, the conditional probability distribution $P(z|d_{new})$ is inferred for a new image d_{new} using the standard folding-in procedure for a new document. Given these estimated mixture weights, the conditional distribution over words given this new image is given by:

$$P(w|d_{new}) = \sum_{k=1}^K \prod_{j=1}^{N_w} P(w_j|z_k)P(z_k|d_{new}), \quad (3.22)$$

where the probability table $P(w|z)$ was estimated from training data.

CHAPTER IV

THE PROPOSED TECHNIQUE

In this chapter, we propose a novel model based on pLSA for image annotation and retrieval. We will call this new model, “two latent aspects pLSA model”, because we use two hidden random variables, the first latent aspect is used for representing that the images on a corpus relate their word, and the second latent aspect is used for representing visual features of each image relating with their word. The graphical model is shown in Figure 4.1.

We obtain an inspiration from the advantage of standard pLSA being semi-supervised learning. It can learn unseen images from the existing parameter of its model, but is poor in the performance. So we adapt pLSA to supervised learning technique to obtain better performance. In another inspiration, the novel image retrieval system should be expert system meaning the system can be used for searching by several query examples, such as image or text, and should also annotate an unseen image, then adding with its caption or tag for image indexing. So we will propose a novel model for image retrieval system that can work in multitasks, such as image search by text or image and image annotation, based on Generative Model using pLSA. Because the advantage of generative model is that we can design the model being useful to several works by a joint probability distribution. Therefore, in generative model by bottom-up reason, we can make our hypothesis which image documents in a corpus cause occurring words, and a word of images causes occurring Bag-of-features. Afterward, we add two latent variables to mapping between the image documents and their word by latent variable z , and between their word and their BoF by latent variable l . In adding latent variable l , a bag-of-feature of each image is mapped by one-to-one mapping with latent variable l , and the words are mapped by many-to-one with latent

variable l . That means each image can be represented by many words liking image annotation concept. Similarly, In adding latent variable z , each bag-of-word of a annotated image is mapped by one-to-one mapping with latent variable z , and image document in corpus are generated by many-to-one mapping with latent variable z . That means the images in corpus are generates by several words of their annotated image. By this hypothesis, we can thus design the model shown in Figure 4.1 which we expect our two latent aspects pLSA model to improve that the performances of several tasks are better than standard pLSA.

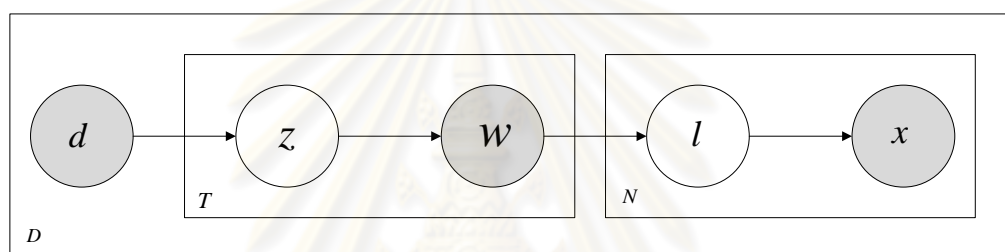


Figure 4.1 our proposed model, two latent aspects pLSA

First, we define the following notations:

- Images are represented by a observed random discrete variable \mathcal{D} that can take the value $d_i, i \in \{1, \dots, D\}$, where D is the number of document in training data set.
- Captions on each image can be represented by observed random variable \mathcal{W} that can take the value $w_j, j \in \{1, \dots, T\}$ where T is the size of a word vocabulary including of several words that are used for labeling an image in image annotation task, and are used for searching images by textual in image retrieval task.
- Bag-of-features, being basis unit in our approach, are presented by a observed visual random variable \mathcal{X} that can take the value $x_n, n \in \{1, \dots, N\}$, where N is the size of visual vocabulary, constructed by K-Mean algorithm, that are used for searching images by an example image.

Under the assumption of probabilistic graphical model in two latent aspects pLSA, all of the observed random variables are conditionally independent distribution given by two hidden variables, so we will define two latent aspects in this following:

- The first latent variables \mathcal{Z} to which we will refer as a visual latent aspect can take the values $l_m, m \in \{1, \dots, M\}$, where M is the number of visual aspects.
- The second hidden variable \mathcal{L} to which we will refer as a word latent aspect can take the values $z_k, k \in \{1, \dots, K\}$, where K is the number of word aspects.

Thus, the joint probability $P(d_i, w_j, x_n)$ of all observed d_i, w_j and x_n is given by the marginalization over all the possible values z_k and l_m :

$$P(d_i, w_j, x_n) = \sum_{k=1}^K \sum_{m=1}^M P(d_i, z_k, w_j, l_m, x_n) \quad (4.1)$$

By the probabilistic graphical model in Figure 4.1, the joint probability of all random variables including of observation and non-observation can expressed in Eq. (4.2)

$$P(d_i, z_k, w_j, l_m, x_n) = P(d_i)P(w_j|z_k)P(z_k|d_i)P(x_n|l_m)P(l_m|w_n) \quad (4.2)$$

Therefore, the joint probability distribution of all observed variable can rewrite as

$$P(d_i, w_j, x_n) = P(d_i) \underbrace{\sum_{k=1}^K P(w_j|z_k)P(z_k|d_i)}_{P(w_j|d_i)} \underbrace{\sum_{m=1}^M P(x_n|l_m)P(l_m|w_j)}_{P(x_n|w_j)} \quad (4.3)$$

There are the two conditional independent assumption expressed in Eq. (4.3). The first term makes each image d_i as a mixture of word latent aspects, defined by the multinomial distribution $P(z|d_i)$. Each word latent aspect z_k is defined by the multinomial distribution $P(w|z_k)$ which gives the probability of each word w_j given by each word aspect z_k . Moreover, in Eq. (4.3), when considering its second term, it also indicate that each w_n as a mixture of visual latent aspects, defined by the multinomial distribution $P(l|w_j)$, where each visual latent aspect is defined by the multinomial distribution $P(x|l_m)$ which gives the probability of each visual word x_n .

Furthermore, the joint probability distribution of pairs between two observations can be computed by marginalization that can be expressed in Eq. (4.4), Eq. (4.5) and Eq. (4.6). These joint probabilities are useful for image annotation and retrieval of which we will explain in below section.

$$P(d_i, w_j) = P(d_i) \underbrace{\sum_{k=1}^K P(w_j|z_k)P(z_k|d_i)}_{P(w_j|d_i)} \quad (4.4)$$

$$P(w_j, x_n) = \underbrace{\sum_{k=1}^K P(w_j|z_k)P(z_k|d_i)}_{P(w_j|d_i)} \underbrace{\sum_{m=1}^M P(x_n|l_m)P(l_m|w_j)}_{P(x_n|w_j)} \quad (4.5)$$

$$P(d_i, x_n) = P(d_i) \sum_{j=1}^T \left[\underbrace{\sum_{k=1}^K P(w_j|z_k)P(z_k|d_i)}_{P(w_j|d_i)} \underbrace{\sum_{m=1}^M P(x_n|l_m)P(l_m|w_j)}_{P(x_n|w_j)} \right] \quad (4.6)$$

4.1 Learning parameters using EM Algorithm

Our model consists of four conditional probabilities, $P(w_j|z_k)$, $P(z_k|d_i)$, $P(x_n|l_m)$, and $P(l_m|w_j)$ which are assumed as multinomial distribution. Their parameters are estimated by the Expectation Maximization Algorithm. For word vocabulary of T different words, $P(w|z)$ is a T -by- K table that stores the parameter of word latent aspects K being multinomial distribution. And the K -by- D table stores the parameters of the D multinomial distribution $P(z|d_i)$ that describes the training document d_i . Moreover, for visual vocabulary of N different visual words, $P(x|l)$ is a N -by- L table that stores the parameter of visual word latent aspects L , which still is multinomial distribution. On the contrary, the L -by- T table is relative between visual word and word, as it stored the parameter of T multinomial distribution $P(l|w_d)$ that describes the training words of word vocabulary.

In order to learning these parameters, in this work, we use EM algorithm including of 2 steps: E-step complete posterior probabilities of two dimension latent

aspects and M-step four parameters are updated based on expectation of the posterior probabilities of E-step.

For E-step, the probability of two latent aspects depending on all observation can simply apply Bayes' rule for Eq.(4.2), which obtain in Eq. (4.7) and Eq. (4.8).

$$P(z_k, l_m | d_i, w_j, x_n) = \frac{P(d_i, z_k, w_j, l_m, x_n)}{P(d_i, w_j, x_n)} \quad (4.7)$$

$$P(z_k, l_m | d_i, w_j, x_n) = \frac{P(w_j | z_k) P(z_k | d_i) P(x_n | l_m) P(l_m | w_n)}{\sum_{k=1}^K P(w_j | z_k) P(z_k | d_i) \sum_{m=1}^M P(x_n | l_m) P(l_m | w_j)} \quad (4.8)$$

$$P(z_k, l_m | d_i, w_j, x_n) = P(z_k | d_i, w_j) P(l_m | w_j, x_n)$$

For the M-step, we have to maximize the expected complete data log-likelihood $\mathbf{E}(\mathcal{L}^c)$ by Eq. (7.13)

$$\mathbf{E}(\mathcal{L}^c) = \mathbf{E}_{z_k, l_m} (\ln P(\mathcal{D}, \mathcal{Z}, \mathcal{W}, \mathcal{L}, \mathcal{X}) | d_i, w_j, x_n) \quad (4.9)$$

$$= \mathbf{E}_{z_k, l_m} \left(\ln \prod_{i=1}^D \prod_{j=1}^T \prod_{n=1}^N P(d_i, z_k, w_j, l_m, x_n)^{n(d_i, w_j, x_n)} \middle| d_i, w_j, x_n \right) \quad (4.10)$$

$$= \sum_{\substack{k=1, m=1, i=1, \\ j=1, n=1, \\ K, M, D, \\ T, N}}^{K, M, D, \\ T, N} \left(n(d_i, w_j, x_n) \ln (P(w_j | z_k) P(z_k | d_i)) \right) (P(z_k, l_m | d_i, w_j, x_n)) \quad (4.11)$$

$$+ \sum_{\substack{k=1, m=1, i=1, \\ j=1, n=1, \\ K, M, D, \\ T, N}}^{K, M, D, \\ T, N} \left(n(d_i, w_j, x_n) \ln (P(x_n | l_m) P(l_m | w_n)) \right) (P(z_k, l_m | d_i, w_j, x_n))$$

Where $n(d_i, w_j, x_n)$ is the count of element x_n correspond to word w_j in document d_i . In order to take care of the normalization constraints in Eq. (4.11), has to be augmented by appropriate Lagrange multipliers τ_i, ρ_k, η_j and β_m , is given by Eq. (4.12)

$$\begin{aligned}
\mathcal{H} = \mathbf{E}(\mathcal{L}^c) &+ \sum_{i=1}^D \tau_i \left(1 - \sum_{k=1}^K P(z_k | d_i) \right) + \sum_{k=1}^K \rho_k \left(1 - \sum_{j=1}^T P(w_j | z_k) \right) \\
&+ \sum_{j=1}^T \eta_j \left(1 - \sum_{m=1}^M P(l_m | w_j) \right) \\
&+ \sum_{m=1}^M \beta_m \left(1 - \sum_{n=1}^N P(x_n | l_m) \right)
\end{aligned} \tag{4.12}$$

Maximization of with respect to the probability mass functions leads to the following set of stationary equations:

$$\sum_{j=1}^T \sum_{m=1}^M \sum_{n=1}^N n(d_i, w_j, x_n) P(z_k, l_m | d_i, w_j, x_n) - \tau_i P(z_k | d_i) = 0 \tag{4.13}$$

$$\sum_{i=1}^D \sum_{m=1}^M \sum_{n=1}^N n(d_i, w_j, x_n) P(z_k, l_m | d_i, w_j, x_n) - \rho_k P(w_j | z_k) = 0 \tag{4.14}$$

$$\sum_{i=1}^D \sum_{k=1}^K \sum_{n=1}^N n(d_i, w_j, x_n) P(z_k, l_m | d_i, w_j, x_n) - \eta_j P(l_m | w_j) = 0 \tag{4.15}$$

$$\sum_{k=1}^K \sum_{i=1}^D \sum_{j=1}^T n(d_i, w_j, x_n) P(z_k, l_m | d_i, w_j, x_n) - \beta_m P(x_n | l_m) = 0 \tag{4.16}$$

After eliminating the Lagrange multipliers and reforming simplify using Eq. (4.8), we can obtain the M-step re-estimation equations:

$$P(z_k | d_i) = \frac{\sum_{j=1}^T n(d_i, w_j) P(z_k | d_i, w_j)}{n(d_i)} \tag{4.17}$$

$$P(w_j | z_k) = \frac{\sum_{i=1}^D n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{j=1}^T \sum_{i=1}^D n(d_i, w_j) P(z_k | d_i, w_j)} \tag{4.18}$$

$$P(l_m | w_j) = \frac{\sum_{n=1}^N n(w_j, x_n) P(l_m | w_j, x_n)}{n(w_j)} \tag{4.19}$$

$$P(x_n | l_m) = \frac{\sum_{j=1}^T n(w_j, x_n) P(l_m | w_j, x_n)}{\sum_{n=1}^N \sum_{j=1}^T n(w_j, x_n) P(l_m | w_j, x_n)} \tag{4.20}$$

Where $n(d_i, w_j) = \sum_{n=1}^N n(d_i, w_j, x_n)$ and $n(w_j, x_n) = \sum_{i=1}^D n(d_i, w_j, x_n)$. From Eq. (4.11), (4.17), (4.18), (4.19) and (4.20) the algorithm can be shown in **Algorithm 4.1** to estimate the parameters of our model.

Algorithm 4.1: Learning a two latent aspect pLSA

Random initialization of $P(z|d)$, $P(z|w)$, $P(x|l)$ and $P(l|w)$ distribution tables

While increase in the likelihood of data $> T$ do

[E-step]

for all (d_{new}, w_j, x_{new}) such that $n(d_{new}, w_j, x_{new}) > 0$, and $\forall k$ **do**

$$P(z_k, l_m | d_i, w_j, x_n) = \frac{P(w_j | z_k) P(z_k | d_i) P(x_n | l_m) P(l_m | w_j)}{\sum_{k=1}^K P(w_j | z_k) P(z_k | d_i) \sum_{m=1}^M P(x_n | l_m) P(l_m | w_j)}$$

end for

[M-step]

for $k \in \{1, \dots, K\}$ and $i \in \{1, \dots, D\}$ **do**

$$P(z_k | d_i) = \frac{\sum_{j=1}^T n(d_i, w_j) P(z_k | d_i, w_j)}{n(d_i)}$$

end for

for $j \in \{1, \dots, T\}$ and $k \in \{1, \dots, K\}$ **do**

$$P(w_j | z_k) = \frac{\sum_{i=1}^D n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{j=1}^T \sum_{i=1}^D n(d_i, w_j) P(z_k | d_i, w_j)}$$

end for

for $m \in \{1, \dots, M\}$ and $j \in \{1, \dots, T\}$ **do**

$$P(l_m | w_j) = \frac{\sum_{n=1}^N n(w_j, x_n) P(l_m | w_j, x_n)}{n(w_j)}$$

end for

for $n \in \{1, \dots, N\}$ and $m \in \{1, \dots, M\}$ **do**

$$P(x_n | l_m) = \frac{\sum_{j=1}^T n(w_j, x_n) P(l_m | w_j, x_n)}{\sum_{n=1}^N \sum_{j=1}^T n(w_j, x_n) P(l_m | w_j, x_n)}$$

end for

compute likelihood of data

end while

4.2 Annotating an image with two-pLSA

Given a new BoF extracted from a new image and the estimated parameters of two-pLSA model $P(l_m|w_j)$. By this process, firstly, the new BoF is matched with $P(x|l_m)$, being the parameter of each cluster in latent variable l_m , to estimate the similarity measurement between BoF and visual latent variable. And secondly, in the cluster l_m , the probability of words $P(l_m|w_j)$ from learning process is used for selecting the set of word by the criterion in Eq. (4.21)

$$P(w_j|x_1, \dots, x_n) \approx \left\{ \sum_{m=1}^M \left[\prod_{n=1}^N P(x_n|l_m)^{BoF_{new}(x_n)} P(l_m|w_j) \right] \right\} \times \sum_{k=1, i=1}^{K, D} P(w_j|z_k) P(z_k|d_i) \quad (4.21)$$

The parameters $P(x_n|l_m)$ and $P(l_m|w_j)$ are estimated by the learning process. These parameters are the conditional probability, grouping words in a latent variable l_m to estimate the probability of the words. Based on Bayes' rule, the probabilities of words are ranked by increasing of their probability to annotate a new image.

4.3 Image Retrieval by Text Query

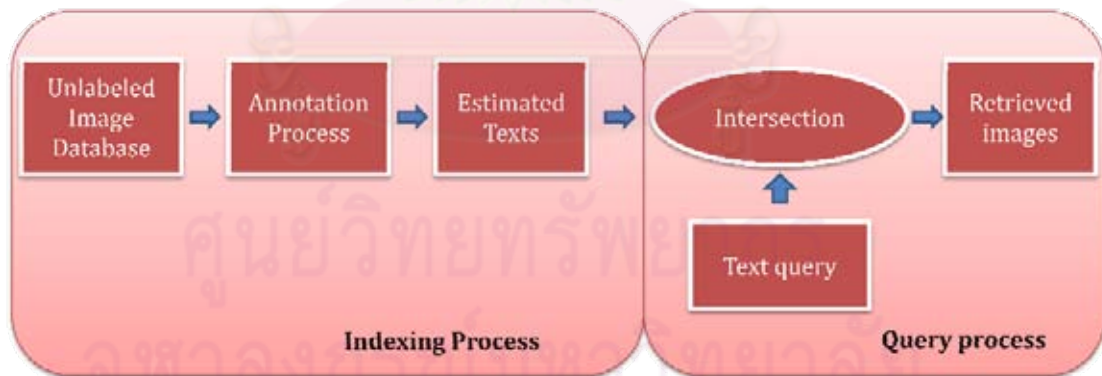


Figure 4.2 Diagram of Image retrieval by Text Query

In Figure 4.2, the estimated words of unlabeled images are estimated and indexed by annotation process. The histogram intersection technique is used to match between estimated words and query word. The retrieved images are ranked with the score of the intersection by increasing.

4.4 Image Retrieval by Image Query

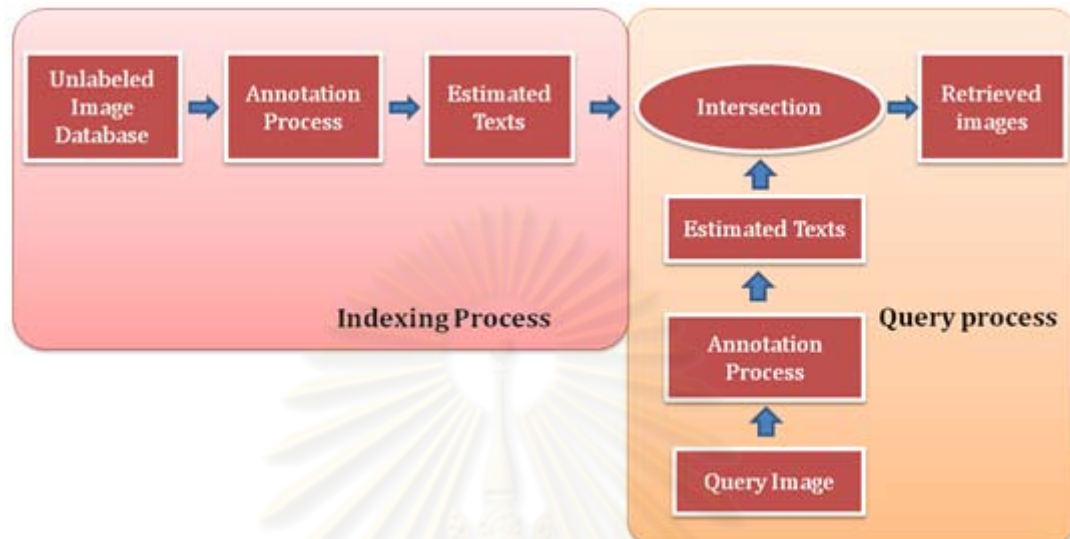


Figure 4.3 Diagram of Image retrieval by Image Query

In Figure 4.3, the estimated words of unlabeled images are estimated and indexed by annotation process. An image query are extracted a set of estimated word by the model, and these estimated word are matching to the estimated word of unlabeled images using histogram intersection technique and the score of matching are sorted by increase to return retrieved image.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER V

EXPERIMENTAL RESULTS

This chapter describes the experimental results of image annotation and retrieval used in this research. Details of performance of constructing the visual vocabulary by varying size of visual vocabulary, and adjusting of the number of latent of variables are elaborated comparing with the existing annotation models, namely naïve Bayes, CMRM, pLSA, and two-pLSA model to evaluate the performance of automatic image annotation. By the automatic image annotation algorithm, the unlabeled images on the corpus are indexed using a set of text. The indexed texts on the corpus are used for matching between a text query and text index.

Based on the visual vocabulary construction, the dimensionality reduction of SIFT features using PCA are evaluated to study the effect in changing the size of visual word and the effect of reducing the dimension of SIFT feature. These effects will directly affect the performance of image retrieval, which will describe in Section 5.1. By selecting the visual word on the visual vocabulary, the experiment in Section 5.2 is to evaluate the effect to criterion of selection which affect to performance of image retrieval. The purpose of this dissertation is to develop the model of image annotation to identify and retrieve images for a large scale image. It is useful for image retrieval on the accuracy performance and speed of searching images, and can be conducive to a query with words or images by the proposed model, called Two-Probabilistic Latent Semantic Analysis. By indexing with words corresponded to image, the designed system can be supported for multiple functions, such as identifying meaningful image automatically, image retrieval by word, and image retrieval with image example , to retrieve images quickly, which are evaluated in Section 5.3, Section 5.4 and Section 5.5. The PASCAL 2008 dataset is evaluated in Section 5.3 in term of the performance of

image annotation and image retrieve with texts and image query. Another dataset, “MIRFlickr25000” is evaluated to study the effect of the general images which are taken by photographers. And the final experiment in Section 5.5 is to measure the visual changing vocabulary construction that effect to the performance of image annotation and retrieval.

5.1 Experiment 1: Dimensionality Reduction of SIFT using PCA for object Categorization

To study reducing the dimensionality number of SIFT, and the reduction on the size of visual vocabulary by adjusting the number of cluster on K-mean process. In this experiment, we investigate the approach of applying PCA to reduce SIFT dimension. The BoF is constructed by vector quantization technique; each center of cluster is a visual word. All center of cluster forms visual vocabulary. Histogram of BoF is used for BoF representation which is further trained by the k-NN model. For testing, a query image's local feature is extracted by PCA-SIFT descriptor and being represented by histogram of BoF. An object category is identified as an output by majority voting using k-NN model

Several simulations have been carried out to verify such objectives. There are three simulation categories. The first two is to investigate PCA-SIFT technique in terms of the number of dimension and mean average precision. The last one is to investigate the effect of visual word reduction toward the performance.

5.1.1. Dataset and simulation

In this experiment, the Caltech-4 dataset consists of images from 6 object categories. This database contains the images in the following categories: airplane (1074 images), background (900 images), car rear (1155 images), car bg (1370 images), face (450 images) and motorbike (826 images). The total numbers of images are 5,775. In the simulations, images are randomly spitted into 2 separated sets; 10% for training, and the remaining for testing purpose. SIFT feature is being extracted from

images by using 4,000 random keypoints. For baseline method, it is a reference method without our proposed PCA-SIFT. The number of dimension is equal to 128 dimensions, and contains 4,000 random keypoints of an image.

For learning process, we use the k-NN, where $k = 10$, for 10 sampling vote for all object categories. However, average precision of each object category is computed by the confidence score to obtain the probability from 10 voting samples.

5.1.2. Performance of PCA-SIFT on dimension reduction and average precision

In these simulations, we evaluate performance of PCA-SIFT in terms of reduction of PCA-SIFT's dimension and average precision. For PCA-SIFT, we vary ϵ for 3 different values which are 0.1, 0.25 and 0.5. The higher value of ϵ implies the lower number of dimension of PCA space. Table 7.1 shows the number dimension of PCA at 3 different values of ϵ . The total dimension of PCA-SIFT space are 24, 40 and 115 for $\epsilon = 0.5, 0.25$ and 0.2, respectively. When comparing to the number dimension of baseline technique, i.e., 128, it shows that PCA-SIFT could reduce the number of dimension up to 80%.

Table 5.1 The number dimension Of PCA

Object	$\epsilon=0.5$	$\epsilon =0.25$	$\epsilon =0.1$	Baseline
airplanes	5	7	16	128
background	4	10	23	128
car (rear)	5	8	17	128
car (bg)	3	6	12	128
face	4	8	21	128
motorbike	3	11	26	128
Dimension	24	40	115	128

Table 5.1 shows the average precision results compared between the baseline and our proposed technique. From the results, the values of mean AP for both

techniques are around the same. When considered together with the number of dimension stated in Table 5.2, it is clear that reducing the dimension does not affect the average precision, i.e., the efficiency of PCA-SIFT does not decrease.

Table 5.2 Average Precision

Object	$\epsilon = 0.5$	$\epsilon = 0.25$	$\epsilon = 0.1$	Baseline
airplanes	0.90	0.91	0.92	0.92
background	0.69	0.71	0.73	0.79
car (rear)	0.92	0.93	0.93	0.93
car (bg)	0.89	0.90	0.91	0.91
face	0.74	0.83	0.86	0.89
motorbike	0.81	0.81	0.82	0.83
Mean AP	0.83	0.85	0.86	0.88

The precision-recall curve of baseline technique is shown in Figure 5.1. This curve shows the ability to retrieve accurately with each recall. Normally, the trend of precision will decrease when the recall increases. In general sense, the best model should be able to retrieve all images according to each category.

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

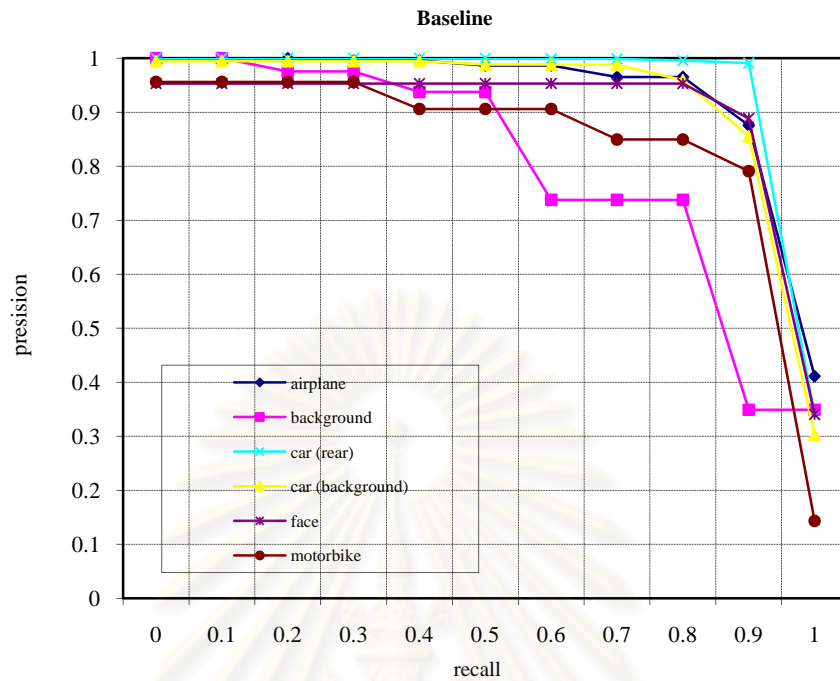
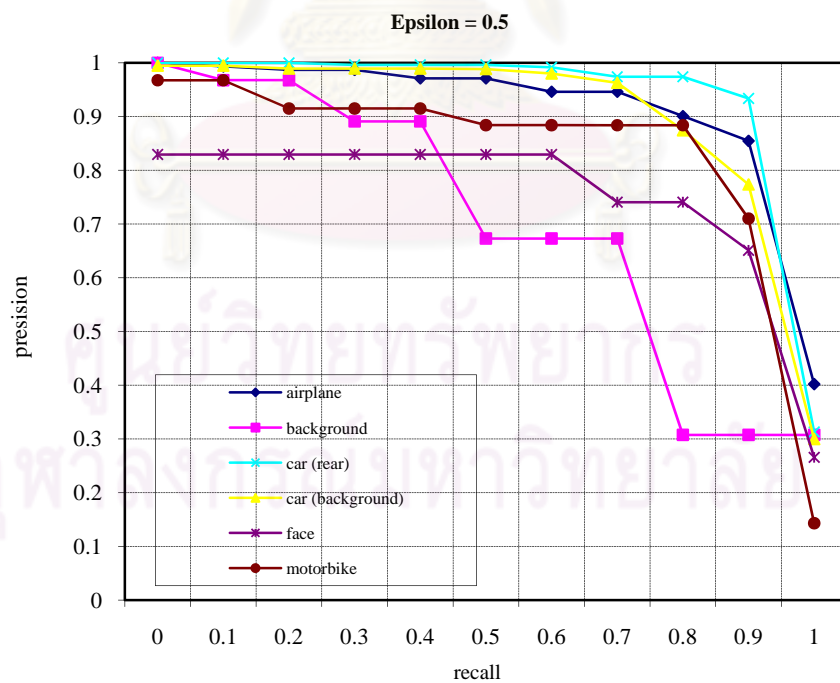


Figure 5.1 Precision-recall curve on baseline technique

Figure 5.2 precision-recall curve of PCA-SIFT for $\epsilon = 0.5$

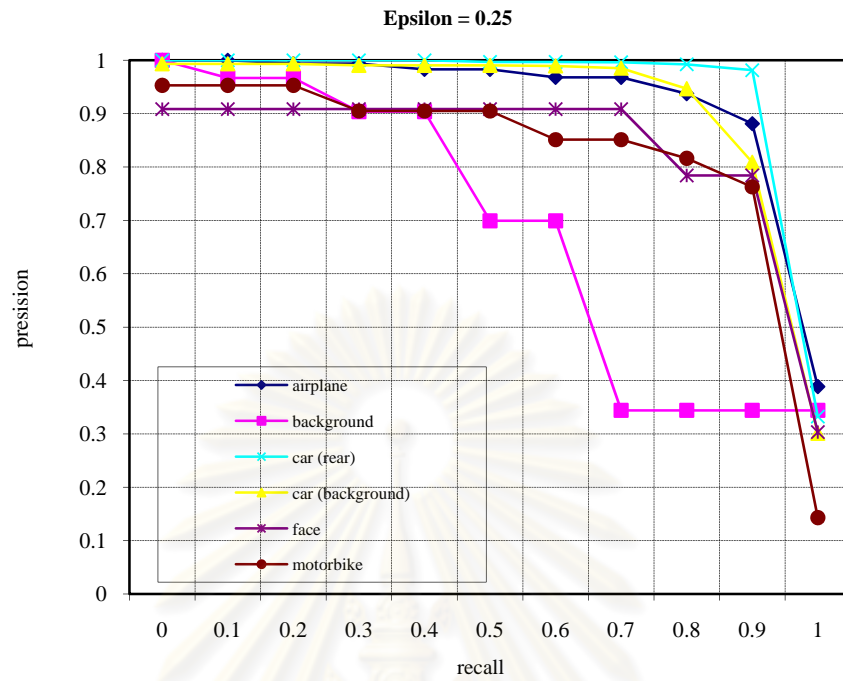


Figure 5.3 Precision-recall curve PCA-SIFT for $\varepsilon = 0.25$

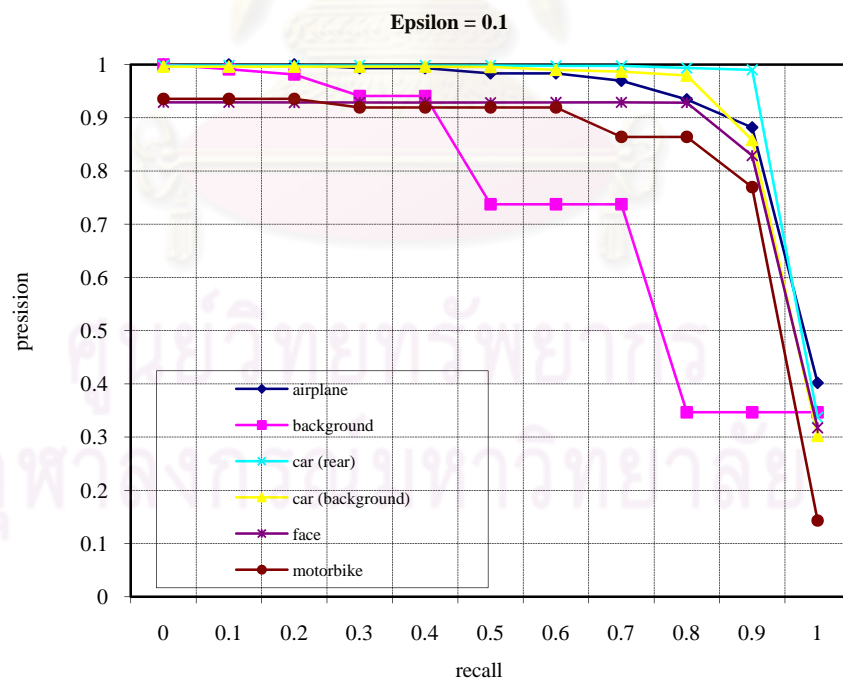


Figure 5.4 Precision-recall curve PCA-SIFT for $\varepsilon = 0.1$

Figure 5.2, Figure 5.3, and Figure 5.4 show the precision-recall curve for three different values of $\varepsilon = 0.5, 0.25$ and 0.1 , respectively. The results from three figures indicate that all techniques including the baseline do not work well with background category images. This is because the background images consist many visual words that make it impossible for the model to identify the correct category.

5.1.3. Performance of PCA-SIFT on the number of visual word

In these simulations, we analyze the effect of the the number of visual word toward the precision. We construct the BoF representation, i.e., 3 visual vocabularies of 60, 600, 6000 and 12000 visual words for baseline technique. A BoF of each category is constructed individually with BoF of size 10, 100, 1000 and 2000 visual words. Finally, we construct a BoF by concatenating individual BoF to obtain 60, 600, 6000 and 12000 visual words. All BoFs are equal to the size of visual word of the baseline technique.

Table 7.3 show the mean AP result on the different number of visual words comparison between the baseline and our proposed PCA-SIFT with concatenated BoF on each category. Simulation results indicate that for small number of visual words, i.e., 60, our proposed technique achieves higher precision than the baseline technique. However, we could not see this improvement when the number of visual words is higher. This suggests the number of visual words play more important role in the retrieval performance. To reduce complexity of the system, the approach toward the reduction of local features' dimension is more promising approach.

Table 5.3 Number of Visual Words and Mean AP

Method	#visual word	mean AP
Baseline	60	0.61
PCA-SIFT $\epsilon = 0.5$	60	0.70
PCA-SIFT $\epsilon = 0.25$	60	0.71
PCA-SIFT $\epsilon = 0.1$	60	0.75
Baseline	600	0.85
PCA-SIFT $\epsilon = 0.5$	600	0.80
PCA-SIFT $\epsilon = 0.25$	600	0.82
PCA-SIFT $\epsilon = 0.1$	600	0.83
Baseline	6000	0.87
PCA-SIFT $\epsilon = 0.5$	6000	0.81
PCA-SIFT $\epsilon = 0.25$	6000	0.83
PCA-SIFT $\epsilon = 0.1$	6000	0.85
Baseline	12000	0.88
PCA-SIFT $\epsilon = 0.5$	12000	0.81
PCA-SIFT $\epsilon = 0.25$	12000	0.85
PCA-SIFT $\epsilon = 0.1$	12000	0.87

5.2 Experiment 2: Selecting Visual word with criterion

Since the visual words are constructed by all parts of the entire image, each visual word is associated with several objects or background parts. Thus some visual words should be selected as the parts of an object. These parts correspond to the target object class and characterize discrimination between an object of interest and other objects. We compare two selection criteria, the maximum probability and entropy criterion and discriminative criterion

5.2.1. *Dataset and Simulation Condition*

In this dissertation, the Caltech-4 dataset consists of images from 6 object categories is used. This database contains the images in the following categories: airplane (1074 images), background (900 images), car rear (1155 images), car back (1370 images), face (450 images) and motorbike (826 images). The total numbers of images are 5,775 images. The images of an object are separated into 2 sets: learning set and testing set. The ratio of the number of learning images to testing images is 0.5, which the number of learning equals to the number of testing. To evaluate the effectiveness of the technique, several simulations have been carried out to verify such objectives. There are two simulation categories. First, we investigate elimination of small scale parameter of SIFT. Second, we investigate the detection technique using the maximum probability and entropy criteria.

5.2.2. *Results using scale parameter by threshold*

In this subsection, we investigate an experiment regarding the noise from object and the background image using a threshold technique of scale parameter. We varies a threshold, T_s , for 5 different values which are 0, 0.2, 0.4, 0.6, 0.8, and 1.0. The threshold equals to zero that use the standard SIFT technique. Our results show in Figure 5.5, Figure 5.6, and Figure 5.7.

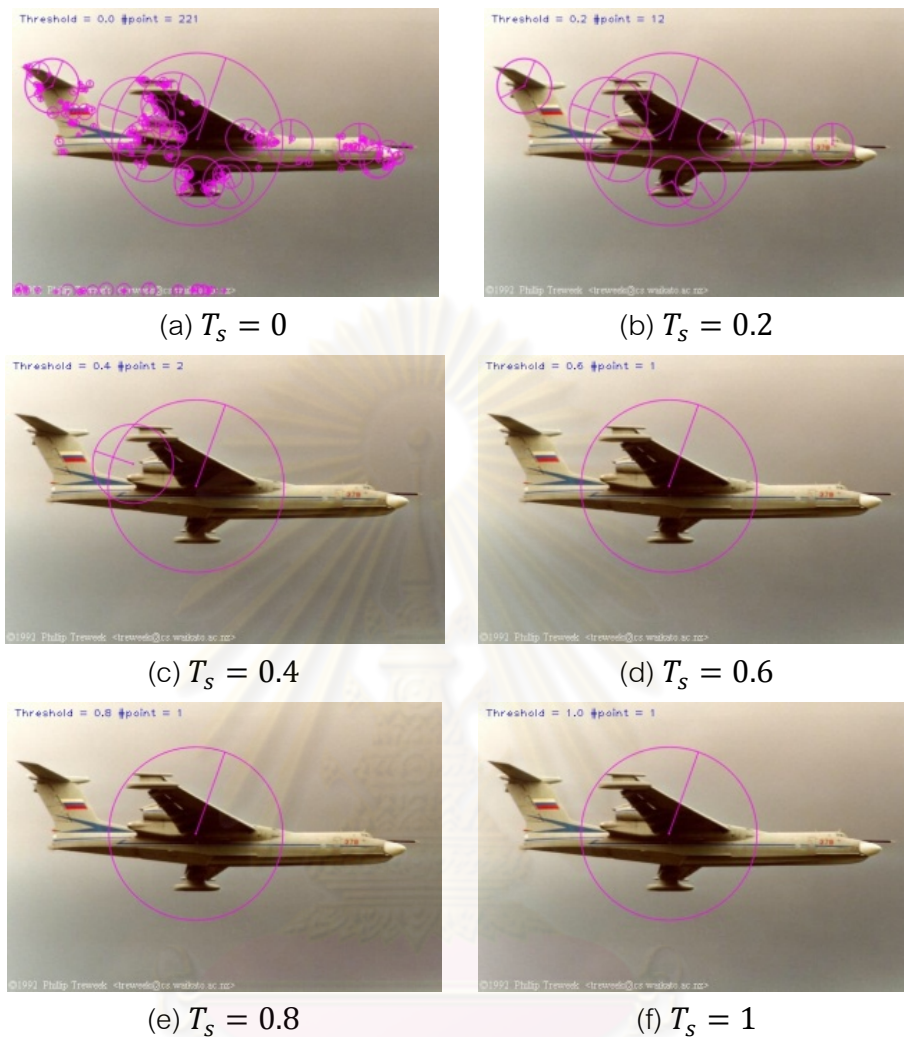


Figure 5.5 The results of eliminated on an airplane image

In Figure 5.5, we show an image in airplane class. When the threshold is increased, the number of points is decreasing. The numbers of points equal to 221, 12, 2, 1, 1, and 1 points with the thresholds, T_s , which are 0, 0.2, 0.4, 0.6, 0.8, and 1.0 respectively. The percentage of point reduction from using standard SIFT is approximately equal to 100, 5, 0.9, 0.4, 0.4 and 0.4 percent respectively. Seeing that an airplane image, the points from our elimination technique cover only the region of airplane

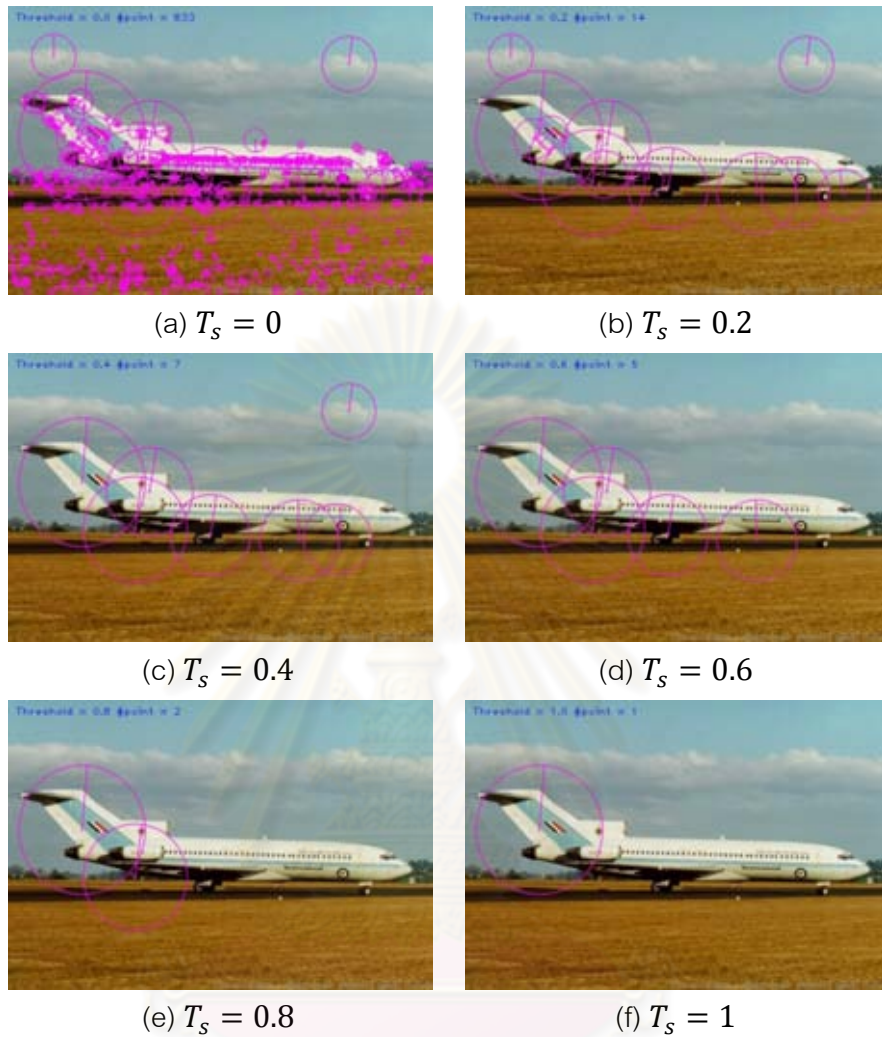


Figure 5.6 The results of elimination on an airplane image with noises

In Figure 5.6, the number of points respectively equal to 833, 14, 7, 6, 2, and 1 point in same airplane class, but an image appears some noise at the background. The percentage of point reduction from using standard SIFT is approximately equal to 100, 1.6, 0.84, 0.72, 0.24 and 0.12 percent respectively. Our technique can eliminate the noise from the background and still retain the point covering the interesting object.

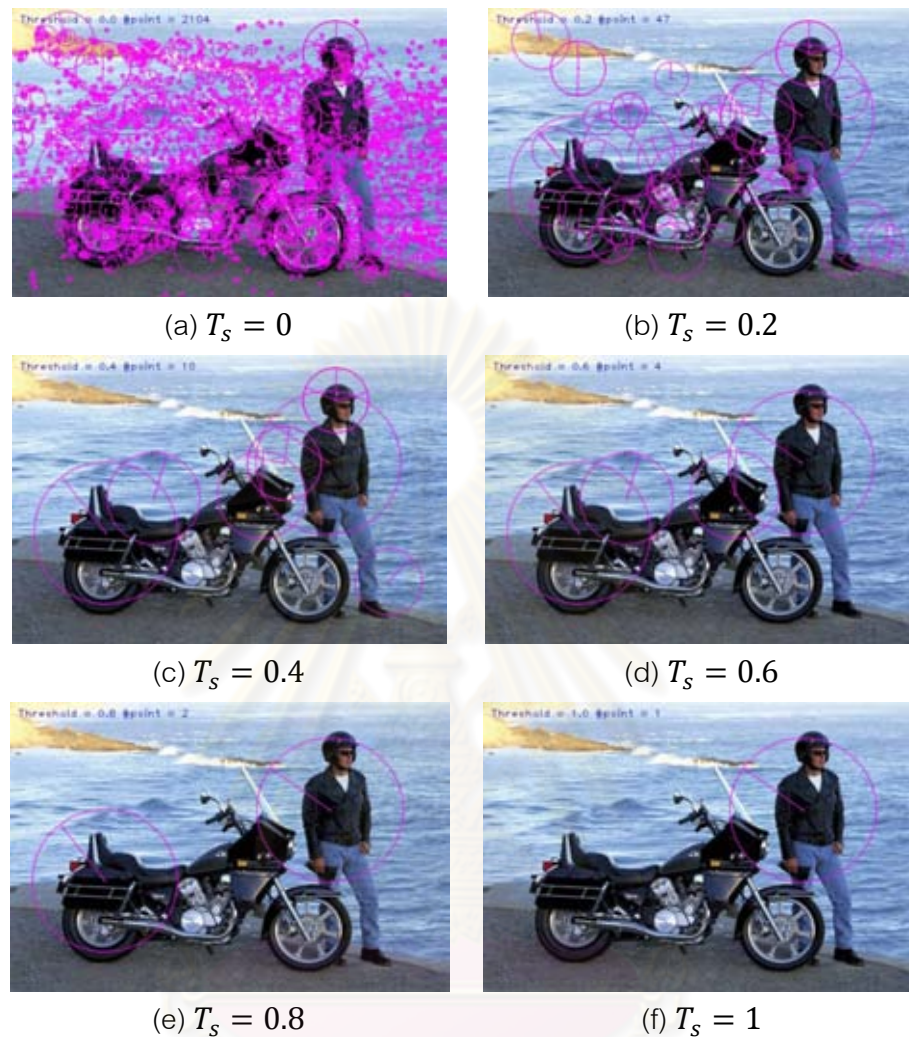


Figure 5.7 The results of elimination on a motorbike and human image

In Figure 5.7, we investigate two objects including of motorbike and a person on an image with clutter background. The number of points respectively equals to 2104, 49, 10, 4, 2, and 1 point. The percentage of point reduction from using standard SIFT is approximately equal to 100, 2.33, 0.48, 0.4, 0.19 and 0.04 percent, respectively.

5.2.3. Maximum Probability and Entropy Criterion for choosing visual words

In the first criterion we use the maximum probability and entropy cluster as parts of object. On each object, the probability of visual words can be defined as the ratio of the number of local features assigned to a visual to the total number of local features. We can write the probability in the form of Eq. (5.1)

$$P(z_i) = \sum_{j=1}^{V^{(u)}} P(z_i | v_j^{(u)}) \quad (5.1)$$

, where $V^{(u)}$ is the number of local features, $v_j^{(u)}$, correspond to the object u from validation set. And z_i is a visual word at i . Afterward, if those probabilities of visual word are larger than the predetermined threshold, those visual words are chosen to parts of the object. The predetermined threshold should equal to $1/K$, where K the number of visual word from K-mean. Others with a small probability value are considered noise. Moreover, the non-common noise can be regarded by the entropy. The entropy is defined as

$$H(z_i) = - \sum_{j=1}^{V^{(u)}} P_j(z_i) \ln (P_j(z_i)) \quad (5.2)$$

The entropy reflects the distribution of a certain visual word in each image within the same object. If the distribution is more uniform, the selected classifier is more common. So the visual words with larger entropies are kept.

We investigate an experiment regarding maximum probability and entropy criterion when the number of local classifiers is increased. In Figure 5.8, when the size of visual word is increased, the recall also increases, but the precision decreases. It shows that this selection is not efficient to detect parts of object accurately, although we increase the number of visual words. Because the selecting set of local classifiers is not suitable to discrimination between foreground and background.

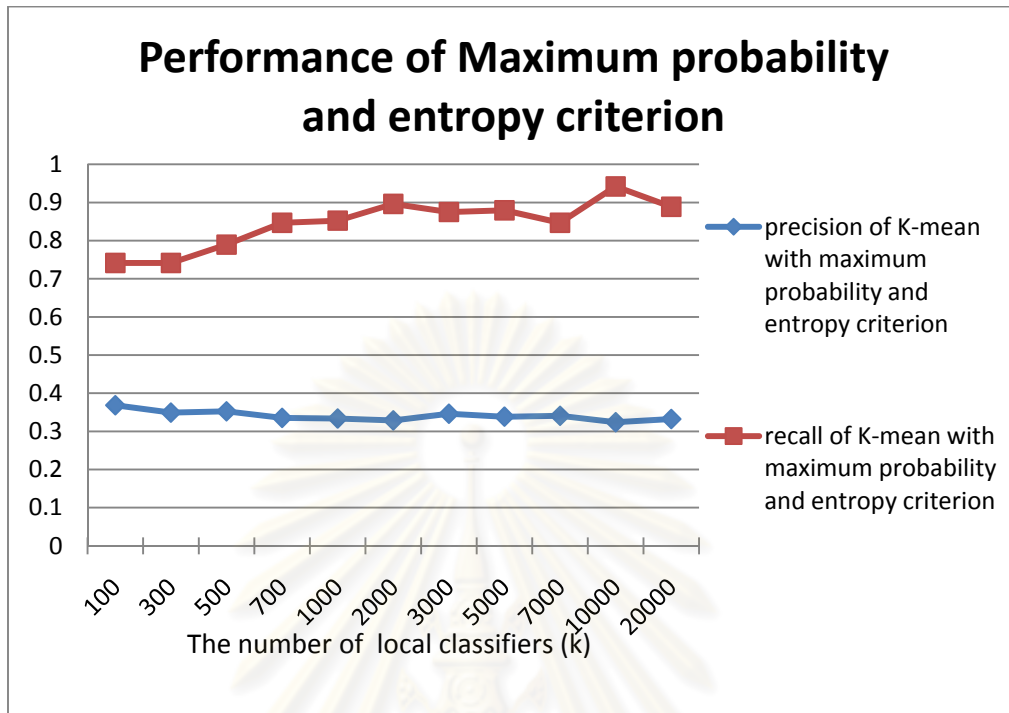


Figure 5.8 The maximum probability and entropy criterion result

5.2.4. Discriminative Criterion for choosing visual words

In the second criteria we choose local classifiers according to their ability to discriminate between object-class and background. This criterion is similar to Bayes' rule where local classifiers are chosen by the probability of positive features more than the probability of negative features. We can write this criterion in form of Eq. (7.25).

$$\sum_{j=1}^{v^{(u)}} P(z_i | v_j^{(u)}) > \sum_{j=1}^{\bar{v}^{(u)}} P(z_i | v_j^{(\bar{u})}), \quad \text{where } i = \{1, \dots, K\} \quad (5.3)$$

, where $\bar{v}^{(u)}$ is the number of negative visual word, $v_j^{(\bar{u})}$, which correspond to the non-interesting object u in images from validation set, and z_i is a visual word at i . Institively, this criterion is well suited for detection purpose because it performs selection by optimal classification rate.

We investigate an experiment regarding discriminative criterion when the number of visual word is increased. In Figure 5.9, when the number of visual word is increased, the recall also decreases, but the precision increases. It shows that this selection is efficient to detect parts of object accurately. However, we must use a larger number of local classifiers to obtain better performance, and when we use the number of local classifier over than 20000, the performance is not better because of trade-off between the precision and recall.

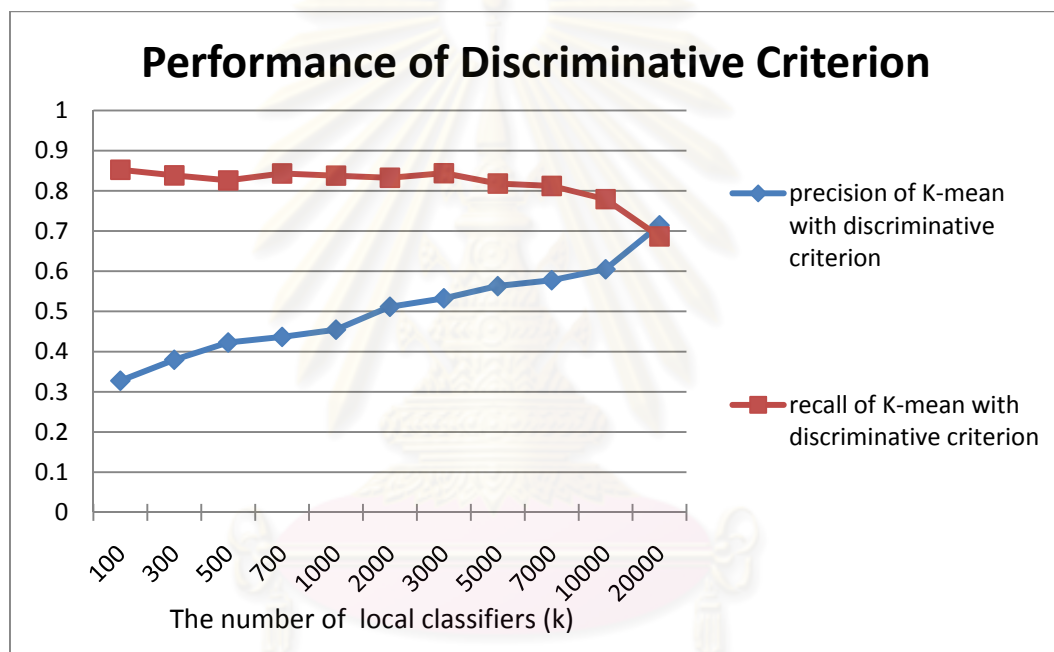


Figure 5.9 The discriminative criterion result

From the experimental results shown in Section 5.2.3, and Section 5.2.4, the technique of discriminative selection can improve the performance in precision of image retrieval where increasing the number of visual words. However the maximum probability and entropy approach is not effect to precision of image retrieval performance although the number of visual word is increased. Therefore, to improve the precision of image retrieval, the discriminative criterion has achieved the satisfied results where choosing a set of visual word by considering the number parts of object more than the number parts of non-object.

5.3 Experiment 3: Performance in PASCAL 2008 Dataset

Objective of this experiment is to evaluate the performance of image annotation with three existing models comparing with our approach in PASCAL 2008 Dataset. Based on the models of automatic image annotation in unlabeled image dataset, the performances of image retrieval are reported in term of querying using text and querying using an image example in this section.

5.3.1. Dataset and Simulation Condition

In this experiment, we used the PASCAL 2008 Dataset [28] which contains a total number of 4,340 annotated images, and 20 words. An image contains at least one word. The data from training/validation in PASCAL dataset are divided into training and testing process by the ratio of 50%.

Table 5.4 Statistics of PASCAL 2008 Image dataset

Word	#image	#word	Word	#image	#word
Aeroplane	236	316	Diningtable	105	110
Bicycle	192	269	Dog	388	477
Bird	305	476	Horse	198	285
Boat	207	336	Motorbike	204	272
Bottle	243	457	Person	2002	4148
Bus	100	129	Pottedplant	180	361
Car	466	840	Sheep	64	145
Cat	328	378	Sofa	134	151
Chair	351	623	Train	151	166
Cow	74	130	TVmonitor	215	274

Table 5.4 summarizes the number of words and images. The PASCAL 2008 dataset is an unbalance dataset which unequally contains the number of word. The word, "Person" is the highest number of both image and word than the others. In this

simulation, the PASCAL 2008 dataset is divided into 2 separated set; 50% for training to construct the visual vocabulary and the models, and remaining for testing purpose. SIFT feature of training data set is extracted from images by Hessian Affine detection. These SIFT features are used for constructing visual vocabulary of BoF prototype by K-mean and images are represented into BoF prototype as new image features. BoF of entire training dataset are used for input feature to construct the models namely Naïve Bayes, Cross Media Relevance Model (CMRM), pLSA and our proposed model, two-pLSA model. To measure the performance among these models, the testing dataset is used for image annotation by mean Average and processing time.

5.3.2. Image Ranking with latent variable \mathbf{z}

The latent variable \mathbf{z} can serve as grouping words which often are found together. It allows to take advantage of browsing the images in retrieval process from training dataset. The ranking images in a training dataset respect to their probability given a latent variable \mathbf{z}_k , to illustrate what this latent variable captures in the dataset. Assuming that $P(\mathbf{d})$ is uniform, $P(\mathbf{d}_i|\mathbf{z}_k)$ becomes proportional to the corresponding to

$$P(\mathbf{d}_i|\mathbf{z}_k) = \frac{P(\mathbf{z}_k|\mathbf{d}_i)P(\mathbf{d}_i)}{P(\mathbf{z}_k)} \propto P(\mathbf{z}_k|\mathbf{d}_i). \quad (5.4)$$

Given each latent variable \mathbf{z}_k , the top-ranked images according to $P(\mathbf{d}|\mathbf{z}_k)$ illustrate its grouping words. Figure 5.10 displays the 11 most probable images from the training set of PASCAL 2008 dataset. The top-ranks images represent latent variable at 0, 1, 4, 6, 10, 12, and 17 related to “chair and sofa”, “bottle and tvmonitor”, “person”, “horse”, “train”, “cow” and “airplane” respectively. That shows that the two-pLSA model is also useful for browsing images.

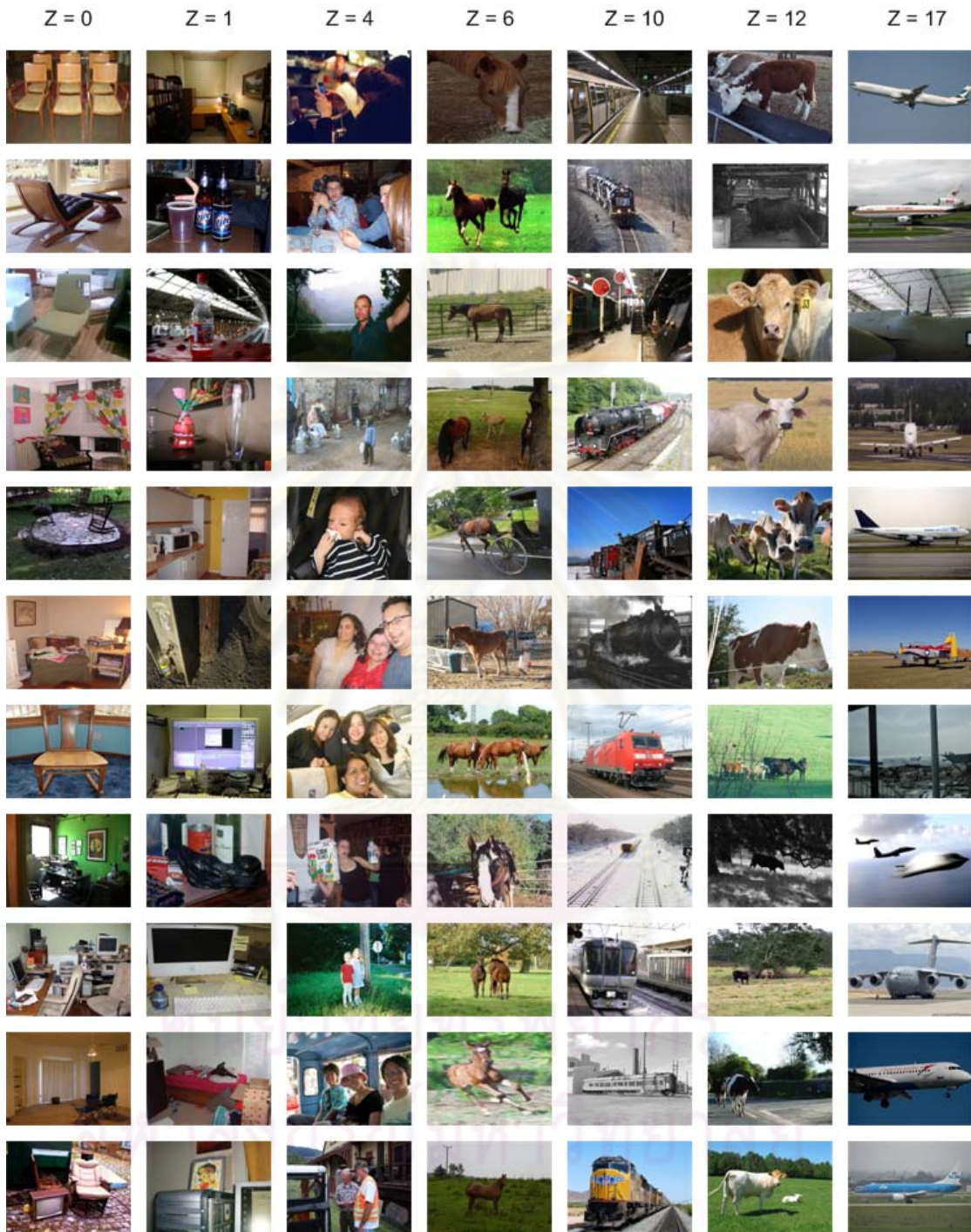


Figure 5.10 The 11 most probable images in 7 latent variables of grouping in 20 latent variables from PASCAL 2008 Dataset variable.

5.3.3. Image Annotation: mean Average Precision Performance and Processing Time

To measurement the annotation quality of the models, we compute the mean Average Precision of the given labels for each image in the test set. In this experiment, we investigate the annotating algorithms by varying the number of visual words which are constructed from K-mean algorithm ranging from 100, 200,300, 400, 500, 600, 700, 800, 900 and 1000. For pLSA, the numbers of latent variable z are equal to 10 and 20 variables. For two-pLSA model, the number of latent variables z is fixed by 20 variables, and the number of latent variable l are 10 and 20 variables. So the comparison results are shown in Figure 5.11.

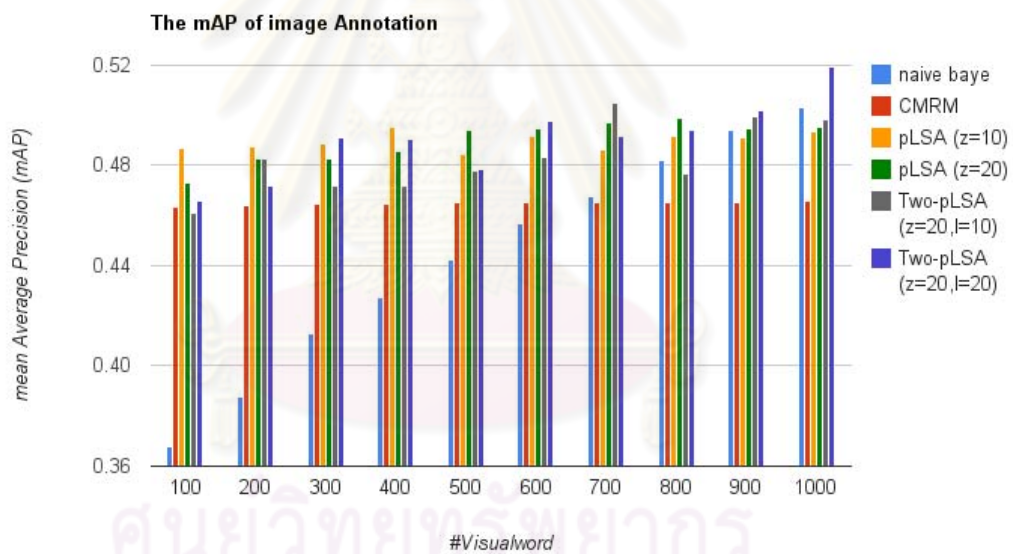



Figure 5.11 mean Average Precision of image annotation in PASCAL 2008 dataset

From Figure 5.11, the mAP value of every model increases as the size of visual vocabulary increases, except for the CMRM that has nearly constant mAP. When considering the number of visual word of $N = 1000$, the naive model obtains the mAP equal to 0.5. In addition the mAP of the other models are less than the of naive except for the two-pLSA model. The mAP of the two-pLSA model equals to 0.51 when the number of both latent variable z and l is set to 20. In case of comparing pLSA model and two-pLSA model, their mAP are slowly increased according to the increasing

number of latent variable. This implies that the number of latent variable is not a necessary factor to improve the mAP performance.

In cases where $N = 900,600$ and 300 visual words, the mAP values of two-pLSA are higher than that of other models. From the experimental results, it can be concluded that our proposed model is suitable for automatic image annotation. Thus the supervised learning technique will achieve better performance than unsupervised learning when compared with the pLSA.



An example of Image Annotation

Manually Label: person, bicycle

Model	Rank1	Rank2	Rank3	Rank4	Rank5	Correct (bicycle)	Average Precision
Naïve bayes	Chair	Bird	Person	Car	Dog	(8)	0.2917
CMRM	Person	Chair	Car	Bottle	Pottedplant	(11)	0.5909
pLSA ($z = 20$)	Person	Sheep	Dog	Bird	Horse	(10)	0.6000
Two-pLSA ($z = 20, l = 20$)	Person	Cat	Car	Bird	Dog	(8)	0.6250

Figure 5.12 An example image of improved annotation performance

In Figure 5.12, the two-pLSA technique improves the ranking of the annotation. The high value of average precision means that the model can annotate correctly in the lowest possible rank. The two-pLSA can annotate the word “bicycle” correctly at the 8th rank, while other models will order this label in the highest rank. Thus the two-pLSA model archives the higher mAP value than the others.

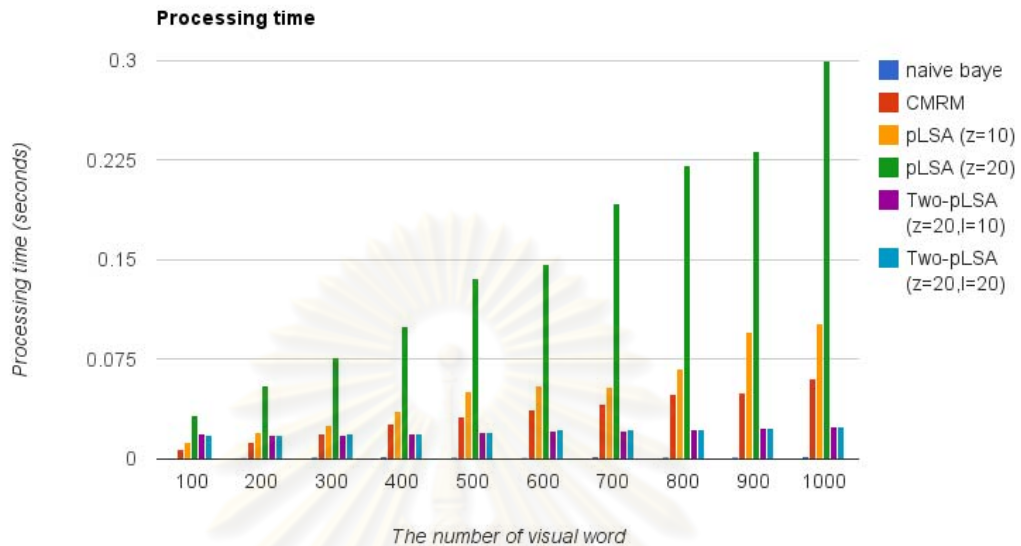


Figure 5.13 Processing time of image annotation per an image of the PASCAL dataset

Moreover, we compare the processing time for each an annotated image of each model, as shown in Figure 5.13. For every model, their processing time are increased according to the size of visual vocabulary. The processing time of naïve bayes model is the lowest. The processing time of two-pLSA model is around 7.2 milliseconds per one image, and also shows that the two-pLSA is faster than others expert the naïve Bayes model. When compared the performance of mAP and processing time, the two-pLSA is still the best model for image annotation.

5.3.4. Text-based Image Retrieval in unlabeled images

In this experiment, the performance of image retrieval by searching with text query is evaluated by varying the size of visual vocabulary in the set of image is unlabeled, but the images are labeled by the annotation model. The models are compared, including the naïve Bayes, CMRM, pLSA and two-pLSA model with the parameter of the different number of latent variables. The unlabeled images in the testing set will be annotated by the words using these models. The probabilities of words are used as the text indexing for image retrieval process. For matching between

two text index, which is estimated by annotation process, the histogram matching technique is used. The results of these models are shown in Figure 5.14.

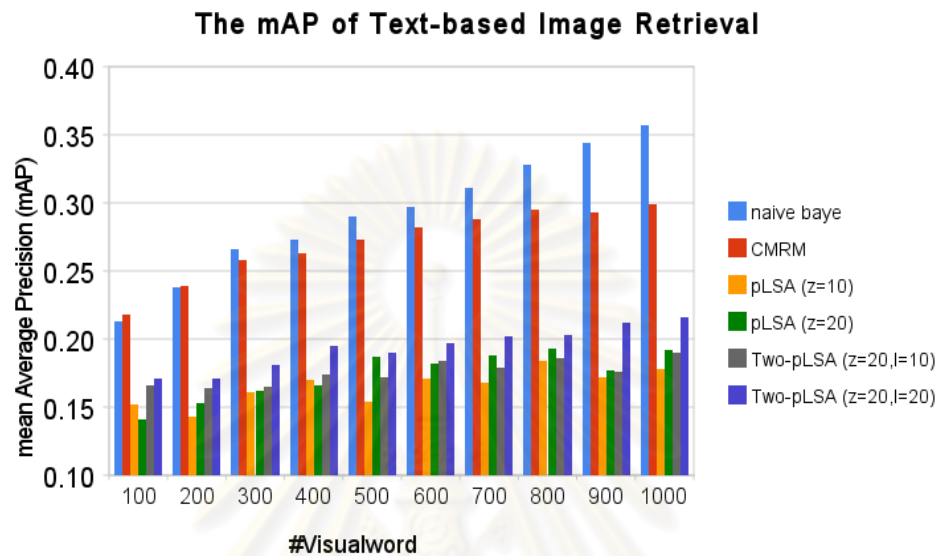


Figure 5.14 mAP of image retrieval by text query in PASCAL 2008 dataset

From Figure 5.14, both of pLSA and two-pLSA model performance are lower than CMRM and naïve bays model, but two-pLSA model is higher than pLSA model. The performance of pLSA and two-pLSA is lower, because the estimated text by both model obtain the lower value than the naïve Bayes and CMRM. Therefore, the score of text histogram intersection obtain the lower value; the performance of image retrieval by text query also will obtain the lower value. The comparison of mAP in each object is shown in Table 5.5. Although, our model cannot archive the best performance using text query, the performance of Table 5.5 is indicated that our model can improve mAP of word more than the mAP of pLSA model, such as, “train”, “person”, “boat”, “horse”, “cow”, “dog”, “aeroplane”, “bus”, “bicycle”, “diningtable”, “bird”, “cat”, “motormike”, and “sheep”.

Table 5.5 Performance of image retrieval with a word query where the size of visual word as 1000 visual words

mAP	Naïve Bayes	CMRM	pLSA (z=20)	Two-pLSA (z=20,l=20)
tvmonitor	0.462	0.332	0.329	0.310
train	0.408	0.298	0.108	0.140
person	0.677	0.728	0.627	0.643
boat	0.321	0.241	0.134	0.259
horse	0.270	0.227	0.078	0.099
cow	0.247	0.157	0.034	0.079
bottle	0.242	0.276	0.197	0.121
dog	0.266	0.301	0.170	0.200
aeroplane	0.569	0.382	0.310	0.456
car	0.358	0.319	0.246	0.245
bus	0.402	0.368	0.128	0.161
bicycle	0.346	0.318	0.199	0.240
chair	0.349	0.251	0.234	0.217
diningtable	0.217	0.227	0.101	0.109
pottedplant	0.210	0.188	0.165	0.073
bird	0.365	0.281	0.165	0.281
cat	0.325	0.278	0.213	0.227
motorbike	0.386	0.322	0.125	0.149
sheep	0.483	0.203	0.066	0.173
sofa	0.225	0.271	0.212	0.122
Average mAP:	0.356	0.298	0.192	0.215

5.3.5. Unlabeled Image Retrieval based on automatic image annotation

In this experiment, the performance of image retrieval by searching with query image is evaluated by varying the size of visual vocabulary as 200, 400, 600, and 1000 respectively. The models are compared, including the naïve Bayes, CMRM, pLSA and two-pLSA with the parameters of the different number of latent variables. The unlabeled images in the testing set will be annotated by words using these models. The probabilities of words are used as the text features for image retrieval process. For matching between two text feature, which are estimated by annotation process, the histogram matching is used.

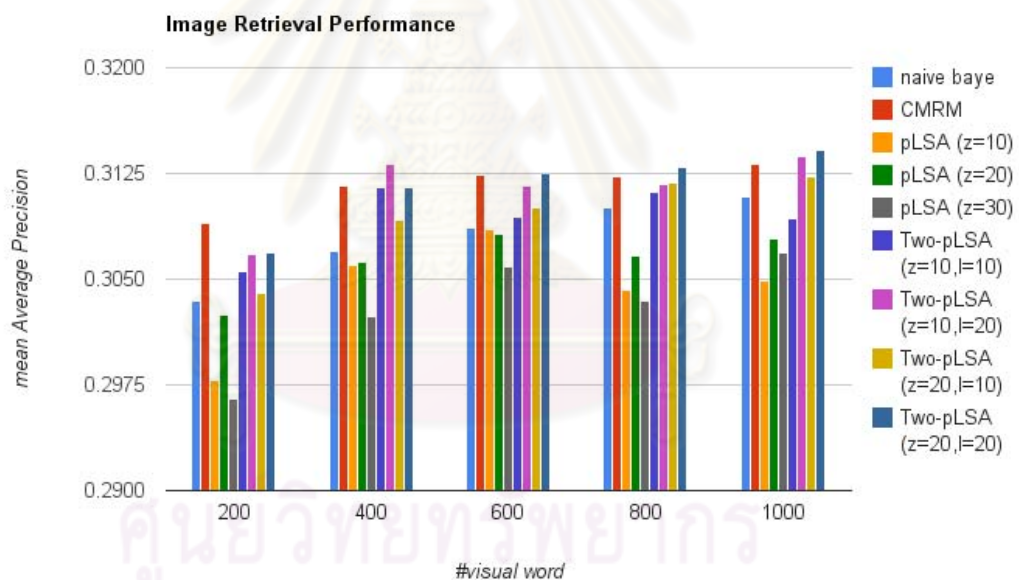


Figure 5.15 mAP Performance of image retrieval using image query for unlabeled image in PASCAL dataset

Figure 5.15, the performance of image retrieval in mAP is shown. The mAP of retrieval in every model is increased when increasing the size of visual vocabulary. The mAP of pLSA increases when the number of latent variable is higher, but decreases when the number of latent variable is less than 20. For the two-pLSA, its mAP is also increase by increasing both the arising number of latent variable z and l .

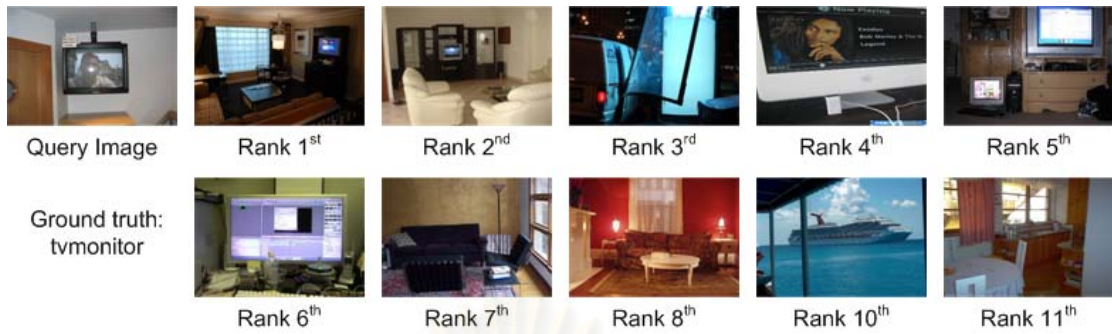


Figure 5.16 Example of image retrieval in the unlabeled image where a ground truth word from PASCAL dataset



Figure 5.17 Example of image retrieval in the unlabeled image where two ground truth word from PASCAL dataset



Figure 5.18 Example of image retrieval in the unlabeled image where three ground truth word from PASCAL dataset

The example of a query image with one word, “tvmonitor”, is shown in Figure 5.16. The correctly retrieved images include of the 1st, 2nd, 4th, 5th and 6th ranks corresponding to one word, “tvmonitor”. It shows that our model can annotated a label correctly, although the small object or part of object. The example of a query image with two words, “person and bicycle”, is shown in Figure 5.17. The correctly retrieved image include of the 3rd and 8th ranks. It shows that the similar shape of object and different view of object, such as bicycle and motorbike, can be retrieved correctly. In case of three ground truth words, “person, bottle and dinningtable”, the retrieved image are shown in Figure 5.18. Using two-pLSA, the correct ranks are in the 2nd, 5th, and 6th. It indicates that the two-pLSA is also suitable for image retrieval, although there are multiple meaning in an images.

Because all of the tested models have estimated the probability of each word, it can be used together. A model is the task of indexing in the database, and then in query process the model can be changed to another model. The experiment results, defining a model is used for indexing process, and another model is served as an annotation model in query process, are shown in Table 5.6.

Table 5.6 The performance in the crossed models where a model as indexing model and another as querying model

Index Query	Naïve Bayes	CMRM	pLSA (z = 20)	Two-pLSA (z=20,L=20)	Average
Naïve Bayes	0.310	0.203	0.214	0.223	0.238
CMRM	0.210	0.313	0.215	0.223	0.240
pLSA (z=20)	0.231	0.237	0.307	0.238	0.253
Two-pLSA (z=20,L=20)	0.205	0.189	0.200	0.314	0.227
Average	0.239	0.236	0.234	0.249	

The results in Table 5.6 observed that the annotation model, which is used in the indexing and querying process, is in the same model that archives the highest performance. In addition, our proposed model for image retrieval is the best model of which mAP is 0.31406. So that, in term of the performance of indexing process, our proposed model achieves the best effectiveness. However the naïve Bayes model usually is a good model when it used in the query process.

5.4 Experiment 4: Performance in MIRFLICKR25000 dataset

Table 5.7 The number of image in each word on MIRFlickr25000 dataset

Words	#images	Words	#images
sky	845	people	330
water	641	city	308
portrait	623	sea	301
night	621	sun	290
nature	596	girl	262
sunset	585	snow	256
cloud	558	food	225
flower	510	bird	218
beach	407	sign	214
landscape	385	car	212
street	383	lake	199
dog	372	building	188
architecture	354	river	175
graffiti	335	baby	167
tree	331	animal	164

In this experiment, we investigate on the performance of image annotation and image retrieval using real images which is taken by photographers and are used in the

image search engine data from Flickr Engine. And it also evaluate in a huge image database that is used in the real world.

5.4.1. Dataset and Simulation Condition

In this experiment, the MIRFlickr25000 dataset is used for evaluate the annotation and retrieval performance. It contains a total of 25000 images with 30 words which were downloaded from social photography sit Flickr.com. The average number of words per image is around 8 words. Table 5.7 shows the most common content-based word. The words can be subdivided in various categories. The most useful tags for research purposes are most likely those that clearly describe the images, preferably with a direct relation to the visual content of the image (e.g. snow, sunset, building, party).

5.4.2. Image Annotation: mean Average Precision Performance and Processing time

Table 5.8 mean Average Precision of image annotation with existing model on MIRFLICKR25000 dataset

Models	N = 1000	N = 5000	N = 10000
Naïve Bayes	0.482	0.538	0.549
CMRM	0.437	0.430	0.429
pLSA (z = 5)	0.443	0.443	0.444
pLSA (z = 10)	0.440	0.438	0.451
pLSA (z = 15)	0.441	0.446	0.444
pLSA (z = 20)	0.449	0.443	out of memory
pLSA (z = 25)	0.441	0.445	out of memory
pLSA (z = 30)	0.449	0.447	out of memory

To measurement the annotation quality of the models, the mean Average Precision in each given label in testing set is computed. The number of visual word is varied as 1000, 5000 and 10000 visual words. Table 5.8 shows the comparison results of existing model on MIRFlickr2005 dataset. The mAP value of naïve model is the best

performance in this dataset. The highest mAP value of the naïve Bayes is 0.54 when the size of visual word is 10000 visual words. In order to increasing the mAP of annotation performance in the naïve Bayes model, the size of visual vocabulary has to increase. In case of CMRM, its mAP has been nearly constant mAP when increasing the size of visual vocabulary. In case of pLSA, the increasing size of visual vocabulary and the increasing number of latent variable do not effect to the annotation performance.

Table 5.9 mean Average Precision of image annotation when the number of vocabulary as 1000 visual words

LZ	5	10	15	20	25
5	0.440	0.439	0.445	0.442	0.437
10	0.444	0.437	0.435	0.442	0.437
15	0.436	0.436	0.441	0.443	0.436
20	0.441	0.442	0.442	0.441	0.437
25	0.441	0.440	0.439	0.440	0.436

Table 5.10 mean Average Precision of image annotation when the number of vocabulary as 5000 visual words

LZ	5	10	15	20	25
5	0.441	0.436	0.437	0.432	0.439
10	0.441	0.434	0.436	0.435	0.436
15	0.454	0.438	0.435	0.434	0.436
20	0.438	0.438	0.438	0.439	0.438
25	0.440	0.439	0.436	0.434	0.434

Table 5.9 and Table 5.10 show the resulting mAP of annotation performance with two-pLSA model when varying the size of visual vocabulary as 1000 and 5000 visual words respectively. These results are indicated that the number of visual and both of latent variable of the two-pLSA model cannot improve efficiency annotation using the

huge images. However, the annotation performances of two-pLSA are nearly equals to the performance of the pLSA.

Table 5.11 processing time (seconds) of annotation with existing model of MIRFlickr25000 dataset

Models	1000	5000	10000
Naïve Bayes	0.001	0.008	0.017
CMRM	0.139	0.584	1.116
pLSA (z = 5)	0.019	0.040	0.080
pLSA (z = 10)	0.026	0.073	0.151
pLSA (z = 15)	0.038	0.108	0.226
pLSA (z = 20)	0.042	0.145	out of memory
pLSA (z = 25)	0.051	0.180	out of memory
pLSA (z = 30)	0.058	0.212	out of memory

However, the complexities of existence models are evaluated by processing time of computing the estimated words using the models as shown in Table 5.11. The naïve model is the lowest complexity model than other models, and the pLSA model spend the most of time for estimated words when increasing the number of latent variable. Figure 5.19 shows the processing time of annotation process when the size of visual vocabulary as 1000 visual words with increasing the number of latent variable. It indicates that the processing time will be increased by the number of latent variable z arises. But the increasing the number of latent variable l does not effect to the complexity of annotation time.

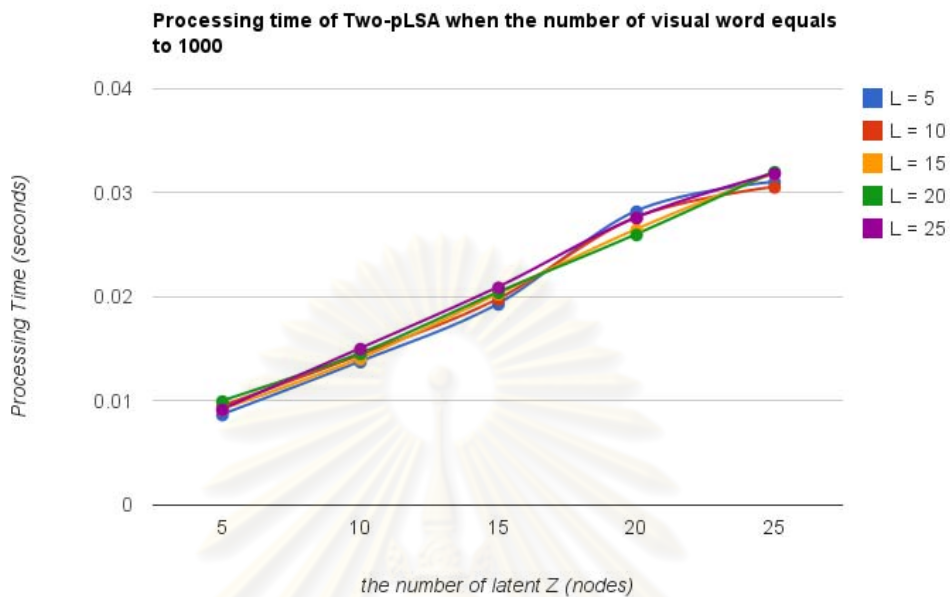


Figure 5.19 The processing time of two-pLSA model when increasing the number of latent variable z and l for image annotation process.

5.4.3. Unlabeled Image Retrieval based on automatic image annotation

Table 5.12 mean Average Precision of image retrieval using a query image with existing model on MIRFlickr25000 dataset

Model	N = 1000	N = 5000	N = 10000
BoF	0.741	0.697	0.704
Naïve Bayes	0.729	0.766	0.717
CMRM	0.702	0.682	0.693
pLSA ($z = 5$)	0.700	0.674	0.697
pLSA ($z = 10$)	0.689	0.698	0.704
pLSA ($z = 15$)	0.698	0.719	0.679

Objective of this experiment is to perform the image retrieval by searching with query image with estimated text using annotation process. By varying the size of visual vocabulary as 1000, 5000 and 1000, the results is compared among BoF matching and

the annotation models, including of naïve Bayes, CMRM, pLSA and two-pLSA model, as shown in Table 5.12, Table 5.13, and Table 5.14.

In Table 5.12, it shows that the naïve Bayes model achieves the best performance in retrieval process. These results indicate that annotation process can improve the performance of image retrieval in term of precisely retrieved images as the mAP of naïve Bayes higher than BoF matching.

Table 5.13 mean Average Precision of image retrieval using a query image with two-pLSA where the size of visual vocabulary as 1000 visual words

LZ	5	10	15	20	25
5	0.640	0.716	0.681	0.640	0.671
10	0.663	0.667	0.665	0.661	0.706
15	0.701	0.683	0.667	0.689	0.673
20	0.687	0.680	0.693	0.704	0.675
25	0.690	0.695	0.709	0.668	0.670

Table 5.14 mean Average Precision of image retrieval using a query image with two-pLSA where the size of visual vocabulary as 5000 visual words

LZ	5	10	15	20	25
5	0.665	0.650	0.670	0.672	0.709
10	0.733	0.684	0.660	0.691	0.678
15	0.689	0.753	0.629	0.648	0.695
20	0.688	0.721	0.683	0.680	0.673
25	0.667	0.681	0.704	0.705	0.728

Considering the increasing number of latent variables in two-pLSA, they effect to little performance of image retrieval. The approximately mAP of two-pLSA when the size of visual vocabulary as 1000, and 5000 visual words are 0.6802 and 0.6867 respectively. However, the complexity of retrieval compared between BoF matching and

estimated text matching depends on the dimensional of indexing in each image. The processing of matching index is evaluated, and is shown in Figure 5.20. The result shows the the retrieved speed of matching with estimated text is faster than the matching with BoF search.

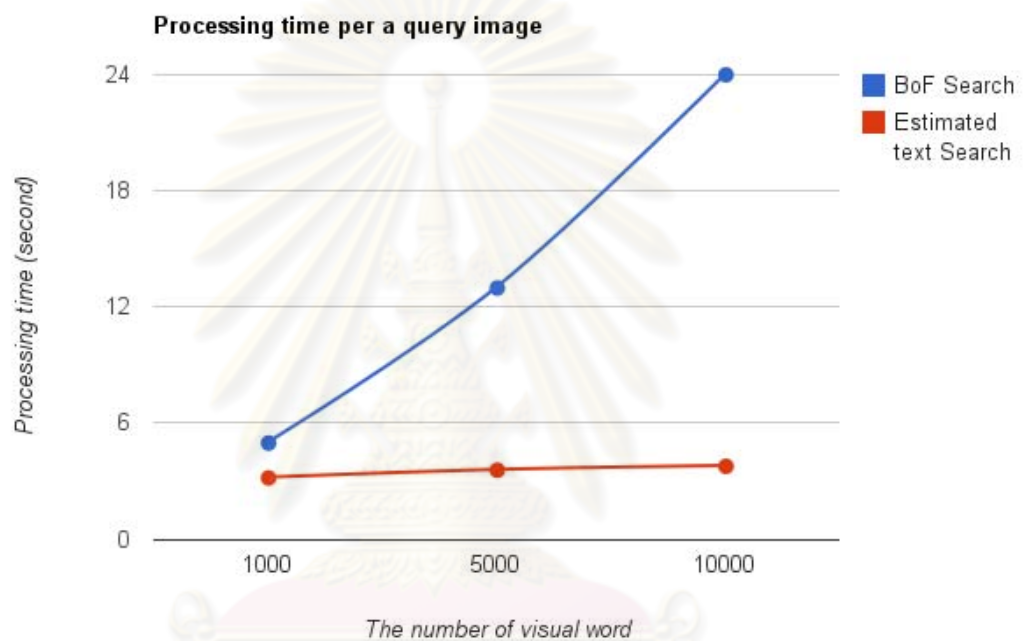


Figure 5.20 The processing time when increasing the size of visual vocabulary compared between the estimated texts search and BoF from a query image

In order to test the performance of indexing with word from automatic image annotation, the process of indexing term from the annotation model serves with various modes of the following: naïve Bayes, CMRM, pLSA when the number of latent z is 15 variables, and two-pLSA model given its number of latent z and equal l to 15 variables. In the process of searching images with an image example, the query image is annotated with a set of specified words from the various modes as well as indexing process. Afterward, the specific meaning of preview with the different model will be switched to indexed and matched in query processes as shown in Table 5.15, and Table 5.16 with the different size of visual vocabulary as 1000 and 5000, respectively.

Table 5.15 The performance in the crossed models of 1000 visual words where a model as indexing model and another as querying model in MIRFlirkc25000 dataset

Index Query	Naïve Bayes	CMRM	pLSA (z = 15)	Two-pLSA (z=15,L=15)	Average
Naïve Bayes	0.732	0.714	0.670	0.659	0.694
CMRM	0.645	0.703	0.684	0.673	0.676
pLSA (z=15)	0.657	0.678	0.691	0.666	0.673
Two-pLSA (z=15,L=15)	0.663	0.696	0.684	0.691	0.6838
Average	0.674	0.698	0.682	0.672	

Table 5.16 The performance in the crossed models of 5000 visual words where a model as indexing model and another as querying model in MIRFlirkc25000 dataset

Index Query	Naïve Bayes	CMRM	pLSA (z = 15)	Two-pLSA (z=15,L=15)	Average
Naïve Bayes	0.704	0.702	0.680	0.654	0.685
CMRM	0.644	0.701	0.671	0.680	0.674
pLSA (z=15)	0.653	0.700	0.699	0.675	0.682
Two-pLSA (z=15,L=15)	0.657	0.699	0.682	0.700	0.6848
Average	0.665	0.701	0.683	0.677	

In order to test the performance of indexing with word from automatic image annotation, the process of indexing term from the annotation model serves with various modes of the following: naïve Bayes, CMRM, pLSA when the number of latent z is 15 variables, and two-pLSA model given its number of latent z and equal l to 15 variables. In the process of searching images with an image example, the query image is annotated with a set of specified words from the various modes as well as indexing process. Afterward, the specific meaning of preview with the different model will be switched to indexed and matched in query processes as shown in Table 5.15, and Table 5.16 with the different size of visual vocabulary as 1000 and 5000, respectively.

Table 5.15 shows the performance of image retrieval is achieved by the different automatic image annotation models where the size of visual vocabulary is 1000 visual words. In term of indexing process, the performance of model are sorted by decreased namely, CMRM, pLSA, naïve Bayes, and two-pLSA model, respectively. It is indicated that the CMRM is the best of indexing process. And in term of query process, the naïve Bayes, Two-pLSA, and CMRM and pLSA are the sorted decreasing performance of image retrieval using an image example. It shows that our proposed model is a model that suitable for image retrieval using query process.

The results obtained from changing the size of visual vocabulary as 5000 visual words are shown in Table 5.16. It reports that the performance of indexing using the automatic models is decreased in CMRM, pLSA, two-pLSA, and naïve Bayes model. And the performance of query using the models are sorted by the most to the least as this following, naïve Bayes, two-pLSA, pLSA and CMRM. Considering the processing time of the annotation process in each model, the CMRM, and two-pLSA usually spend more time. So both of CMRM and pLSA is not suitable for image retrieval system, and the naïve Bayes and two-pLSA model also is the best model in image retrieval system without labeled in any images.

5.5 Experiment 5: Performance in MIRFLICKR25000 dataset when the constructed visual vocabulary by PASCAL 2008 dataset

Because the constructing of new visual vocabulary often spent a lot of times, the performance of visual vocabulary created from dataset itself is compared with the visual vocabulary constructed by another dataset as describing in this experiment. Where the sizes of visual vocabulary are 1000 and 5000 visual words, the dataset, PASCAL 2008, is used for creating the visual vocabulary. There are three tables shown the performance of image annotation, retrieval by text query, and retrieval by image example.

Table 5.17 Performance of image annotation compared between two visual vocabularies

Models	N=1000		N=5000	
	Flickr25000	Pascal 2008	Flickr25000	Pascal 2008
Naïve Bayes	0.482	0.483	0.538	0.539
cmrm	0.437	0.436	0.430	0.437
pLSA (z = 15)	0.441	0.448	0.446	0.454
two-pLSA(z = 15,l=15)	0.441	0.440	0.435	0.438

The performance results of image annotation with the models namely naïve Bayes, CMRM, pLSA, and two-pLSA are shown in Table 5.17. The mAPs of the models compared between the visual vocabulary created by the itself and another visual word created by another dataset are not much different. It indicated that the visual vocabulary, which is constructed by another dataset, can be used to an image representation based on Bag-of-Feature representation. It means that the new visual vocabulary is not necessary, and the existing visual vocabulary can be used at the same. The performance of the image annotation is sorted decreasing namely, naïve bayse, pLSA, two-pLSA and CMRM respectively.

Table 5.18 Performance of image retrieval by a text query compared between two visual vocabularies

Models	N=1000		N=5000	
	Flickr25000	Pascal 2008	Flickr25000	Pascal 2008
Naïve Bayes	0.328	0.333	0.328	0.344
cmrm	0.301	0.296	0.288	0.285
pLSA (z = 15)	0.247	0.244	0.252	0.246
two-pLSA(z = 15,l=15)	0.217	0.220	0.227	0.227

In Table 5.18, the performance of image retrieval by text query is shown. It shows that another visual vocabulary does not affect to the effectiveness of image retrieval. The value of mAP is closely to the visual vocabulary created by its dataset. And

the mAP of model is sorted by decreasing namely, naïve Bayes, CMRM, pLSA and two-pLSA. It also is indicated that the probabilistic model is not suitable for find the images in case of text search.

The performance of image retrieval by image example is shown in Table 5.19. It shows that, in case of BoF search, the size of visual vocabulary is decreased, the mAP of it performance is also decreased. And the mAP of changing visual vocabulary is decrease where the size of visual vocabulary as 1000 visual words. However, using the annotation model in different visual vocabulary can improve the performance of image retrieval by image example where the size of visual vocabulary is increased.

Table 5.19 Performance of image retrieval by an image query compared between two visual vocabularies

Models	N=1000		N=5000	
	Flickr25000	Pascal 2008	Flickr25000	Pascal 2008
BoF	0.741	0.679	0.697	0.698
Naïve Bayes	0.729	0.739	0.766	0.740
cmrm	0.702	0.698	0.682	0.711
pLSA (z = 15)	0.698	0.700	0.719	0.661
two-pLSA(z = 15,l=15)	0.667	0.692	0.629	0.655

All of experimental results in this chapter show that the size of visual vocabulary is the most importance in term of efficient of precision and processing time in image annotation and retrieval process. The increasing number of visual word affects directly the mAP of every performance and also effects to spend the time of annotating unlabeled image. The performance of the probabilistic model depends on the initial parameters setting. Because the initial setting in EM-Algorithm of both pLSA and two-pLSA normally used the randomly setting, so the performance of precision in annotation and retrieval depends on that setting. The number of latent variable has affected little to performance in term of precision performance. Base on the experimental results of MIR

Flickr25000 dataset is found that the naïve Bayes is the best model of image annotation and retrieval.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

CHAPTER VI

CONCLUSIONS

This dissertation presents a model called “two-pLSA model” to solve the problem of automatic annotation and retrieval based on Bag-of-Feature using SIFT feature. The performance was analyzed both in term of the computational complexity and the annotation and retrieval performance. The experiments on comparing existing annotation techniques, namely naïve Bayes, cross media relevance model, and pLSA model with our proposed model only were conducted to obtain an actual efficiency of the annotation model.

From experiment in Section 5.1, an object categorization framework utilizing principal component analysis with bag-of-feature is proposed. The experimental results shows that the PCA-SIFT can reduce the dimensionality of SIFT space with comparable efficiency as baseline technique. And it also can reduce the dimension of SIFT up to around 80% with the same average precision compared to baseline technique.

From experiment in Section 5.2, an approach to detect the parts of interest object is proposed. By firstly removing the noise parts of image using the threshold technique on scaling parameter of SIFT feature, the visual vocabulary then is constructed using k-mean algorithm basing on bag-of-feature model. Afterward the visual words are selected using based on the maximum probability and entropy criteria and choosing entropy criteria using the transiting number of selecting visual words. Experimental results show that the entropy criterion is robust to background clutter and eliminate the parts of irrelevant object.

From experiment in Section 5.3, the strength of the two-pLSA lies on the efficient annotation annotating performance and the faster speed of processing time for

identifying the meaning than the other models. Moreover, in term of image retrieval by text query and image query, the textural features of the unlabeled image are estimated by the models and then used for the index in search process. The performance of image retrieval task using two-pLSA is better than the other models. Based on the experimental results, the size of visual vocabulary is the critical factor that increases the efficiency of the image annotation and retrieval. For pLSA model, the number of latent variable z as a factor affecting the efficiency of the annotation and retrieval is minimal. However, the pLSA model is not the best model for image retrieval task because its mAP of the image retrieval is less than the other models. For the two-pLSA model, the concept of our proposed model is separated into two latent variables. The first latent variable z used for grouping words which often found together, while the second latent variable l used for grouping visual words which often found together at each word. Therefore both the number of latent variable will affect the performance of annotation and retrieval. In case of the image retrieval task, the number of latent variable z will not increase the efficiency of retrieval. By increasing the number of latent variable l , the performance of retrieval is also increased. On the other hand, in the image annotation task, the effect of increasing the number of latent variable z impacts less to increase the efficiency of the annotation than increasing the number of latent variable l .

From the experiment in Section 5.4, the MIRFlickr25000 dataset is used for evaluate the annotation and retrieval performance. The mAP value of naïve model is the best performance in this dataset. The naïve model also is the lowest complexity model than other models, and the pLSA model spend the most of time for estimated words when increasing the number of latent variable. In term of indexing process, the performance of model are sorted by decreased namely, CMRM, pLSA, naïve Bayes, and two-pLSA model, respectively. It is indicated that the CMRM is the best of indexing process. And in term of query process, the naïve Bayes, Two-pLSA, and CMRM and pLSA are the sorted decreasing performance of image retrieval using an image example.

By the experiment in Section 5.5, the performance of visual vocabulary created from dataset itself is compared with the visual vocabulary constructed by another dataset. It indicated that the visual vocabulary, which is constructed by another dataset, can be used to an image representation based on Bag-of-Feature representation. It means that the new visual vocabulary is not necessary, and the existing visual vocabulary can be applied at the same result.

Considering all of the experimental results in the dissertation, they conclude: (1) the size of visual vocabulary is the most importance in term of efficiency precision and processing time in the image annotation and retrieval process. (2) Our designed image retrieval based on text-based indexing using automatic image annotation can support to search various examples such as text or image. (3) Using the automatic image annotation, the text-based indexing for unlabeled image is applied the proposed model to improve the performance in term of robustly change the visual vocabulary. However, the limitation of our proposed model is that the number of latent variable in pLSA and two-pLSA has affected little to performance in term of precision of image annotation and retrieval. In another way, the performances of both pLSA and two-pLSA depend on the initial parameter setting. The normally setting in EM-algorithm of both used the randomly setting is difficult to find optimal solution to obtain the better performance of image annotation. However, the proposed model is promising for image annotation and retrieval system, which are demonstrated by the evaluation of the prototype system in PASCAL 2008 and MIRFlickr25000 dataset.

In future works, our model can be implied the object segmentation algorithm. Based on the grouping visual word of each object in latent variable l can used as a criterion of identify the SIFT feature belonging to the part of interest object. Moreover, the technique of visual vocabulary construction instead of K-mean can improve both performance of image annotation and retrieval such as fuzzy c-mean or another dictionary learning approach.

REFERENCES

- [1] Russell, B. C., and Torralba A., LabelMe: a database and web-based tool for image annotation, Int'l. J. Computer Vision, 77, 1-3, (May 2008): 157-173.
- [2] Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N., Supervised Learning of Semantic Classes for Image Annotation and Retrieval, IEEE Trans. Pattern Analysis and Machine Intelligence, 29, 3, (Mar. 2007): 394-410.
- [3] Hofmann, T., Unsupervised Learning by Probabilistic Latent Semantic Analysis, Machine Learning, 41, 2 (2001): 177-196.
- [4] Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., and Tuytelaars, T., A Thousand Words in a Scene, IEEE Trans. Pattern Analysis and Machine Intelligence, 29, 9, (Sept. 2007): 1575-1589.
- [5] Monay, F., and Gatica-Perez, D., Modeling Semantic Aspects for Cross-Media Image Indexing, IEEE Trans. Pattern Analysis and Machine Intelligence, 29, no. 10, (Oct. 2007): 1802-1817.
- [6] Bosch, A., Zisserman, A., and Munoz, X., Scene Classification Using a Hybrid Generative/Discriminative Approach, IEEE Trans. Pattern Analysis and Machine Intelligence, 30, 4, (Apr. 2008): 712-727.
- [7] Blei, D., and Jordan, M., Modeling Annotated Data, Proc. Int'l Conf. Research and Development in Information Retrieval, (Aug 2003).
- [8] Fei-Fei, L., and Perona, P., A Bayesian Hierarchical Model for learning Natural Scene Categories, Int'l IEEE Conf. Computer Vision and Pattern Recognition. 2, (June 2005): 20-25.

- [9] Webber, M., Welling, M., and Perona, P., Unsupervised Learning of Models for recognition, Proc. European Conf. Computer Vision, (2000).
- [10] Fergus, R., Zisserman, A., Perano, P., Object class recognition by Unsupervised Scale-Invariant Learning, Int'l. IEEE Conf. Computer Vision and Pattern Recognition, (2003).
- [11] Wang, X., Zhang, L., Li, X., and Ma, W., Annotating Images by Mining Large Search Results, IEEE Trans. Pattern Analysis and Machine Intelligence, 30, 11, (Nov. 2008): 1919-1932.
- [12] Viola, P., and Jones, M., "Robust Real-Time Face Detection, Int'l J. Computer Vision, 57, (2004): 137-154.
- [13] Viola, P. and Jones, M., Rapid Object Detection Using a Boosted Cascade of Simple Features, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, (Dec 2001).
- [14] Torralba, A., Murphy, K. P., and Freeman, W. T., Sharing Visual Features for Multiclass and Multiview Object Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 29, 5, (May. 2007): 854-869.
- [15] Sivic, J., and Zisserman, A., Video Google: A text retrieval approach to object matching in Videos, Proc. IEEE Int'l Conf. Computer Vision, (2003).
- [16] Mikolajczyk, K., and Schmid, C., A Performance Evaluation of local descriptors, IEEE Trans. Pattern Analysis and Machine Intelligence, 27, 10, (Oct. 2005).
- [17] Lowe, D.G., Distinctive Image Feature from Scale-Invariant Keypoints, Int'l J. Computer Vision, 60, (2004): 91-110.

- [18] Willamowski, J., Arregui, D., Csurka, G., Dance, C.R., and Fan, L., Categorizing Nine Visual Classes using Local Appearance Descriptors, Int'l workshop Learning for adaptable Visual Systems, (2004).
- [19] Jiang, Y.G., Ngo, C. W., and Yung, J., Toward Optimal Bag-of-features for Object Categorization and Semantic Video Retrieval, Proc. ACM Int'l. Conf. Image and Video Retrieval, (2007).
- [20] Marszalek, M., and Schmid, C., Spatial weighting for bag-of-features, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, (2006).
- [21] Nowak, E., Jurie, F., and Triggs, B., Sampling Strategies for Bag-of-Features Image classification, Proc. European Conf. Computer Vision, (2006).
- [22] Nister, D. and Stewenius, H., Scalable Recognition with a Vocabulary Tree, Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, (2006).
- [23] Xu, F., Zhang, L., Zhang, Y-J., and Ma, W-Y. Salient Feature Selection for Visual Concept Learning, Advance in Multimedia Information Processing -PCM 2005, (2005): 617-628.
- [24] Dorko, G., and Schmid, C., Selection of Scale-Invariant Parts for Object Class Recognition, Int'l IEEE Conf. Computer Vision, 1, (Oct 2003): 113-117.
- [25] Zhao, Z., and Elagmml, A., A Statistically Selected Part-Based Probabilistic Model for Object Recognition, Int'l. Workshop Intelligent Computing in Pattern Analysis/Synthesis, (Aug 2006).
- [26] Duda, R., Hart, P., and Stock, D., Pattern Classification. John Wiley and Sons, 2001.
- [27] Bishop, C. M., Pattern Recognition and Machine Learning. Springer, 2006.
- [28] M. Eeringham and L. Van-Gool, C. K. Williams, J. Winn, and A. Zisserman, The Pascal Visual Object Classes Challenge 2008 (VOC 2008), [online]

Available from: <http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html>.

[2009/06/30].

- [29] Niblack, W., Barber, R., Equitz, W., Flicker, M., Glasman, E., Petkovic, D., Yanker, P., and Faloutsos, C., The QBIC project: query images by content using color, texture, and shape, Proc. of the SPIE Conf. Storage and Retrieval for Image and Video Databases, (Feb. 1993).
- [30] Jain, A. and Vailaya, A., Image Retrieval using Color and Shape, Pattern Recognition J., 29, (Aug. 1996): 1233-1244.
- [31] Pentland, A., Picard, R., and Sclaroff, S., Photobook: Content-Based Manipulation of Image Databases, Int'l J. Computer Vision, 18, 3, (June 1996.): 233-254
- [32] J. Smith and S. Chang, VisualSEEK: A Fully Automated Content Based Image Query System, ACM Multimedia, pp. 87-98, 1996.
- [33] Carson, C., Belongie, S., Greenspan, H., and Malik, J., Blobworld: Image segmentation using expectation-maximization and its application to image querying, IEEE Trans. Pattern Analysis and Machine Intelligence, 24, 8, (Aug 2002): 1026-1038.
- [34] Dempster, A. P., Laird, N. M., and Rubin, D. B., Maximum likelihood from incomplete data via the EM algorithm, J. Royal Statist. Soc. B, 39, (1977): 1-38.
- [35] Agarwal, S., Awan, A., and Roth, D., Learning to Detect Objects in Images via a Sparse, Part-Based Representation, IEEE Trans. Pattern Analysis and Machine Intelligence, 26, 11, (Nov 2004): 1475-1490.
- [36] Smeaton, A. F., Over, P., and Kraaij, W., Evaluation campaigns and TRECVID, MIR '06. ACM Press, (2006).

- [37] Duygula, P., Barnard, K., De Freitas, N., and Forsyth, D., Object Recognition as Machine Translation: Learning a lexicon for a fixed image vocabulary, Proc. European Conf. Computer Vision, (2002): 97-112.
- [38] Zhang, R., Zhang, Z., Li, M., Ma, W-Y., Zhang, H-J., A probabilistic semantic model for image annotation and multi-model image retrieval, Multimedia Systems, 12, (2006): 27-33.
- [39] Jeon, J., Lavrenko, V. and Manmatha, R., Automatic Image Annotation and Retrieval using Cross-Media Relevance Models, Proc. Int'l Conf. Research and Development in Information Retrieval (SIGIR), (Aug. 2003).
- [40] Larenko, V., Manmatha, R., and Jeon, J., A Model for Learning the Semantics of Pictures, Proc. of Advances in Neural Information Processing Systems, (Dec. 2003).
- [41] Feng, S.L., Manmatha, R., and Lavrenko, V., Multiple Bernoulli Relevance Models for Image and Video Annotation, Int'l Conf. Computer Vision and Recognition, 2 (July 2004): II-1002-II-1009.
- [42] Huang, P., Bu, J., Chen, C., Liu, K., and Qiu, G., Improve Image Annotation by combining Multiple Models, Int'l IEEE Conf. Signal-Image Technologies and Internet-based system, (2008).
- [43] Marukatat, S., A New Model for Image Annotation, Advances in Knowledge Discovery and Data Mining, 5012, (2008).
- [44] Pham, T-T., Maillot, N. E., Lim, J-H., Chevallet, J-P., Latent Semantic Fusion Model for Image Retrieval and Annotation, Proc. ACM Conf. Information and Knowledge Management, (2007): 439-444.

- [45] Jin, R., Chai, J. Y., and Si, L., Effective Automatic Image Annotation Via A coherent Language Model and Active Learning, Proc. ACM Int'l Conf. Multimedia, (2004): 892-899.
- [46] Chang, E., Goh, K., Sychay, G., Wu, G., CBSA: Content-based Soft Annotation for multimodal Image Retrieval Using Bayes Point Machines, IEEE Trans. Circuits and Systems for Video Technology, 13, 1, (2003).
- [47] Li, B.T., Goh, K., and Chang, E., Confident-Based Dynamic Ensemble for Image Annotation and Semantic Annotation and Semantic Discovery, Proc. ACM Int'l Conf. Multimedia, (2003): 195-206.
- [48] Lui, R., Wang, Y., Baba, T., Muasumoto, D., and Nagata, S., SVM-based active feedback in image retrieval using clustering and unlabeled data, J. Pattern Recognition, 41, (2008): 2645-2655.
- [49] Zhou, X. S., Huang, T.S., Relevance Feedback in Image Retrieval: a comprehensive review, Multimedia System, 8, 2, (2003): 536-544.
- [50] Tao, D., Tang, X., Li, X., and Rui, Y., Direct kernel biased discriminant analysis: a new content-based image retrieval relevance feedback algorithm, IEEE Trans. Multimedia, 8, 4, (2006.): 716-726.
- [51] Cox, I. J., Miller, M.L., Minka, T.P., Papathomus, T.V., and Yianilos, P.N., The Bayesian image retrieval system, PicHunter: theory implementation, and psychophysical experiment, IEEE. Trans. Image Processing, 9, 1, (2000).
- [52] Tong, S., and Chang, E., Support Vector Machine active learning for image retrieval, Proc. Int'l ACM Multimedia, (2001)107-118.
- [53] Mori, Y., Takahashi, H., and Oka, R., Image-to-word transformation based on dividing and vector quantizing images with words, In Proc. of Int'l

Workshop on Multimedia Intelligent Storage and Retrieval Management,
(Oct 1999).



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

VITAE

Nattachai Watcharapinchai received the B.Eng degree with the 2nd class honor in electronic engineering from King Mongkut's Institute of Technology Ladkrabang, Thailand, in 2002, and the M.Eng degree in electrical engineering from Chulalongkorn University, Thailand, in 2005. He is currently working toward the Ph.D degree at department of electrical engineering at Chulalongkorn University, Thailand. His research interests include computer vision, machine learning and multimedia retrieval.

List of conference publications

- Nattachai Watcharapinchai, Supavadee Aramvith, Supakorn Siddhichai, Sanparith Marukatat, Dimensionality Reduction of SIFT using PCA for Object Categorization, Electrical Engineering Conference 31th (EECON'29), Nakornnayok, Thailand, October 2008 (Thai Conference).
- Nattachai Watcharapinchai, Supavadee Aramvith, Supakorn Siddhichai, Sanparith Marukatat, Dimensionality Reduction of SIFT using PCA for Object Categorization, The International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2008), Dec 8-11 2008, Bangkok, Thailand.
- Nattachai Watcharapinchai, Supavadee Aramvith, Supakorn Siddhichai, Sanparith Marukatat, Selecting Parts of object using the Maximum Probability and Entropy Criteria for Object Detection, 2008 International Workshop on Smart Info-Media Systems (SISB 2008), Dec 8 2008, Bangkok, Thailand
- Nontarat Bumrungrat, Nattachai Watcharapinchai, Supavadee Aramvith, Thanarat Chalidabhongse, Real-time Face Cateloger using Cooperative Cameras, International Symposium on Multimedia and Communication Technology 2008, ISMAC 2008, Bangkok, Thailand.

- Nattachai Watcharapinchai, Supavadee Aramvith, and Supakorn Siddhichai, Comparative Analysis of Probabilistic Models for Image Annotation, 2010 International Symposium on Multimedia and Communication Technology (ISMAC 2010), Sep 8-9 2010, Manila, Philippines.
- Nattachai Watcharapinchai, Supavadee Aramvith, and Supakorn Siddhichai, Two-Probabilistic Latent Semantic Analysis for Image Annotation, 2010 International Workshop on Smart Info-Media Systems in Asia (SISA 2010), Sep 8-9 2010, Manila, Philippines.
- Nattachai Watcharapinchai, Supavadee Aramvith, and Supakorn Siddhichai, Two-Probabilistic Latent Semantic Analysis for Image Annotation and Retrieval, The 2nd International Workshop on Video Event Categorization, Tagging and Retrieval, Nov 8-12 2010, Queenstown, New Zealand.



ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย