Combler le fossé sémantique en recherche d’image via la transformation des concepts sémantique en une représentation visuelle

Presented by:

**Oussama Aiadi**

Committee members

<table>
<thead>
<tr>
<th>Role</th>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>President</td>
<td>Mohamed-Khireddine Kholladi</td>
<td>Prof- El Oued University</td>
</tr>
<tr>
<td>Examiner</td>
<td>Okba Kazar</td>
<td>Prof- Biskra University</td>
</tr>
<tr>
<td>Examiner</td>
<td>Ahmed Korichi</td>
<td>MCA- Ouargla University</td>
</tr>
<tr>
<td>Examiner</td>
<td>Fatima Zohra Laallam</td>
<td>MCA- Ouargla University</td>
</tr>
<tr>
<td>Supervisor</td>
<td>Mohammed Lamine Kherfi</td>
<td>MCA- Ouargla University</td>
</tr>
</tbody>
</table>

Academic Year: 2016 – 2017
Dedication

This thesis would be incomplete without mentioning the support of my beloved and dear parents. I dedicate this thesis to my parents for their endless love, infinite support and great encouragements throughout my life.

I also dedicate this thesis to the lights of my life: my brothers and sisters. I appreciate their support and interest.

I dedicate this thesis to all my family,

do all my friends

To all those who were giving me any kind of support.
Acknowledgments

First and foremost, praise and thanks to ALLAH the sacred and the mighty for giving me the ability and helping me in accomplishing this thesis.

Then, I would like to express my sincerest gratitude to my supervisor Dr. Mohammed Lamine Kherfi for his continuous support, help, guidance, motivation and patience throughout this thesis. I deeply appreciate his immense knowledge, constructive feedback, assistance, which considerably improved my experience.

My sincere thanks also go to Dr. Ahmed Korichi who provided me the opportunity to join the PhD study. I acknowledge him for his encouragements, motivation, advice and insight.

I also greatly appreciate the collaboration and the cooperation of my dear friend Bilal Khaldi. I deeply acknowledge him for their assistance, discussions and efforts.

A very special thanks go to the honorable members of my dissertation committee who accepted to examine and evaluate this thesis namely, Prof. Mohamed-Khireddine Kholladi from El Oued university; Prof. Okba Kazar. and Dr. Djeffal Abdelhamid from the University of Biskra; Dr. Ahmed Korichi and Dr. Laallam Fatima Zohra from the university of Ouargla.

I gratefully acknowledge all the professors within the Department of computer science and information technology, especially whom teaching me during my curriculum. I am also very grateful to all the members of our group of research and to all my colleagues.

Last but not least, I would like to thank with gratitude the two Prof. Cheriet and Dr. Hadjem for their advices, support and assistance, which were very useful in completing this work.
Abstract

Semantic gap between low-level image features and high-level user semantics is one of the common issues which degrade the performance of Content-Based Image Retrieval (CBIR). The main concern of this thesis is to bridge this gap, and thus to improve the performance of retrieval systems. The proposed approach consists in a process of supervised learning which aims to learn a visual model for each semantic concept. For a given concept (e.g., Horse), we associate the concept with a set of images (i.e., training images), where the web search engines are used for providing those images. Then, so as to faithfully learn the concept, we automatically prune the training images from outliers. Afterwards, we automatically identify the visual appearances within the training set, and then a Gaussian Mixture Model (GMM) was used to model the concept. During retrieval, user can naturally formulate his query using a textual query. The learnt visual model (i.e., GMM), corresponding to the query concept, is used to detect the pertinent images.

The proposed approach presents a multitude of advantages. First, it allows to use a textual query in retrieving images from collections where the text is entirely absent. Second, it takes into account the intra-variation within training images and the presence of outliers as well. In addition, it is fully automatic, as no human intervention is required. Furthermore, it is unlimited to any pre-defined concept, as training images are automatically downloaded from the web. Because we formulate the retrieval as a probabilistic supervised classification problem, we apply the proposed approach to the task of date fruit recognition.

Experimental results showed that the proposed approach has achieved a high retrieval precision as well as a high recognition accuracy. Furthermore, both retrieval and recognition were performed with a high speed. Moreover, our approach has proven its strength against several methods from the state-of-the-art.

Keywords: Content-Based Image Retrieval (CBIR), semantic gap, Gaussian Mixture Model (GMM), Supervised learning, Outlier detection, date fruit recognition.
Résumé

Le fossé sémantique est l’un des problématiques qui dégrade la performance des systèmes de recherche d’image par le contenu (CBIR). L’objectif de cette thèse est de combler le fossé sémantique afin d’améliorer la performance de CBIR. L’approche proposée consiste à un processus d’apprentissage supervisé qui vise à apprendre un modèle visuelle pour chaque concept sémantique. Nous avons utilisé les images du web afin d’accomplir l’apprentissage, cela permet d’être ne pas limité par un concept prédéfini. Puis, on a éliminé les images aberrantes afin d’apprendre correctement les différents concepts sémantiques. Les apparences visuelles au sein des images qui présentent le même concept sont automatiquement identifiées. Par ailleurs, un model du mélange de Gaussienne (GMM) est utilisé pour modeler chaque concept. Durant la recherche, l’utilisateur peut fournir une requête textuelle. Les modèles visuelles extraites lors de l’apprentissage sont utilisées ultérieurement pour détecter les images pertinentes à la requête.

L’approche proposée a plusieurs avantages. Premièrement, elle permet d’utiliser une requête textuelle pour chercher dans des collections où le texte est entièrement absent. Deuxièmement, elle prend en considération l’intra-variation des images ainsi que l’existence des images aberrante. En outre, elle est purement automatique. Nous avons appliqué notre approche, qui formule le problème de recherche comme un problème de classification supervisée, pour la classification des dattes.

Les résultats expérimentaux montrent que notre approche a atteint une haute précision ainsi qu’un taux de reconnaissance élevé. En outre, le processus de recherche et de reconnaissance sont effectuées en grande vitesse. En comparant l’approche proposé avec d’autres méthodes, elle a montré une performance élevé.

Mot-clés : Recherche d’image par le contenu, fossé sémantique, Mixture de la Gaussiens, Apprentissage supervisé, détection des outliers, reconnaissance des dattes.
الملخص

تعتبر الهواء المعنية بين مفردات المستعمل و الخصائص البصرة للصورة أحد أكبر العوائق بالنسبة لعمل محركات البحث عن الصورة عن طريق المحتوى. إن الهدف الرئيسي لهذه الأطروحة هو جسر هذه الهوة، وبالتالي تحسين عمل محركات البحث عن طريق المحتوى. إن الطريقة المقترحة لحل هذا المشكل تقوم أساساً على تقنيات التعليم المراقب للاطلاع. لقد قمنا بجلب الصور المخصصة للتعليم من الويب، ما يعني أن طريقتنا غير محدودة بأي مفردة معروفة مسبقاً بعد ذلك فمنا يتنبأ هذه الصور من الشواب، وذلك بغرض تعلم المفردات بشكل صحيح. بعد ذلك فمنا يتفحص المظاهر البصرية الموافقة لكل مفردة، ثم استعملنا نموذج غوس من أجل تدمج كل مفردة بشكل منفصل. أثناء البحث يستطيع المستعمل أن يبحث عمداً يريد و بحرية تامة و ذلك عن طريق تشكيل طلبه عن طريق المفردات. ثم نستعمل النماذج البصرية من أجل تحديد الصور الموافقة للمطلوب.

تقدم الطريقة المقترحة عديد الميزات، فهي تأخذ في الاعتبار الاختلاف الداخلي ضمن الصور بالإضافة إلى نزع الشواب المعيبة أيضاً، كما أنها تعمل بشكل تام بدون أي تدخل من طرف الإنسان. كما تجد الإشارة إلى أن طريقتنا استعملت أيضاً لحل مشكلة الفرز الألاني للحالة.

أظهرت النتائج التجريبية أن طريقتنا بلغت درجة عالية من الدقة في كل من البحث و الفرز. يضاف إلى هذا أن كلتا العمليتين تمتا بسرعة عالية، و أن النتائج أظهرت تفوق الطريقة المقترحة على طرق أخرى.

الكلمات المفتاحية: البحث عن الصورة عن طريق المحتوى، الهواء المعنية، نموذج غوس، التعليم المراقب، عزل الحالات الدخيلة، التعرف على التمور.
# Table of Contents

ACKNOWLEDGMENTS ...........................................................................................................III

ABSTRACT .............................................................................................................................. IV

TABLE OF CONTENTS ........................................................................................................... VII

LIST OF FIGURES ................................................................................................................. XII

LIST OF TABLES .................................................................................................................... XVI

CHAPTER I. INTRODUCTION

I.1 INTRODUCTION .................................................................................................................. 1

I.2 PROBLEMATIC .................................................................................................................... 2

I.3 OVERVIEW ON THE RELATED WORK ............................................................................... 4

I.4 MOTIVATION ....................................................................................................................... 5

I.5 CONTRIBUTIONS ................................................................................................................ 6

I.6 THESIS STRUCTURE .......................................................................................................... 8

CHAPTER II. CONTENT-BASED IMAGE RETRIEVAL (CBIR): PRINCIPLES, FUNDAMENTAL NOTIONS, APPLICATIONS AND COMMON ISSUES

II.1 INTRODUCTION .................................................................................................................. 10

II.2 PRINCIPLE OF CBIR SYSTEM ........................................................................................ 11

II.3 MAJOR COMPONENTS OF A CBIR SYSTEM ...................................................................... 12

II.3.1 Features extraction ........................................................................................................ 12

II.3.2 Query Formulation ........................................................................................................ 13

II.3.2.1 Query by providing the values of the visual features ................................................. 15

II.3.2.2 Query by image(s) example(s) .................................................................................. 15

II.3.2.3 Query by sketch .......................................................................................................... 16

II.3.3 Similarity measures ........................................................................................................ 16

II.3.3.1 Minkowski distances ................................................................................................. 17

II.3.3.2 Kullback-Leibler divergence: .................................................................................... 17

II.3.3.3 Jeffery divergence ....................................................................................................... 18

II.3.3.4 Relative deviation ...................................................................................................... 18
II.3.3.5 Histogram intersection ................................................................. 18
II.3.3.6 Quadratic distance ........................................................................ 19
II.3.3.7 Earth Movers Distance (EMD)....................................................... 19
II.3.4 Examples of CBIR systems ................................................................. 20
II.3.4.1 QBIC .......................................................................................... 20
II.3.4.2 BlobWorld .................................................................................. 20
II.3.4.3 SIMPLIcity .................................................................................. 21
II.3.4.4 Atlas WISE ................................................................................ 21
II.3.4.5 ImageScape ................................................................................ 21
II.3.4.6 PicHunter .................................................................................... 21
II.3.5 Application fields of CBIR ................................................................. 22
II.3.6 Common issues of CBIR ................................................................. 23
II.3.6.1 Feature extraction and selection ................................................... 23
II.3.6.2 Page zero problem .................................................................. 24
II.3.6.3 Semantic gap ............................................................................ 24
II.4 CONCLUSION ...................................................................................... 26

CHAPTER III. BRIDGING THE SEMANTIC GAP IN CBIR: A STATE OF THE
ART REVIEW

III.1 INTRODUCTION ................................................................................ 28
III.2 CATEGORIZATION OF METHODS AIMING TO BRIDGE THE SEMANTIC GAP ........ 28
III.3 AUTOMATIC IMAGE ANNOTATION APPROACH ........................................... 29
III.3.1 Automatic image annotation based on supervised learning ................. 30
III.3.1.1 A baseline method for auto-image annotation ................................ 30
III.3.1.2 Automatic image annotation using semantic relevance .................. 31
III.3.1.3 Textual query of personal photos facilitated by large-scale web data... 33
III.3.1.4 Improving automatic image annotation with Google semantic link .... 34
III.3.1.5 Multi-label learning for image categorization ............................... 36
III.3.2 Automatic image annotation based on unsupervised learning ...........................................36
  III.3.2.1 Co-occurrence model .................................................................................................37
  III.3.2.2 Machine Translation Model (TM) .............................................................................37
  III.3.2.3 Cross Media Relevance Model (CMRM) .................................................................38
  III.3.2.4 Continuous Relevance Model (CRM) .......................................................................39
  III.3.2.5 Methods incorporating semantic relatedness ............................................................39
III.3.3 Automatic image annotation based on semi-supervised learning ...................................40
  III.3.3.1 Annotation refinement and completion for image retrieval method ..........................40
III.4 REGION-BASED APPROACH ..............................................................................................41
  III.4.1 Image retrieval based on attention-driven salient edges and regions extraction ..............41
  III.4.2 Adaptive region matching for region-based image retrieval .......................................42
III.5 RELEVANCE FEEDBACK APPROACH ..............................................................................43
  III.5.1 Relevance feedback for image retrieval by Combining SVM and GM .......................44
III.6 VISUAL OBJECT RECOGNITION .......................................................................................45
  III.6.1 Automatic date fruit recognition ..................................................................................45
    III.6.1.1 Methods for automatic date fruit grading .................................................................46
    III.6.1.2 Methods for automatic date variety recognition ......................................................48
III.7 CONCLUSION ......................................................................................................................49

CHAPTER IV. A SUPERVISED PROBABILISTIC-BASED APPROACH TO FIGHT AGAINST THE SEMANTIC GAP IN CBIR

  IV.1 INTRODUCTION ................................................................................................................50
  IV.2 OVERVIEW ON THE PROPOSED APPROACH ...............................................................51
  IV.3 STEPS OF THE PROPOSED APPROACH FOR BRIDGING THE SEMANTIC GAP .............54
    IV.3.1 Training images gathering ..........................................................................................54
    IV.3.2 Features extraction .....................................................................................................54
      IV.3.2.1 Color features .........................................................................................................55
      IV.3.2.2 Texture features .....................................................................................................57
      IV.3.2.3 Shape features .......................................................................................................61
| IV.3.2.4 Keypoint-based features | .............................................................. 65 |
| IV.3.3 Dimensionality reduction | .............................................................. 68 |
| IV.3.4 Outlier detection | ........................................................................ 70 |
| IV.3.5 Detecting the optimal number of appearances | ............................................ 72 |
| IV.3.5.1 Calinski-Harabasz index | ................................................................ 73 |
| IV.3.5.2 Davies-Bouldin index | ..................................................................... 74 |
| IV.3.6 Normality test using Mardia test | ................................................................. 76 |
| IV.3.7 Modeling of semantic concepts using Gaussian Mixture Models | ............... 77 |
| IV.3.8 Image retrieval | ........................................................................ 79 |
| IV.4 APPLICATION OF THE PROPOSED APPROACH TO THE TASK OF PATTERN RECOGNITION | ............................................................... 80 |
| IV.5 CONCLUSION | ........................................................................ 81 |

CHAPTER V. A COMPUTATIONAL MULTI-TASK SYSTEM FOR IMAGE RETRIEVAL AND PATTERN RECOGNITION

| V.1 INTRODUCTION | .............................................................. 82 |
| V.2 EXPERIMENTAL SETUP | ........................................................................... 82 |
| V.2.1 Datasets | .................................................................................. 82 |
| V.2.1.1 Datasets used to test retrieval | ................................................................. 82 |
| V.2.1.2 Dataset used to test recognition | ..................................................................... 85 |
| V.2.2 Performance metrics | .................................................................................. 87 |
| V.2.3 Parameters tuning | .................................................................................. 88 |
| V.3 EXPERIMENTAL RESULTS | ........................................................................... 88 |
| V.3.1 Retrieval | .................................................................................. 88 |
| V.3.1.1 Retrieval results on the web dataset | ................................................................. 88 |
| V.3.1.2 Retrieval results on COIL-100 | ........................................................................ 97 |
| V.3.1.3 Comparison with the state of the art methods using IAPR TC-12 dataset | 99 |
| V.3.2 Recognition | .................................................................................. 101 |
| V.3.2.1 First experiment: measuring the overall performance | ........................................... 102 |
| V.3.2.2 Second experiment: measuring the recognition speed | ........................................... 103 |
| V.3.2.3 Third experiment: comparison with state of the art methods | ....................................... 104 |
V.3.2.4 Fourth experiment: test the recognition strength of different combinations .................................................................................................................................................. 106

V.4 CONCLUSION ........................................................................................................................................................................................................................................... 107

GENERAL CONCLUSION AND PERSPECTIVES ................................................................................................................................. 109

BIBLIOGRAPHY ........................................................................................................................................................................................................................................... 112
List of Figures

**Figure 1.** Top retrieved images by ‘Google Images’ by submitting the image at top as a query ................................................................. 3

**Figure 2.** Architecture of a typical CBIR system ......................................................... 11

**Figure 3.** Top retrieved images by Google Images for the query ‘apple’ ...................... 14

**Figure 4.** Top retrieved images by Google Image by submitting the image at top as a query ............................................................................................................. 15

**Figure 5.** 3D PET/CT image retrieval system (Fathabad and Balafar 2012) ............... 22

**Figure 6.** Both visual and textual descriptions for an image ........................................... 25

**Figure 7.** Images representing the semantic concept ‘lunch’ ........................................ 25

**Figure 8.** Semantic gap between low-level image features and high-level semantics ...................................................................................................................... 26

**Figure 9.** The proposed framework in (Liu, Xu et al. 2011) .......................................... 33

**Figure 10.** The process of blobs generation from training images (Duygulu, Barnard et al. 2002) ................................................................................................. 38

**Figure 11.** Semantic similarity between certain concepts (Gong, Li et al. 2010)........ 40

**Figure 12.** Tag completion for auto-annotation method (Wu, Jin et al. 2013) ............. 40

**Figure 13.** Flowchart of a region-based for image retrieval (Feng, Xu et al. 2010) 42

**Figure 14.** A computer-vision-based system for date grading (Pourdarbani, Ghassemzadeh et al. 2015) ......................................................................................... 46

**Figure 15.** Date samples in different hardness degree (Manickavasagan, Al-Mezeini et al. 2014) ........................................................................................................... 47

**Figure 16.** Two types of defects, bird flicks and bruises (Al-Ouali 2011) ............. 48
FIGURE 17. THE FLOWCHART OF THE PROPOSED APPROACH .......................................................... 53

FIGURE 18. TOP GATHERED IMAGES FOR CERTAIN SEMANTIC CONCEPTS ........................................ 55

FIGURE 19. RGB-BASED COLOR HISTOGRAM FOR THREE IMAGES .................................................. 56

FIGURE 20. EXAMPLES FOR TEXTURE IMAGES FROM THE Outex DATASET (Ojala, Maenpaa et al. 2002) ............................................. 57

FIGURE 21. EXAMPLE OF LBP CALCULATION .................................................................................. 59

FIGURE 22. DIFFERENT STEPS TO CALCULATE THE WLD (Chen, Shan et al. 2010) ...................... 60

FIGURE 23. EXAMPLE OF WLD HISTOGRAMS FOR THREE DIFFERENT IMAGES ....................... 61

FIGURE 24. Hu MOMENTS EXTRACTED FROM AN IMAGE AND THE SAME IMAGE ROTATED WITH 180 DEGREE ......................................................................................... 63

FIGURE 25. MAJOR AND MINOR AXIS LENGTH FOR A DATE SAMPLE ............................................. 65

FIGURE 26. CALCULATION OF DIFFERENCE OF GAUSSIAN (DoG) (Lowe 2004) ......................... 66

FIGURE 27. DETECTION OF MINIMA AND MAXIMA (Lowe 2004) .................................................. 66

FIGURE 28. IMAGE REPRESENTATION USING BAG OF WORDS (BoW) ........................................... 68

FIGURE 29. IMAGES FROM THE CLASS 'POTATO', WHICH ARE DETECTED AS OUTLIERS USING PLDOF .................................................................................................................. 72

FIGURE 30. DETECTED VISUAL APPEARANCES OF THE SEMANTIC CONCEPT 'POTATO' .......... 75

FIGURE 31. REPRESENTATIVE SAMPLES FROM THE OUR DATASET ............................................. 83

FIGURE 32. TYPICAL SAMPLES FROM THE IAPR TC-12 DATASET .................................................. 84

FIGURE 33. TYPICAL SAMPLES FROM THE COIL-100 DATASET .................................................... 84

FIGURE 34. TYPICAL SAMPLES FROM DIFFERENT VARIETIES ....................................................... 85

FIGURE 35. TYPICAL VARIETIES WITH A LARGE INTRA-VARIATION. SAMPLES WITHIN THOSE VARIETIES ARE DIFFERENT IN TERMS OF MATURITY STAGES, HARDNESS DEGREE AND SHAPE (A) Bayd Hmam (B) Hamraya and (C) Tinisin ..................................................... 86
Figure 36. Typical varieties with a small inter-variation, we can notice the high visual resemblance (a) Three samples from Bayd Hamam and three others from Litima, (b) Three samples from Dfar lgat and three others from Degla Bayda and (c) Three samples from Tarmount and three others from Tantbucht......86

Figure 37. Some outlier samples from different varieties ........................................86

Figure 38. Precision per class yielded by the first subset of feature combinations89

Figure 39. Precision per class yielded by the second subset of feature combinations .................................................................90

Figure 40. Precision per class yielded by the third subset of feature combinations .................................................................90

Figure 41. Recall per class yielded by the first subset of feature combinations ...91

Figure 42. Recall per class yielded by the second subset of feature combinations92

Figure 43. Recall per class yielded by the third subset of feature combinations ..92

Figure 44. Precision-Scope curve of the proposed approach .................................................93

Figure 45. Top five retrieved images for the first subset of classes .....................94

Figure 46. Top five retrieved images for the second subset of classes...............94

Figure 47. Top five retrieved images for the third subset of classes .................95

Figure 48. Top five retrieved images for the fourth subset of classes.............95

Figure 49. Retrieval strength of the different combinations on the COIL-100......97

Figure 50. Top five retrieved images for the first five classes .........................98

Figure 51. Precision-scope curve for the COIL-100 dataset .............................98

Figure 52. Top five retrieved images from the IAPR TC-12 (first subset of concepts) .............................................................................100

Figure 53. Top five retrieved images from the IAPR TC-12 (second subset of concepts) .............................................................................101
FIGURE 54. THE ACCURACY PER VARIETY YIELDED BY EACH METHOD .................. 106

FIGURE 55. THE ACCURACY PER VARIETY YIELDED BY EACH COMBINATION. ..................... 107
List of Tables

**Table 1.** State of the art methods for date varieties recognition .................................48

**Table 2.** Summarizes the values assigned to the different parameters .........................88

**Table 3.** Average precision and recall for all the combinations .................................93

**Table 4.** Comparison of the proposed approach with the conventional classifiers. 96

**Table 5.** Processing time required by each component in the proposed approach (for retrieval) ..............................................................................................................................................96

**Table 6.** Comparison with state of the art methods in terms of MAP .............................99

**Table 7.** Confusion matrix .................................................................................................102

**Table 8.** Processing time required by the major components of the proposed system ..............................................................................................................................................104

**Table 9.** Experimental settings of the related methods .................................................104

**Table 10.** Average recognition accuracy yielded by each method ...............................105

**Table 11.** Average recognition accuracy yielded by each combination .................107
Chapter I. INTRODUCTION

I.1 Introduction

In the last two decades, there has been an explosive growth of technology touching the different aspects of people’s daily life. The novel technologies have drastically changed the people’s life style, interests, habits and communication way. This last is among the main aspects concerned with this growth. Numerous sophisticated devices (Ex. laptops and smartphones) that allow acquiring and exchanging photos and videos have been appeared. Due to the large diversity of features they provide, these devices become the preferable choice for peoples in their daily communication. Indeed, this choice can be explained by many facts. First, instead of traveling for thousands of kilometers to deliver a message to its recipient, only some touches on the keypad of a cell phone or a computer is sufficient to perform this hard task. Thus, preserve a great deal of time, effort and cost. Second, such devices can easily be manipulated; one can navigate through the web, chat with their friends, play games, take photos and record videos in few simple steps. Third, the low prices of these devices compared to the features they offer, make them available for a large proportion of people. Recently, and with the prevalence of the social networks and the immense increasing of the number of their users, the use of those devices has noticeably increased. Many of the users believes the old saying “a picture is worth a thousand of words” and use pictures to transfer their messages and express their feelings. They take photos in the different events (wedding, holiday, trip…), places they visit (zoo, cites, forests…) and activities they practice (swimming, soccer…). This has led to an immense increasing of the size of personal photo albums as well as the
number of images in the web. For the user, the classical question: does this image exist? Is changed to be: how can I quickly locate this image?

One way to locate a target image is by manually browsing image collections. In nowadays, such a solution is far away from realization, especially with the large-scale collections such as the ones we have in our own computers. This heightened the need to develop tools and techniques that help users to quickly and accurately locate images they look for. Hence, many image retrieval engines have been appeared. They can roughly be classified into two categories: text-based image retrieval (TBIR) and content-based image retrieval (CBIR). In TBIR, annotations are manually assigned to images by human annotators. User can naturally express his needs using a textual query; relevant images are then identified based on matching the query with the annotations. Given the manner it operates, TBIR has achieved a great success, as human is, in the same time, who assign the annotations and judge the retrieval results. However, a tedious and time-consuming work is required to manually annotate images especially when dealing with large-scale datasets. Besides, in some cases, annotations do not reflect image content, as they are ambiguous, confused with noisy ones or even completely irrelevant. In some other cases, images can be found without any kind of annotations, and thus making TBIR unusable.

CBIR is the alternative solution to overcome the TBIR shortcomings. It takes an image (or images) as a query and returns images which are visually similar to the query. Despite the efforts that have been made to refine the CBIR performance, much more efforts are needed to tackle two main issues. The first one is the image representation, given the big variety of image features that have been proposed in the literature, what is the combination of features capable to accurately describe image content? To nowadays, this is still an open issue because new features are proposed every day. The second one is the semantic gap between low-level features and high-level user semantics.

I.2 Problematic

CBIR is applied to solve a wide range of problems, including critical applications such as medical diagnosis and crime prevention. Besides, it is used in several other applications, including image retrieval from personal collections, remote sensing, industrial and military applications. Hence, CBIR systems should perform well and satisfy their users by
assuring a high performance. Unfortunately, most existing CBIR systems (La Cascia, Taycher et al. 1997, Rui, Huang et al. 1997, Kherfi, Ziou et al. 2002, Kherfi, Ziou et al. 2003, Tao, Tang et al. 2006) engines don’t meet the necessary requirements and their performance is still unsatisfactory, as they suffer from serious limitations. One of the most common limitations is the semantic gap between low-level image features and the high-level user semantics. The semantic gap occurs because of the possible contradiction between two interpretations, one by user and the other by the machine, which may be assigned to the same image (Smeulders, Worring et al. 2000, Liu, Zhang et al. 2007, Yang and Zhu 2012). In other words, the semantic gap is the discrepancy between what the user wants and the retrieved images by the engine. For the sake of clarity, let us illustrate the semantic gap with a real example. Figure 1 shows the top retrieved images by the well-known and the widely used engine ‘Google Images’ when submitting the image at top as a query. The first observation we can make is the homogeneity of visual characteristics between the query and the retrieved images. Although the query image presents an image of a ‘bird’, none of the retrieved images contain a ‘bird’. This can be explained by the fact that CBIR doesn’t interest with the semantics associated the query image, but rather with the visual features of the query image.

Figure 1. Top retrieved images by ‘Google Images’ by submitting the image at top as a query
I.3 Overview on the related work

In the literature, several attempts have been made to tackle the semantic gap issue (Feng, Manmatha et al. 2004, Kherfi and Ziou 2006, Guillaumin, Mensink et al. 2009, Chen, Xu et al. 2012, Verma and Jawahar 2012, Ballan, Uricchio et al. 2014, Yang and Cai 2014). They can be categorized on the basis of several criterions. For instance, according to the application for which they are intended, those methods can broadly be classified into those intended for the web applications and those dedicated to local applications.

Based on the used techniques, they can roughly be classified into three main approaches. The first approach includes methods that attempt to automatically infer the relationship between low-level image features and high-level user semantics (Lavrenko, Manmatha et al. 2003, Jia, Yu et al. 2008, Liu, Xu et al. 2011, Verma and Jawahar 2013, Murthy, Maji et al. 2015), which commonly known as automatic image annotation. Automatically inferring such a relationship is, in fact, of great interest because it enables the engine to determine the semantics (i.e., concepts) associated with images. Thus, be capable to successfully retrieve the relevant images for a particular query. To do so, diverse machine learning techniques and algorithms, such as Support Vector Machines (SVM), are used. Machine learning methods consist in two main stages, namely training and testing. The training stage (also known as learning stage) aims to model the relationship between image features and user semantics. The outcome of the training stage is a visual model that summarizes the relationship between image features and user semantic concepts. The extracted model can then be used to tag (i.e., annotate) yet annotated images.

The second approach involves methods arguing that information extracted from image regions could be more useful than that extracted from the whole image (Feng, Xu et al. 2010, Yang and Cai 2014, Yang, Lv et al. 2014). The motivation behind those methods is that users are mainly concerned with objects present in the image. Region-based methods used segmentation techniques, such as Normalized cuts (Shi and Malik 2000), to perform segmentation. Then, similarity techniques are used to determine the degree of relevance between the gallery image regions and those of the query.

A third approach assembles methods aiming to reduce the semantic gap using the relevance feedback technique (Bulo, Rabbi et al. 2011, Marakakis, Siolas et al. 2011, Qian, Tan et al. 2016). The general principle of those methods is to involve the user in the process
of bridging the semantic gap. The relevance feedback is an iterative procedure in which user judges the retrieval results, in each iteration, and the engine tries to improve these results, taking into account the user feedback. Hence, in such methods, the user is the decisive player. 

It is worth noting that in this section we give only an overview on the related work, while we devote a whole chapter (Chapter 3) to detail the different aspects of methods concerned with bridging the semantic gap.

I.4 Motivation

In spite of the significant efforts that have been made by researchers to solve the semantic gap issue, much more efforts are still needed. This issue is yet an open issue because of the numerous issues arose when researchers tried to solve it. Below, we briefly summarize the main issues:

- Some of the existing methods (Jeon, Lavrenko et al. 2003, Feng, Manmatha et al. 2004, Liu, Wang et al. 2007, Xiang, Zhou et al. 2009) model the relationship between visual features and semantic concepts based on their co-occurrence in a local annotated dataset. In fact, this is impractical for several reasons. First, those datasets are manually annotated by human annotators; these annotations (also called ‘labels’) are often ambiguous, noisy, inconsistent or even completely false (Wu, Jin et al. 2013). In addition, appearance frequency of the annotations over the dataset is extremely imbalanced (Verma and Jawahar 2012). All the pre-mentioned problems can potentially lead to significantly distort the outcome models, and thus negatively affect the quality of annotation results. Moreover, make very complicated to generalize such solutions to other cases in which images are either different in terms of visual features or labels frequency.

- Methods which leverage web images and used them for training (Jia, Yu et al. 2008, Wang, Zhang et al. 2008, Wang, Zhang et al. 2008, Dai, Wang et al. 2012), in turn, encounter many other barriers and difficulties. First, as mentioned above, labels assigned to web images may be noisy, ambiguous or unsuitable. Second, the presence of outlier images within the training set could probably lead to a minor performance. Furthermore, although the huge amount of images
used for training and the high dimension of visual features extracted from images, performance of web-based methods is yet insufficient.

- Several methods from the state of the art, such as the ones in (Murthy, Can et al. 2014), are computationally expensive because of the complicated learning techniques they adopted. Therefore, they cannot be applied to solve critical applications where the response time is amongst the crucial factors to be taken into consideration.

- In the methods adopting the relevance feedback technique, user should participate in the process of reducing the semantic gap. Most users tend to provide only a quite limited feedback, which could degrade the quality of the retrieval results (Yang and Zhu 2012). In addition, for some users, this is boring and time-consuming, especially if several iterations are needed to achieve the desirable results.

- The performance of the region-based methods is highly dependent on the quality of segmentation results. Thus, any bad-segmented images could strongly influence the performance of the method.

Given the above issues, we are motivated to propose a new solution to bridge the semantic gap between low-level image features and high-level user semantics. The next section summarizes our contribution and shows how we addressed the above issues.

### 1.5 Contributions

Because of the full independence between image visual characteristics and user semantic concepts, CBIR suffers from the semantic gap issue. In this thesis, we propose a new fully automatic approach for bridging the semantic gap in CBIR. To reach such an aim, the proposed approach use machine learning techniques in order to overcome the drawbacks of the previous methods. The core idea behind our proposed approach is to use a set of images labeled with a particular concept in order to extract (i.e., learn) the concept’s visual representation (i.e., visual model). During retrieval, user can naturally express his needs using
a textual query. Then, the query is replaced by the corresponding visual model which is used to perform the retrieval process.

The major contributions of this thesis reside mainly in the following aspects:

- Allows user to use a textual query in retrieving images from the unlabeled (i.e., un-annotated) collections. In such a way, we overcome the two main drawbacks of each of CBIR and TBIR, which are the semantic gap and the lack of annotations with images, respectively. In addition, formulating the query using text, rather than searching for a query image that visually match the targeted images, allows the user to naturally express his needs.

- Training images are gathered from the web, thus, the proposed approach is not limited to any pre-defined concept.

- We use outlier detection methods to solve the problem of the presence of outlier images within the training set. In addition, we limit ourselves in using only the top images returned by the web search engines, thereby maximizing the chance to acquire the most representative images.

- At the opposite of several state of the art methods, neither prior segmentation nor relevance feedback are required by our approach. In contrast, it is computationally fast and operates in a fully automatic manner.

- To the best of our knowledge, none of the existing methods take into account the intra-variation of images representing the same concept. The proposed approach, however, takes into account this important aspect and exploits it to enhance the retrieval results.

- Because we formulate the retrieval as a supervised classification problem, our approach is capable to perform the recognition task. In particular, the proposed approach has successfully applied to visual object recognition. More specifically, we use our approach for automatic date fruit recognition, where superior results have been achieved.
• For both retrieval and recognition, the proposed approach has outperformed several traditional classification methods as well as several recently-proposed methods from the state of the art.

I.6 Thesis structure

In addition to the first chapter which presents the general introduction, the thesis is organized as follows:

• The second chapter is dedicated to introduce the background and the general context of the work. We review the important notions related to the Content-Based Image Retrieval (CBIR). These notions include query formulation, image representation, similarity measures and the main applications of CBIR as well as the common issues encountered by CBIR systems.

• The third chapter is devoted to survey the literature methods concerned with the semantic gap issue. We review several state of the art methods that are close to ours. The presented methods belong to three different approaches. As the proposed method is applied for recognition tasks, we also review the state of the art methods that interest with visual object recognition, especially those concerned with date fruit recognition.

• The fourth chapter presents our contribution and the system we design to bridge the semantic gap. The different components constituting the system are explained and detailed. In this chapter, we explain how web images, used for training, are gathered and how we prune them from outliers. Then, we explain, in details, our contribution in modeling the semantic concepts, and the use of models for both retrieval and recognition tasks.

• In the fifth chapter, we first detail the experimental setup on which we carried out the experiments. The experimental setup includes the used datasets, the performance metrics and the parameters tuning. Then, we review the
experiments we conducted to prove the effectiveness of our approach. The experiments section is divided into two main parts; the first part is intended to test the retrieval performance, while the second is devoted to investigate the performance of the proposed approach in the recognition tasks. Each of the conducted experiments is designed to measure a specific aspect from the proposed method. The main covered aspects include, the overall performance in both retrieval and recognition, computational complexity, strength of different feature combinations and comparison with the state of the art methods.

At the end of the thesis, we draw the main conclusions of the work and we introduce some perspectives and future works.
Chapter II. CONTENT-BASED IMAGE RETRIEVAL (CBIR): PRINCIPLES, FUNDAMENTAL NOTIONS, APPLICATIONS AND COMMON ISSUES

II.1 Introduction

Researches in the field of image retrieval have started in the early of 1970’s (Rui, Huang et al. 1999). The earliest efforts have been focused on searching images by text (i.e., Text-Based Image Retrieval (TBIR)) (Chang and Fu 1980, Chang and Fu 1980, Chang, Yan et al. 1988). The retrieval process is carried out by matching the textual descriptions assigned to images with the textual queries. Afterwards, with the emerging of the applications in which text is partially or entirely absent, TBIR cannot perform the task of image retrieval because of the lack of textual descriptions with images. Examples of such applications include person recognition by face or fingerprint, military purposes, medical diagnosis and image retrieval from unlabeled personal photo collections. Hence, Content-Based Image Retrieval (CBIR) systems have been appeared to satisfy this need.

Content-Based Image Retrieval (CBIR) is an active research field, since the 1990’s, which attracted much attention from researchers. This is because of the wide range of applications it covers as well as the several open issues it encloses. The aim of this chapter is to present the architecture of typical CBIR systems and their major components including
query formulation, image description and similarity measures. We also present the fields in which the CBIR is applied in addition to the common issues encountered by CBIR systems.

II.2 Principle of CBIR system

CBIR is the technique that aims to automatically locate relevant images to query image(s) based on the visual content of images (Liu, Zhang et al. 2007). In other words, CBIR use the low-level image features such as color, texture and shape in order to automatically locate, within an image database, the semantically-relevant images for query images. A high degree of accuracy and a high speed in retrieval (i.e., short response time) are two main desirable criterions in CBIR systems. To satisfy the last concern (i.e., speed in retrieval), an offline process, which is called indexation, is conducted. This process consists in extracting the visual features of images and then indexing them based on their features. Image indexation should be performed in a manner that allows to quickly achieve targeted images during retrieval process (Berchtold, Keim et al. 2001, Kherfi and Ziou 2007).

After images having indexed, user can perform the retrieval process online. Suppose that we want to search and retrieve images from a database of images that assembles images tagged with different textual descriptions. Suppose also that the different features describing the different aspects of images were extracted offline. A typical CBIR system is illustrated in Figure 2.

![Figure 2. Architecture of a typical CBIR system](image-url)
Let us explain the retrieval steps depicted in Figure 2:

1. The user submits a query, which is supposed to be similar to the targeted ones.
2. The system measures the similarity between the query/queries images and the database images on the basis of the low-level features.
3. The system ranks the images according to their degree of similarity to the submitted query/queries, and then displays the top ones to the user.

In the next sub-sections, we provide details about the major components of the typical CBIR system depicted in Figure 2.

II.3 Major components of a CBIR system

II.3.1 Features extraction

The most intuitive approach to seek images from an image database is to directly compare the query image with those of the database. That is, comparing pixel values from the query image with their respective in the image collection. For a CBIR, applying such an approach is, in fact, computationally expensive and ineffective as well (Huang and Tran 2016). What being desired is a compact representation that permits to effectively describe the image and significantly reduce the computation cost. Image features, known also as image signatures, are numerical values which encode the image and describe its content from diverse aspects, including color, texture, shape and key-points. Therefore, before being capable to match images and to distinguish the relevant from the irrelevant ones, CBIR should perform the process of feature extraction.

In the literature, a large variety of image features were proposed (Pass and Zabih 1996, Lin, Wang et al. 1997, Chen, Liu et al. 2010, Banerji, Verma et al. 2011, Liu, Zhao et al. 2012). They can be classified into different categories. The first category interests with describing the color aspect and it includes color histogram (Swain and Ballard 1991), color moments (Flickner, Sawhney et al. 1995), color coherence vector (Pass and Zabih 1996), etc. The second category groups methods which are concerned with representing texture aspect, it involves Gray-Level Co-occurrence Matrix (GLCM) (Haralick and Shanmugam 1973), Gabor filters (Weldon, Higgins et al. 1996), Local Binary Patterns (LBP) (Ojala, Pietikainen et al.
CHAPTER II. CBIR: PRINCIPLES, FUNDAMENTAL NOTIONS, APPLICATIONS AND COMMON ISSUES


As image features are amongst the major components in most retrieval and recognition systems, including ours, an overview of the used image features in this work is given in chapter 4.

II.3.2 Query Formulation

The query in the earlier TBIR systems is formulated using either keywords or sentences. Formulating the query in such a way could potentially lead to serious limitations. For instance, TBIR systems cannot be used to retrieve images from the unlabeled collections, simply because of the lack of textual descriptions with images. Unfortunately, most existing collections, especially the personal ones, are unlabeled because a high cost, in terms of time and manual labor, is required to label them. In other situations, using text could yield minor results because the image labels suffer from the subjectivity of the human labeler. Thus, the TBIR performance is extremely dependent to quality and the availability of textual information with images (Wu, Jin et al. 2013). All the above issues have motivated the researchers to introduce a new way that allows user to query the system and retrieve images.
In CBIR, user is required to supply a query that is supposed to be similar to the ones he wants. This query can be formulated either by images (Kelly, Cannon et al. 1995, Das, Riseman et al. 1997, Ma and Manjunath 1999, Brunelli and Mich 2000) or by sketch (Lew, Lempinen et al. 1997, Mukherjea, Hirata et al. 1999) or even by supplying the values of visual features (Wei, Li et al. 1998). Before going ahead in explaining each of those formulations, let us give an example which justifies why queering using images is, in some cases, better than queering using text. Figure 3 shows the top retrieved images by ‘Google Images’ for the query ‘apple’.

Even the user intention was to search an ‘apple fruit’, but the retrieved images seem to be confused with other images such as image representing the logo of ‘apple’ trade mark. However, by submitting the image at the top of Figure 4 as a query, we see that the top retrieved images are strongly relevant to the query.
II. CBIR: PRINCIPLES, FUNDAMENTAL NOTIONS, APPLICATIONS AND COMMON ISSUES

Figure 4. Top retrieved images by Google Image by submitting the image at top as a query

II.3.2.1 Query by providing the values of the visual features

In some CBIR systems, as in ((Wei, Li et al. 1998), user can formulate the query by supplying the numerical values of each feature. Obviously, this is unpractical, as it is very difficult for an ordinary user to determine numerical values for the different type of features. Suppose that the system use Hu moments as features to describe the shape, who is capable to tune the values of this descriptor to match the desirable shapes. In such a case, even a specialist cannot do that.

II.3.2.2 Query by image(s) example(s)

In this technique, the CBIR system suggests a set of images to the user, and then the user selects the ones that resemble the images he wants. Note that in certain systems user can provide image(s) and ignores those suggested by the system. In the case where the user submit more than a single image, logical connectors, such as AND, OR and NOT, are used to combine the query.
Query by image example may take several forms. For instance, it can be performed based on the region of interest instead of queering using the image as a whole (Del Bimbo, Mugnaini et al. 1998, Ko, Peng et al. 2001, Yong and Gang 2009, Abuhaiba and Salamah 2012). This is because user, in sometime, is interested with only a particular object within the image. It can also be performed by using positive along with negative examples (Muller, Muller et al. 2000, Kherfi, Ziou et al. 2003, Kherfi and Ziou 2006).

II.3.2.3 Query by sketch

In this case, the user designs a sketch that look alike with the images he looking for (Lew, Lempinen et al. 1997, Mukherjea, Hirata et al. 1999). To do so, the system should provide the user with a set of tools that allow him to draw the sketch. In another case, the system can provide the user with a set of icons representing different objects such as ‘car’, ‘sun’, ‘tomato’, ‘Bicycle’, ‘ball’, etc; the user can then use those icons to construct the query image (Lew, Lempinen et al. 1997).

This way of formulation could be useful in the case where the query image is simple, whereas it is very difficult to design the complex images. Moreover, some users cannot make sketches in real life, and thus they cannot make them using an electronic pen with limited drawing possibilities.

II.3.3 Similarity measures

Various computer vision tasks are relying on measuring the similarity between images. A non-exhaustive list of those tasks involves image classification (Mao and Jain 1992, Laine and Fan 1993, Ojala, Pietikäinen et al. 1996), retrieval (Forsyth, Malik et al. 1996, Manjunath and Ma 1996, Pentland, Picard et al. 1996) and unsupervised segmentation (Jain and Farrokhnia 1991, Hofmann, Puzicha et al. 1998). In the context of image retrieval, similarity measures are used to compare images with each others in order to identify the pertinent images for a particular query image. A given measure can be applied on the numerical values yielded by the process of image features extraction. Note that the selection of the similarity measures has a great influence on CBIR performance, while the optimal measure can be determined as a function of the selected feature (Long, Zhang et al. 2003).
comparison between various similarity measures in conducted in (Rubner, Puzicha et al. 2001).

We can distinguish two families of measures, the first family covers the measures in which all the bins of an image descriptor (i.e., numerical values describing the image) contribute equally in calculating the similarity degree. A second family includes measures which tend to give more importance to the bins having comparable values.

Given two images A and B, where $I_1$ and $I_2$ are two $M$-dimensional feature vectors extracted from A and B, respectively. Hereafter, we provide some similarity measures that appertain to the first family.

### II.3.3.1 Minkowski distances

Minkowski distances encompass the set of functions given by

$$D_p(I_1, I_2) = \left( \sum_{i=1}^{M} |I_1(i) - I_2(i)|^p \right)^{1/p}$$

where $p \geq 1$, $D_1$ represents the Manhattan distance, $D_2$ represents the Euclidean distance and $D_\infty$ represents the maximum distance. This distance can easily be implemented and computed as well.

### II.3.3.2 Kullback-Leibler divergence:

The Kullback-Leibler divergence is defined as

$$D_{KL}(I_1, I_2) = \sum_{i=1}^{M} I_1(i) \log \frac{I_1(i)}{I_2(i)}$$

This distance measures the dissimilarity of two different distributions.
II.3.3.3 Jeffery divergence

The Jeffery divergence is symmetric, is defined by

$$D_{JD}(I_1, I_2) = \sum_{i=1}^{M} I_1(i) \log \frac{2I_1(i)}{I_1(i) + I_2(i)} + I_2(i) \log \frac{2I_2(i)}{I_1(i) + I_2(i)},$$

(3)

This divergence is symmetric and it is also known as Jensen Shannon.

II.3.3.4 Relative deviation

The Relative deviation indicates how much a vector is deviating from the other, is given by

$$D_{RD}(I_1, I_2) = \sqrt{\frac{\sum_{i=1}^{M} (I_1(i) - I_2(i))^2}{\frac{1}{2} \left( \sum_{i=1}^{M} I_1(i)^2 + \sum_{i=1}^{M} I_2(i)^2 \right)}}$$

(4)

II.3.3.5 Histogram intersection

This measure was proposed by (Swain and Ballard 1991), it calculates the joint bins between the histograms (i.e., vectors) and it ignores the values contained in only one histogram, it is defined as

$$D_{Hist}(H_1, H_2) = 1 - \frac{\sum_{i=1}^{M} \min(H_1(i), H_2(i))}{\sum_{i=1}^{M} H_2(i)}$$

(5)

A value of $D_{Hist}$ near the 1 indicates that high similarity of the histograms.

Measures belonging to the second family include the following:
II.3.3.6 Quadratic distance

In this measures, a matrix of similarity, referred to as \( A = [A_{i,j}] \), is introduced in order to calculated the similarity between the different vector bins, where \( A_{i,j} \) represents the distance between the bin \( i \) and \( j \), respectively. The quadratic distance is given by

\[
D_Q(I_1, I_2) = \sqrt{(I_1 - I_2)^T A(I_1 - I_2)}
\]  

(6)

And \( A_{i,j} \) is calculated as follows

\[
A_{i,j} = 1 - \frac{D_Z(i,j)}{d_{max}}
\]  

(7)

Where \( D_Z(i,j) \) represents the euclidean distance between the bins \( i \) and the bin \( j \), and \( d_{max} \) represents the maximum distance

II.3.3.7 Earth Movers Distance (EMD)

The Earth Mover Distance (Rubner, Tomasi et al. 1998) is defined as the measure the minimum work should be performed to transform one distribution to another. It is considered as a special case of the transportation problem, it is defined as

\[
D_{EMD}(I_1, I_2) = \frac{\sum_{i,j} d_{i,j} g_{i,j}}{\sum_{i,j} g_{i,j}}
\]  

(8)

Such that \( d_{i,j} \) denotes the distance between the bins \( i \) and the bin \( j \), and \( g_{i,j} \geq 0 \) is the optimal flow between two distributions, where the total cost \( \sum_{i,j} d_{i,j} g_{i,j} \) is minimized according to the following constraints

\[
\sum_i g_{i,j} \leq I_2(j), \forall j
\]  

(9)
II.3.4 Examples of CBIR systems

Since the 1990’s, several CBIR systems have been introduced. Below, we briefly review some of them:

II.3.4.1 QBIC

Query By Image Content (QBIC) (Niblack, Barber et al. 1993) is one of the earliest CBIR systems, it is developed by IBM. QBIC uses the different visual features including a 256 RGB-based color histogram and the average color vector in RGB, LAB and YIQ color spaces. It also adopts texture features involving Tamura features (Tamura, Mori et al. 1978). Shape features are also used, including area, eccentricity and major axis length. QBIC system supports different ways of query formulations such as query by example as well as query by sketch. To speed up the process of retrieval, \( R^\ast - trees \) indexing scheme is adopted.

II.3.4.2 BlobWorld

BlobWorld was introduced in 1999 by the computer science division in the University of Berkeley (Carson, Thomas et al. 1999). It uses a histogram of 218 bins for color representation and shape is represented by the area, eccentricity and orientation extracted from the image regions. Besides, texture and location features were also considered. To measure the similarity between images, many distinct measures are used such as quadratic and Euclidean distance. During retrieval, user selects a blob (region) and indicates its importance along with the importance of each of the visual features it represents. Note that more than on blob can be selected.
II.3.4.3 SIMPLIcity

SIMPLIcity (Wang, Li et al. 2001) is the acronym of Semantics Sensitive Integrated Matching for Picture Libraries. In this system, images are represented with a set of regions that roughly correspond to the set of objects contained in the image. Each region is characterized with color, texture, shape and location features. To facilitate the retrieval process, images are automatically organized in semantic categories.

II.3.4.4 Atlas WISE

Atlas WISE (Kherfi, Ziou et al. 2003) is a web image retrieval engine which uses both textual and visual features to describe and index images. The visual features used to represent images are color histograms and edge-orientation histograms. Before launching retrieval, Atlas WISE collects web images from the popular pages such as Google and Yahoo. Collected images are then indexed according to their visual and textual features. Textual features of images are extracted from their tags, captions and page titles. To perform retrieval, the engine adopts the relevance feedback technique that combines both positive and negative examples.

II.3.4.5 ImageScape

In this system (Lew, Lempinen et al. 1997), query can be formulated using sketch, where a dedicated interface is provided. User can also bring icons and put them in the suitable position on the canvas. The available icons involve sky, grass, water, sand, stone and others. Several pixel-level features, including Laplacian, invariant moments and Fourier descriptors, are used to describe images.

II.3.4.6 PicHunter

PicHunter (Cox, Miller et al. 2000) is developed by NEC Research Institute. It describes images using color histogram and color spatial distribution together with textual annotations. Some other features, including HSV histogram and HSV color autocorrelogram, are also employed, where Minkowski distance is used to match features. PicHunter adopts a probabilistic relevance feedback mechanism in order to predict the desired images by the user.
Numerous other CBIR systems were also proposed. For further information, reader can refer to (Veltkamp and Tanase 2001).

II.3.5 Application fields of CBIR

Image retrieval based on content covers a large and diverse variety of interesting and very useful applications. In medicine, image retrieval is utilized to find out and display the relevant previous cases relevant to a specific case. ASSERT (Shyu, Brodley et al. 1999) and 3D PET/CT (Figure 5) (Fathabad and Balafar 2012) are retrieval engines intended for medical applications. Automatic Search and Selection Engine (ASSERT) (Shyu, Brodley et al. 1999) is used for the retrieval of high resolution tomography images of the lung.

![Figure 5. 3D PET/CT image retrieval system (Fathabad and Balafar 2012)](image)

As for the judiciary applications, image retrieval techniques can help the police targeting to verify the presence of a suspect face within a database of criminals. In the case where the criminal is yet unknown, image retrieval can assist the police in quickly narrowing down the number of potential suspects. This can be reached by drawing a sketch for the suspect face and search its relevant within the database containing the criminals faces. Image retrieval can also be used for the military purposes such as recognizing the engine of enemies through the radar images. Journalists use image retrieval for several purposes such as illustrating articles and adverts.
Beside, retrieving images based on content covers also other interesting applications including museum management, weather forecasting, biometrics, remote sensing, geographic information system and architectural design (Gudivada and Raghavan 1995).

II.3.6 Common issues of CBIR

CBIR presents a multitude of advantages and can be used in certain situations in which the TBIR cannot be used. In the case of the lack of textual descriptions with images, CBIR is the only available alternative. In addition, it provides another interesting advantage compared to the TBIR, which its ability to accurately locate certain kinds of images such as the complicated ones which can be tagged with numerous keywords. It potentially yields better than TBIR in the case of images with very specific features, as shown in Figure 3. Moreover, searching images based on their content could be better than searching with text because this latter is, in sometimes, affected by the noise and human subjectivity.

However, CBIR suffers from a number certain of limitations and common issues which should be addressed to improve its performance, following, we summarize the common issues encountered by CBIR systems.

II.3.6.1 Feature extraction and selection

As explained above, visual features are amongst of the major components is any CBIR system and they play a decisive role in the quality of retrieval outcomes. Until nowadays, several features, describing images from distinct aspects, are proposed (Khan, Van de Weijer et al. 2013, Satpathy, Jiang et al. 2014, Khan, Anwer et al. 2015) and new others are being proposed every day. The challenge behind that is to find the feature that best describes the image content. In particular, the main challenge is to detect the feature that is capable to capture the maximum from the semantics associated with images. Hence, features extraction remains an open issue, and a lot of efforts are remains to be done.

Another issue related to image features is to determine for a particular application the most appropriate feature combination, which is commonly known as feature selection (Kherfi, Ziou et al. 2003, Armanfard, Reilly et al. 2016). This is because two different or slightly different combinations may yield noticeably different outcomes.
Several algorithms for feature selection have been proposed (Chakraborti and Chatterjee 2014, Bouchrika, Harrati et al. 2015, Armanfard, Reilly et al. 2016).

II.3.6.2 Page zero problem

During retrieval, most CBIR systems suggest to user some sample images (page zero) from which he selects the ones relevant to what he looking for. Although, in some cases, none of the suggested images meet the user needs, this problem is known as page zero problem (Ziou and Boutemedjet 2006). Numerous approaches have been adopted to overcome this problem. For instance, the system can provide the user with typical image samples from all the categories, otherwise, user can supply an external image as a query. Certain systems provide user with a set of icons that it can use to formulate his query such as in ImageScape (Lew, Lempinen et al. 1997).

In (Kherfi, Ziou et al. 2003), it has been showed that utilizing positive along with negative examples can help to alleviate the problem, other solutions are presented in (La Cascia, Sethi et al. 1998, Ziou and Boutemedjet 2006).

II.3.6.3 Semantic gap

The existing techniques in the field of computer vision allow describing images either by their content or by semantics extracted from their surrounding text. Figure 6 shows both the visual and textual description for the image at left.
CBIR relies mainly on matching the visual features of images, and it doesn’t consider the semantics associated with images. However, relying only on the visual features could probably lead to poor results. This is because, in certain cases, there is only a slight visual resemblance between images representing the same semantic concept. To prove this pretend, let us illustrate this by an example.

**Figure 6.** Both visual and textual descriptions for an image

**Image:**
Ghazal in the ground

**Figure 7.** Images representing the semantic concept ‘lunch’
Figure 7 shows images that represent the semantic concept ‘lunch’. The first observation we can make is that there is only a little visual resemblance between those images. Supposing that a CBIR system is queried with the first image at left in Figure 7, without any doubt, the system will consider most of the remaining images as irrelevant. Although it is obviously for us, as humans, that all these images represent ‘lunch’, for any CBIR system, whatever the features and the similarity measures it uses, those images will be considered as irrelevant to the query. This contradiction between the human judgment and the CBIR provided results is commonly known as the semantic gap issue (Figure 8). The semantic gap is the discrepancy between two interpretations, one by the user and the other by the machine, that could be assigned to the same image (Smeulders, Worring et al. 2000, Liu, Zhang et al. 2007, Yang and Zhu 2012).

Figure 8. Semantic gap between low-level image features and high-level semantics

One way to alleviate the semantic gap is by making CBIR capable to detect the semantics associated with images. In such a case, the human judgment will certainly meet the CBIR outcomes. The main focus of the present work is to bridge the semantic gap in CBIR in a fully automatic manner with the minimum of requirements in terms of computational complexity. In the next chapter, we survey, in details, the state of the art methods focusing on reducing the gap between user semantics and image features.

II.4 Conclusion

Content-Based Image retrieval is an active area of research, which have attracted a lot of attention from researchers over the two last decades. This chapter has been mainly devoted to state the background of our work. In this chapter, we have presented the
fundamental notions related to this area as well as the common issues encountered by researchers. In particular, we have explained the principle of a typical CBIR system in addition to its major components, involving feature extraction, query formulation and similarity measures. We have provided details about each component separately. Examples of some existing CBIR systems have been reported along with a brief definition for each of them. Then, we have demonstrated the strong need for CBIR systems by presenting a certain number of useful and interesting applications of these systems. Throughout the chapter, a multitude of CBIR problems have been stated together with the proposed solutions for them.

We have concluded the chapter by explaining the common issues of CBIR systems. We have explained and illustrated the semantic gap issue which is our main concern in this thesis. We also showed that many issues, including the semantic gap, are remains in questions.
Chapter III. BRIDGING THE SEMANTIC GAP IN CBIR: A STATE OF THE ART REVIEW

III.1 Introduction

By considering the interesting applications of CBIR systems, alleviating the negative effect of semantic gap is highly recommended, especially in the case of the critical applications such as the medical diagnosis. In the last two decades, significant efforts have been devoted to address the semantic gap issue. Nevertheless, despite those efforts, the semantic gap is still an open issue.

In the remainder of this chapter, we present an overview on the related work to this thesis. At first, we start by categorizing the works concerned with reducing the semantic gap. Then, we separately provide details for methods belonging to each category. As our proposed approach is also applied for the task of visual object recognition, especially date fruit recognition, we devote a section that discusses the works investigating this task.

III.2 Categorization of methods aiming to bridge the semantic gap

The state of the art methods that aim to alleviate the semantic gap can be classified according to several criterions. (Yang and Zhu 2012) classified these methods, based on the technique they adopt, into three main approaches which are: automatic image annotation approach, region-based approach and relevance feedback approach. The target of the first
category of methods is to make CBIR able to automatically detect semantic concepts associated with images, as operating using only the visual features could probably lead to fall in the semantic gap (Jin, Shi et al. 2004, Carneiro, Chan et al. 2007, Wang, Zhou et al. 2008, Zhang, Huang et al. 2010, Zhao, Lu et al. 2013, Ballan, Uricchio et al. 2014, Murthy, Maji et al. 2015, Jing, Wu et al. 2016, Murthy, Sharma et al. 2016). Automatically assigning an image with a set of concepts that describe its semantic content is known as automatic image annotation.

The second category assembles methods assuming that using global features extracted from the entire image make difficult for CBIR to locate targeted images by user (Parashar 2009, Manipoonchelvi and Muneeswaran 2011, Zhang, Islam et al. 2012, Belloulata, Belallouche et al. 2014, Yang and Cai 2014, Gallas, Barhoumi et al. 2015, Manipoonchelvi and Muneeswaran 2015, Chaudhuri, Demir et al. 2016). This is because features of different objects present in image are pooled in the same feature vector, and therefore features of desired object will be confused with the undesired ones. They proposed to rather opt for the local features extracted from regions.

Another category of methods propose to involve user in the process of retrieval so as to understand what he wants exactly (Marakakis, Siolas et al. 2011, Su, Huang et al. 2011, Yang, Nie et al. 2012, Wang, Han et al. 2015, Qian, Tan et al. 2016). Those methods use the technique commonly known as relevance feedback (RF). The aim of this technique is to iteratively understand the user query.

### III.3 Automatic image annotation approach

As stated in the previous chapter, the origin of the semantic gap problem is the use of low level image features which cannot reflect the semantics contained in the query image. Automatic image annotation methods aim to bridge the semantic gap based on identifying the set of concepts present in images. To do so, learning techniques, which aim to make the machine capable to automatically perform learning, can be used. Although it is easy for a human to learn a semantic concept, computationally performing such a task is quite complicated. Broadly, machine learning algorithms can be categorized into three main categories: supervised, semi-supervised and unsupervised. In the following, we explain the principle of each of those categories and we review the methods belonging to each of them.
III.3.1 Automatic image annotation based on supervised learning

Supervised learning uses labeled training data to predict response measure (i.e., outcome) for new unlabeled data (Jain, Murty et al. 1999). The core idea behind the supervised learning is to optimize a particular function, using the available techniques, so that for a yet seen sample, the optimized function can accurately determine the suitable outcome. Classification algorithms, involving parametric and non-parametric ones, are typical formulation of supervised learning task. Now, let us give an overview on certain representative methods, from this category, which are close to ours:

III.3.1.1 A baseline method for auto-image annotation

In this work (Makadia, Pavlovic et al. 2008), authors started from a reasonable critic of the previous methods and asked the question: why they didn’t justify the need for the complex models they proposed? Simpler methods could probably achieve a comparable performance. Alternatively, authors proposed a baseline method for automatic image annotation based on the hypothesis arguing that visually similar images are more likely to share annotations. In particular, a set of low-level image features as well as a combination of some basic distances are firstly used to find the k-nearest neighbors of an unannotated image $I$. Then, a label transfer mechanism is used to transfer annotations from the $k$-nearest neighbors to $I$. Details of this method are given in the following:

Given unannotated image, the method is asked to predict labels for it. A first step consists in finding the $k$-nearest neighbor of $I$ in a training set $T$:

- Extract a set of low level image features from $I$ and $T$. The extracted features are:
  - Color features from 3 color spaces namely: RGB, HSV and LAB.
  - Texture: Gabor and Haar wavelets.
- Employ a set of well-known distances to measure the distance between $I$ and images in $T$. Those distances are:
  - $KL$-divergence, $Chi2$-statistic, $L1$-distance and $L2$-distance.
The distances are evaluated on the training set $T$ in order to determine for each feature the distance that can appropriately used with it. For example, it was found that $KL$-divergence best fit the LAB features.

In the case of lack of a labeled training set, the method opt for two techniques to combine distances:

- Joint Equal Contribution (JEC): allows all distances to contribute equally.
- $L_1$–Panelized Logistic Regression (known also as Lasso): a method proposed in (Tibshirani 1996).

Use the extracted features in addition to the combination of the distances to find the $k$-nearest neighbors of $I$ in $T$.

A second step is to Transfer $n$ labels from the set of $k$-nearest neighbors $\text{knearst} = \{l_i, i = 1, \ldots, k\}$ to $I$:

- Given that $I_1$ is the most similar to $I$, annotations of $I_1$ is ranked according to their appearance frequency in $T$.
- Transfer the $n$ annotations having the highest scores to $I$. if $|I_1| < n$ then:
  - Rank annotations from $I_2$ to $I_k$ depending on:
    - Co-occurrence with annotations of $I_1$.
    - Appearance frequency through $I_2$ and $I_k$.
  - Transfer the $n - |I_1|$ annotations having the highest scores to $I$.

### III.3.1.2 Automatic image annotation using semantic relevance

(Zhao, Lu et al. 2013) proposed to introduce the semantic relevance between concepts when predicting labels for an unlabeled image. For instance, the concept “street” is most likely to occur with the concept “building” in the same image. Taking into account such a relation could improve the annotation performance. In addition, authors argue that using global features of images do not allow to well reflect local attributes of images. Therefore, they opt for the use of features extracted from image regions instead of those extracted from the entire image.

Formally, supposing that we have a set of words $W = \{w_i, i = 1, \ldots, n\}$, let us define $T_i$ the set of images labeled with $w_i$ and $V = \{v^i_m, m = 1, \ldots, p\}$, where $p$ is the number of regions in $T_i$. Feature vectors in $V$ are supposed to be normally distributed. This formulation
is known as Multiple Instance Learning (MIL) in which the relevant instances of a given label are pooled in the same bag. Instead of labeling each instance within the bag, only the label of interest is assigned for the entire bag.

Given an image \( I \) where \( R = \{r_j : j = 1, ..., b \} \) is the set of regions that form it. The probability of labeling \( I \) with \( w_k \) is computed as in Eq.(12)

\[
p(w_k|I) = \sum_{r_j \in I} p(w_k|r_j)p(r_j|I)
\]

Such that

\[
p(w_k|r_j) = \frac{p(r_j|w_k)p(w_k)}{p(r_j)}
\]

Where \( p(w_k) \) is the prior probability of \( w_k \), and \( p(r_j|w_k) \) is the likelihood that \( r_j \) belongs to the set of regions \( V \).

\( p(r_j|I) \) represents the importance of \( r_j \) in \( I \). A weighted method is used to compute this probability according to region position and brightness.

To improve the annotation performance, Eq.(12) is updated to introduce the semantic relevance of concepts, as in Eq.(14)

\[
p(w_k|r_j) = \frac{1}{1 + |S(w_k)|} \left( p(w_k|r_j) + \sum_{w_l \in S(w_k)} p(w_l|r_j)p(w_k|w_l,r_j) \right)
\]

Where \( S(w_k) \) denotes the set of concept that are in semantic relevance with \( w_k \). \( S(w_k) \) is computed statistically on the basis of the co-occurrence of concepts pairs using a training set. \( p(w_k|w_l,r_j) \) denotes the semantic contribution of \( w_l \) to \( w_k \) in the region \( r_j \), such that \( w_l \in S(w_k) \). It is calculated over a training set.
III.3.1.3 Textual query of personal photos facilitated by large-scale web data

(Liu, Xu et al. 2011) proposed a web-based framework to facilitate the process of retrieval in the consumer’s unlabeled photo collections, as depicted in Figure 9.

![Diagram of proposed framework](image_url)

**Figure 9.** The proposed framework in (Liu, Xu et al. 2011)

From Figure 9 we notice that the proposed framework is composed of several modules. To further understanding those modules and how the retrieval process is carried out, we summarize the different modules of the framework in the following:

- Leverage millions of images from the web together with their rich associated textual descriptions.
- The user supplies a textual query for the system, e.g., “horse”.
- The inverted file method is used to identify relevant and irrelevant images to the query. Irrelevant images (negative examples) are the ones which don’t contain neither “horse” nor its descendents of WordNet.
- Relevant and irrelevant images are used to train three classifiers namely SVM, KNN and decision stumps.
- After training, images which are supposed to be pertinent to the query are ranked and then returned to the user.
- Consumer images may be significantly different from those of the web, which can negatively affects the retrieval results. To deal with this issue, two cross-domain techniques are proposed.

### III.3.1.4 Improving automatic image annotation with Google semantic link

(Xu, Pan et al. 2014) proposed a model which they refer to as Google Semantic Link based image Annotation Model (GSAM). As its noun indicates, GSAM automatically assigns unlabeled images with a set of semantically coherent concepts. The idea behind GASM is to enhance the annotation quality by taking into account the semantic relevance between concepts. For example, concepts like “sky” and “cloud” are most likely to co-occur in the same image.

Given the following:

1. A labeled training set of images $T$.
2. A training set $W$ composed of about 100 billion pages indexed in Google.
3. A set of vocabulary $V = \{x_1, ..., x_k\}$.

GASM is composed of two main steps which are:

1. For each concept in $V$, calculate the set $R(x)$ representing the set of concepts relevant to the concept $x_i$ as follows:

   \[
   GD(x_i, x_j) = \frac{\max\{\log f(x_i), \log f(x_j)\} - \log f(x_i, x_j)}{N_G - \min\{\log f(x_i), \log f(x_j)\}} \tag{15}
   \]

   - Calculate the Google Distance (GD) between each pair of concepts $x_i$ and $x_j (x_j \in V)$, GD is given by Eq.(15), where
   
   1. $f(x_i)$ and $f(x_j)$ denotes the number of pages that contains $x_i$ and $x_j$ respectively.
CHAPTER III. BRIDGING THE SEMANTIC GAP IN CBIR: A STATE OF THE ART REVIEW

2. \( f(x_i, x_j) \) denotes the number of pages that contains both \( x_i \) and \( x_j \).
3. \( N_c \) denotes the total number of pages indexed by the system (100 billion).

- Estimate the relevance between \( x_i \) and \( x_j \) using the probability as in Eq.(16)

\[
p_T(x_j|x_i) = \exp\left(-\frac{G_D(x_i, x_j)}{\delta}\right)
\]

(16)

where \( \delta \) is a distance parameter.

- Calculate the relevance between \( x_i \) and \( x_j \) over \( T \), namely \( p_T(x_j|x_i) \) as calculating \( p_W(x_j|x_i) \) the relevance between \( x_i \) and \( x_j \) over \( W \).

- Fuse \( p_W(x_j|x_i) \) and \( p_T(x_j|x_i) \) to get the overall relevance between \( x_i \) and \( x_j \), the fusion is done as shown in Eq.(17)

\[
p(x_j|x_i) = \begin{cases} 
\lambda_1 p_T(x_j|x_i) + \lambda_2 p_W(x_j|x_i) & \text{if } p_W(x_j|x_i) > 0 \\
0 & \text{elsewhere} 
\end{cases}
\]

(17)

Where \( \lambda_1 \) and \( \lambda_2 \) are two parameters that determine the trade-off between \( p_T(x_j|x_i) \) and \( p_W(x_j|x_i) \).

2. For each \( x_i \in R(x) \), the relevance of an un-annotated image \( I \) to a the concept \( x \) is calculated according to Eq.(18)

\[
r(x_i, I) = \sum_{i=1}^{R} r(x|x_i)p(x_i|I)
\]

(18)

Where the posterior probability \( p(x_i|I) \) is calculated using the SVM classifier.
Beside those studies explained above, many other methods have been proposed (Wang, Zhou et al. 2008, Poblete, Bustos et al. 2010, Zhang, Huang et al. 2010, Chiang 2013, Ballan, Uricchio et al. 2014, Murthy, Maji et al. 2015).

### III.3.1.5 Multi-label learning for image categorization

There exist another category of methods, which is analogous to supervised auto-annotation methods, and which are commonly known as multi-label image learning. Multi-label learning is defined as the task of constructing classification models for images assigned with multiple labels (Tsoumakas and Zhang 2009). For instance, (Zhang and Wu 2015) proposed an algorithm namely multi-label learning with Label specific FeaTures (LIFT). LIFT is based on clustering analysis of the negative as well as the positive instances (i.e., images) related to a specific label. In (Sun, Tang et al. 2014), sparse structure of label dependency was used in order to improve the task of multi-label image classification. They proposed to take into consideration only the most prominent correlations between labels and avoid over-fitting the multi-label classifier with the unnecessary correlations e.g., there is no correlation between ‘train’ and ‘bird’. Similarly, (Wang, Yan et al. 2009) proposed a multi-label sparse coding framework for automatic image annotation.

### III.3.2 Automatic image annotation based on unsupervised learning

Unsupervised learning is used to identify patterns within an unlabeled data i.e., identify the input data structure with lack of response measure (i.e., outcome) (Arbelaitz, Gurrutxaga et al. 2013). In other words, unsupervised learning aims to find out patterns and relationships (i.e., correlations) among data. Clustering and dimensionality reduction are typical examples of unsupervised learning. Relevance models are unsupervised generative models which estimate the joint probability distribution of image regions and semantic concepts, over a training set, in order to be capable to predict labels for unlabeled images (Zhou, Cheung et al. 2011). Hereafter, we provide details about some state of the art relevance models.
III.3.2.1 Co-occurrence model

One of the earliest efforts in the field of automatic image annotation back to the work proposed by (Mori, Takahashi et al. 1999). Given a set of images $T$, all the images firstly pre-processed by dividing them into tiles. After that, the relatedness probability of a label $L$ to a cluster $C$ is estimated according to Eq.(19)

$$p(L|C) = \frac{N(L,C)}{\sum_i N(L,C)}$$  \hspace{1cm} (19)

Such that $N(L,C)$ is the appearance frequency of $L$ with an image tile from $C$. The probability for an unannotated $I$ to be labeled with a concept $L$ is calculated using Eq.(20)

$$p(L|I) = \frac{1}{|I|} \sum_{t \in I} p(L|C_t)$$ \hspace{1cm} (20)

Where $C_t$ is the closest cluster to the tile $t$ from $I$, and $|I|$ is the number of tiles in $I$. Then, scores obtained by all labels are sorted in a deceasing order, and the top labels are assigned to $I$.

III.3.2.2 Machine Translation Model (TM)

Authors in this work (Duygulu, Barnard et al. 2002) believed that estimating relevance of labels with regions is more useful than that of labels with tiles. Hence, they constructed a visual vocabulary referred to as ‘blobs’. An illustration of the resulting blobs from a set of images is shown in Figure 10. In fact, blobs construction is analogous to tile construction, the only difference is that images are firstly segmented into regions instead of dividing them into tiles. Then, a machine translation model, originally proposed for linguistic translation, is applied to perform translation between ‘blobs’ and ‘keywords’.
III.3.2.3 Cross Media Relevance Model (CMRM)

In (Jeon, Lavrenko et al. 2003), a relevance model for automatic image annotation, referred to as Cross-Media Relevance Model (CMRM), is introduced. At the opposite to the work in (Duygulu, Barnard et al. 2002) which assume a one-to-one relationship between blobs and labels in an image, (Jeon, Lavrenko et al. 2003) assumes simply that for a given image, a set of labels \( L = \{l_1, l_2, \ldots, l_N \} \) is related to a set of blobs \( B = \{b_1, b_2, \ldots, b_M \} \). Thus, an image \( I \) within the training set \( T \) has a dual representation \( I = \{b_1, b_2, \ldots, b_M, l_1, l_2, \ldots, l_N \} \).

Given unannotated image, in order to predict labels for \( I \), the joint probability distribution of labels and blobs is estimated as expectation aver \( T \), as follows

\[
P(I|l_i) \approx P(I|b_1, b_2, \ldots, b_N) \tag{21}
\]

\[
P(I|b_1, b_2, \ldots, b_N) = P(I|l_i, b_1, b_2, \ldots, b_N) = \sum_{I \in T} P(I)P(I|l_i, b_1, b_2, \ldots, b_N), \tag{22}
\]

Under the assumption that the events of observing \( l_i \) and \( b_1, b_2, \ldots, b_N \) are independent, Eq.(22) is re-re-written as
\[ P(l_i, b_1, b_2, \ldots, b_N) = \sum_{i \in T} P(I)P(l_i|I) \prod_{j=1}^{N} P(b_j|I) \]  \tag{23}

Where \( P(I) \) is the prior probability which is kept uniform over \( T \), and \( P(l_i|I) \) and \( P(b_j|I) \) are estimated using the smoothed maximum likelihood.

### III.3.2.4 Continuous Relevance Model (CRM)

Results of (Jeon, Lavrenko et al. 2003) were subsequently improved by (Lavrenko, Manmatha et al. 2003), which proposed a model referred to as Continuous Relevance Model (CRM). Similarly to CMRM, CRM uses the joint probability distribution to predict labels for the unlabeled images. However, in CRM, images are represented with continuous feature vectors extracted from image regions instead of the blobs.

Formally, CRM calculate the joint probability distribution of a set of labels \( L = \{l_1, l_2, \ldots, l_N\} \) and a set of regions \( R = \{r_1, r_2, \ldots, r_M\} \) according to Eq.(24)

\[ P(L, R) = \sum_{i \in T} P(I) \prod_{l_i \in L} P(l_i|I) \prod_{r_i \in R} P(R|I) \]  \tag{24}

Where \( P(I) \) is kept uniform over, \( P(L|I) \) is estimated using a smoothed multinomial distribution and \( P(R|I) \) is non-parametric kernel-based density estimate.

### III.3.2.5 Methods incorporating semantic relatedness

The previous reported works (i.e., unsupervised methods) consider concept-image relationship, while they totally neglect that of concept-concept. Considering this latter relationship could significantly help in improving the annotation performance. The works of (Liu, Wang et al. 2007) and (Gong, Li et al. 2010) consider the concept-concept relationship along with that of concept-image. Figure 11 shows example of pairwise semantic similarity between concepts yielded by the work in (Gong, Li et al. 2010).
### III.3.3 Automatic image annotation based on semi-supervised learning

Semi-supervised learning falls between supervised and unsupervised learning. In this kind of learning, both labeled and unlabeled data are employed, where the amount of labeled data is relatively much less than the labeled data. Compared to the first two categories (i.e., supervised and unsupervised), a few number of semi-supervised methods can be found in the literature.

#### III.3.3.1 Annotation refinement and completion for image retrieval method

The work in (Wu, Jin et al. 2013) is one of the most representative semi-supervised methods. Commonly, the manual annotations provided by users are noisy, inconsistent and suffer from missing. Thus, improving the quality of manual annotations will certainly lead to enhance the retrieval results. The method in (Wu, Jin et al. 2013) targets mainly two aims, completing the missing labels and rectifying the noisy ones.

![Figure 12. Tag completion for auto-annotation method (Wu, Jin et al. 2013)](image-url)
To reach such an aim, a partially labeled image database is provided, and then the relation between the initial tags and the available images is represented by a matrix, referred to as tag matrix. Columns of the matrix represent tags, whereas, rows correspond to images, each entry of the matrix is assigned a value, either 0 or 1. The value 1 denotes the presence of the concept in the image, while the value 0 denotes the absence. The aim of the method is to achieve the optimal values for the matrix by considering both the visual similarity between images and the observed tags. Achieving the optimal matrix means that the missed tags are completed and the noisy ones are rectified. Hence, to solve this optimization problem, a new algorithm is proposed, so that after having the matrix completed, it can be exploited for text-based retrieval. Figure 12 illustrates the steps of the method.

III.4 Region-based approach

Region-based methods (Parashar 2009, Manipoonchelvi and Muneeswaran 2011, Zhang, Islam et al. 2012, Belloulata, ) assume that user interests with only a part of the image, whereas, it neglects the remaining parts. In the following, we report the details of two typical region-based methods which are recently proposed.

III.4.1 Image retrieval based on attention-driven salient edges and regions extraction

The aim of this work (Feng, Xu et al. 2010) is to effectively retrieving images based on capturing (i.e., understanding) the human visual perception of images. To do so, an improved saliency map, based on the existing selective visual attention models, was proposed. Authors justified the need for the saliency map by the fact that it contains useful information about the location of interesting parts of the image. Three images are generated from each image, namely, segmented image, edge image, saliency map image. The saliency value of a pixel is $x$ is calculated according to Eq.(25)

$$SP(x) = \sum_{l=1}^{L} \sum_{y \in \theta x} (y_{Cl}S_{Cl}^{l}(x,y) + y_{O}S_{O}^{l}(x,y))$$

(25)
Where \( y \) is a pixel belonging to the \( 3 \times 3 \) window centered at \( x \), namely \( \theta_x \). \( S^c_l(x,y) \) and \( S^o_l(x,y) \) are the color intensity contrast and orientation contrast between \( x \) and \( y \). \( \gamma_c \) and \( \gamma_o \) are weighting coefficients. \( l \) denotes the \( l^{th} \) level image in a Gaussian image pyramid that linearly combining the contrasts from each channel, where the number of levels equal to 3.

The saliency map image was then used along with the edge image and segmented image to generate the salient edge image and salient region image, respectively. Finally, Salient Edge Histogram Descriptor (SEHD) and Region Adjacency Graph (RVG) are extracted from each of salient edge and salient region images, respectively. After having SEHDs and RVGs matched to each other for each image, resulting values are linearly combined to measure the overall similarity between images. Figure 13 depicts the flowchart of the method.

![Figure 13. Flowchart of a region-based for image retrieval (Feng, Xu et al. 2010)](image)

**III.4.2 Adaptive region matching for region-based image retrieval**

In this work (Yang and Cai 2014), Region-Based Image Retrieval (RBIR) is performed after determining the meaningful regions, which could be the user target, within an image. At first, Region Importance Index (RII) is calculated for each region on the basis of both region percentage from the whole image and region position. Given the set of regions \( I = \{r_1, r_2, \ldots, r_N\} \) related to the image \( I \), RII is calculated according to Eq. (26)
\[ RII(r_i) = \frac{(r_i)_{area}}{A_{area}} \cdot (1 - V) \] (26)

Where

\[ V = \sqrt{\frac{(r_{ix} - x)^2 + (r_{iy} - y)^2}{\sqrt{L(A)^2 + H(A)^2}}} \] (27)

Such that \( \frac{(r_i)_{area}}{A_{area}} \) represents the percentage of the region \( r_i \), and \((r_{ix} - r_{iy})\) is the barycentric coordinates of the region \( r_i \), \((x,y)\) is central coordinates of an image \( A \), \( L(A) \) and \( H(A) \) are length and height of \( A \), respectively.

The remaining steps of the methods are the following:

- Taking an image \( I \) as a query, and segment the image using MS-Ncut.
- Calculate the RII for all regions of \( I \), for an \( r_i \in I \), if \( RII(r_i) \geq \text{Threshold} \), \( r_i \) is considered as Semantic Meaningful Region (SMR).
- If there exist SMR regions within \( I \), select the region having the maximum value of RII. Otherwise, extract visual features from all the regions of \( I \).
- Calculate distance between \( I \) and all the images using the Adaptive Region Matching (ARM), which is designed especially to deal with the problem of regions interference.
- Display the retrieved images after ranking them.

### III.5 Relevance feedback approach

Methods belonging to this approach attempt to iteratively understand the user intention by involving him in the retrieval process. Let us give an overview about certain methods adopting the relevance feedback technique so as to mitigate the semantic gap.
III.5.1 Relevance feedback for image retrieval by Combining SVM and GM

In this work (Marakakis, Siolas et al. 2011), Gaussian Models (GMs) were used for image representation and Support Vector Machines (SVM) were used for relevance feedback. Thus, the presented method is a combination of discriminative and probabilistic approaches. At first, Color-SIFT features were extracted from the database images, and then Bag of Features (BoF) technique was used to represent each image. Distribution of feature vectors that correspond to a particular image were modeled using a GM, where parameters are estimated using MAP-EM algorithm. The probability density function (Pdf) of a GM is given by

\[ P(x|\Theta) = \sum_{i=1}^{K} w_i G(x|\theta_j) \]  

(28)

Such that

\[ P(x|\Theta) = \sum_{i=1}^{K} w_i G(x|\theta_j) \]  

(29)

\[ \Theta = \{(\theta_j, w_i), j = 1, ..., k\}, \quad \theta_j = (\mu_j, \Sigma_j) \]  

(30)

\[ G(x|\theta_j) = \frac{1}{\sqrt{|\Sigma_j|}} \frac{1}{(2\pi)^M} e^{-\frac{1}{2}(x_i - \mu_j)^T\Sigma_j^{-1}(x_i - \mu_j)} \]  

(31)

Where \( x \) is a data point in the \( M \)-dimensional space, and the pair \((\mu_j, \Sigma_j)\) are the mean vector and the covariance matrix, respectively. An SVM classifier is trained in each round of the RF with the positive and the negative examples supplied by the user. In order to be able to incorporate images, which are represented with GMs, within the SVM, a kernel
function is employed. This kernel is based on an approximation of Kullback-Leibler divergence.

**III.6 Visual object recognition**

Object recognition is the task of automatically assigning a given object, contained in an image, to a set of pre-defined categories (Chen, Song et al. 2015). Numerous studies have investigated this task and several methods have been proposed (Zhu, Fu et al. 2010, Song, Chen et al. 2011, Choi, Torralba et al. 2012, Dehghani, Moloney et al. 2016, Shang, Chen et al. 2016). Certain methods have tried to recognize different types of visual objects (Zhang, Berg et al. 2006, Bosch, Zisserman et al. 2007, Bosch, Zisserman et al. 2007, Boiman, Shechtman et al. 2008, Yang, Yu et al. 2009), while certain others have focused on recognizing a specific type of objects such as in (Nilisback and Zisserman 2006, Nilisback and Zisserman 2008, Harish, Hedge et al. 2013, Elhariri, El-Bendary et al. 2014, Amlekar, Gaikwad et al. 2015).

The approach we propose in this thesis aims mainly to mitigate the semantic gap between image features and semantic concepts. However, as it has been already been mentioned, our proposed approach has also been successfully applied for the task of date fruit recognition. In the following, we provide an overview on the related work on date fruit recognition.

**III.6.1 Automatic date fruit recognition**

which are (Aiadi, Khaldi et al. 2016, Aiadi and Kherfi 2016), respectively. Let us take a look on those methods:

### III.6.1.1 Methods for automatic date fruit grading

![Computer-vision-based system for date grading](image)

**Figure 14.** A computer-vision-based system for date grading (Pourdarbani, Ghassemzadeh et al. 2015)

Several criterions were adopted for date grading, quality and maturity level are probably the two main criterions. Hereafter, we present details about methods opted for each of these criterions.

#### III.6.1.1.1 Date grading based on maturity level

The work of (Pourdarbani, Ghassemzadeh et al. 2015) aims to classify samples of the Berhee variety into three maturity levels. An online computer-vision-based system is designed for this purpose (Figure 14). A sample being classified is fed to the conveyor belt, and then a camera is used to take a snapshot for the sample. After having performed the necessary pre-processing operations, such as noise filtering and background subtraction, and extracting the sample features, the sample grade is determined based on its color properties. The sample is then orientated to the suitable port after the interface circuit receives a signal from the sensor.
III.6.1.1.2 Date grading based on quality

A sample quality can be measured according to certain factors, hardness degree and percentage of defects are among those factors. In (Manickavasagan, Al-Mezeini et al. 2014), linear discriminate analysis (LDA) and stepwise discriminate analysis (SDA) were used to grade dates into hard, semi-hard and soft (Figure 15). A total of 39 features (13 from each of the RGB channels), including color and texture, were used to describe samples. Figure 15 shows some samples in different hardness degree.

![Figure 15. Date samples in different hardness degree (Manickavasagan, Al-Mezeini et al. 2014)](image)

In (Mohana and Prabhakar 2015), date samples were graded into six grades depending on their hardness degree. A number of pre-processing operations were firstly conducted to remove reflection and to separate the fruit area from the background. A variety of features were used to represent sample images, LBP and primitive shape features are among those features. Three classifiers, namely k-nearest neighbors (KNN), support vector machines (SVM) and LDA were employed, while the first one has achieved the best performance.

Since that the external properties of date samples are among the crucial factors adopted by consumer in determining the fruit quality, (Al-Rahbi, Manickavasagan et al. 2013) have designed a system for detection of date defects, especially date cracks. In particular, samples were classified into three grades based on the percentage of cracks within the surface. Color and size features were used along with a Linear Discriminant Analysis (LDA) classifier to accomplish this task. In (Djeffal, Babahenini et al. 2010, Djeffal, Rais et al. 2012), a date sorting system is developed, where mean color, caliber and homogeneity features were used to describe date samples.

With the aim of speed up the process of date sorting, (Al Ohali 2011) has introduced a prototypical system that uses date defects as features for sorting. Two main defects were estimated, namely, bird flicks and bruises (Figure 16), the first one is estimated in terms of
color intensity, while the second one is estimated on the basis of the sample shape. A Back
Propagation Neural Network (BPNN) has been used to perform the classification process.

Figure 16. Two types of defects, bird flicks and bruises (Al Ohali 2011)

III.6.1.2 Methods for automatic date variety recognition

Compared to date grading, a very few number of methods have been proposed for
date variety recognition. Table 1 summarizes the techniques used by each of the proposed
methods for date variety recognition.

Table1. State of the art methods for date varieties recognition

<table>
<thead>
<tr>
<th>Features</th>
<th>(Muhammad 2015)</th>
<th>(Haidar, Dong et al. 2012)</th>
<th>(Fadel 2007)</th>
<th>(Hobani, Thottam et al. 2003)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weber Local Descriptor (WLD), Local Binary Patterns (LBP)</td>
<td>Color mean and standard deviation (from the 3 RGB channels), Area, Perimeter, Eccentricity, Major and Minor axis length, Entropy (from the 3 RGB channels) and Energy measures extracted from the Grey-Level Co-occurrence Matrix (GLCM)</td>
<td>Color mean and standard deviation, for each of the three RGB channels</td>
<td>physical measurements: length, diameter, weight, volume, moisture content, and water activity, and the color attributes.</td>
</tr>
<tr>
<td>Classifiers</td>
<td>Support Vector Machine (SVM)</td>
<td>Artificial Neural Network (ANN), K-Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA)</td>
<td>Probabilistic Neural Network (PNN)</td>
<td>Artificial Neural Network (ANN)</td>
</tr>
</tbody>
</table>
It should be noted that the method in (Hobani, Thottam et al. 2003) has utilized the physical measurements of dates, including moisture content and water activity. Indeed, using such measurements is impractical and makes very difficult to automate the method. In addition, (Haidar, Dong et al. 2012) has reported that the ANN has taken a great deal of time before yielding the optimal network. Such a high computational cost is, in fact, amongst the main barriers to meet the exigencies of a real-time recognition system.

III.7 Conclusion

In this chapter, we have focused our attention on the works attempted to narrow the semantic gap in CBIR. We have presented the different existing approaches, namely relevance feedback, region-based and automatic annotation. We have given details about several methods appertaining to each of these approaches. In addition, we have put the light on the studies investigating the task of visual object recognition, as the proposed approach has successfully been applied to the task of object recognition, especially date fruit recognition.
Chapter IV. A SUPERVISED PROBABILISTIC-BASED APPROACH TO FIGHT AGAINST THE SEMANTIC GAP IN CBIR

IV.1 Introduction

As it has previously shown, mitigating the semantic gap in an urgent need, particularly for the crucial and critical applications. To nowadays, the semantic gap remains an open issue in spite of the considerable amount of efforts that have been devoted to reduce it. Over the last decades, numerous approaches have been proposed for this end, as it has been shown in the previous chapter. Unfortunately, the yielded results are still unsatisfactory and a lot of work remains to be done, so that the competition is still opened in order to achieve a convincing performance.

In the rest of this chapter, we present our proposed approach for narrowing the semantic gap. At first, we start by giving an overview on the proposed approach as well as presenting the major contributions we make. Because we make use of several and diverse techniques, we review each of them in details.
IV.2 Overview on the proposed approach

In the present thesis, we propose a supervised probabilistic-based approach for bridging the semantic gap in CBIR. Because the semantic gap occurs because of the discrepancy of semantic concepts and visual features, we suggest, therefore, to map each semantic concept with its visual representation (i.e., visual model). In other words, the aim is to identify and model, for each concept, the common visual features describing its relevant images, in a way that allows to accurately locate the pertinent images of the concept during retrieval.

However, certain challenges have to be confronted in order to successfully reach this aim, which are:

1. Training images
   Because our approach utilizes supervised learning techniques, a sufficient number of training images have to be available for all the concepts being modeled.

2. Feature selection
   A large diversity of image features have been proposed, what is the most appropriate combination of features allowing to faithfully describe the image content.

3. Outlier images
   Some training image sets comprise outliers, which could distort the visual model, and thus decreasing the retrieval performance. The challenge is how to effectively prune these sets by removing outliers and keeping inliers.

4. Modeling method
   By considering the intra-variation, in terms of visual features, of images representing the same semantic concept, it is challenging to take this into consideration in the modeling process. For instance, images representing the concept “Apple” have several visual appearances, there exists “red apples”, “Green apples” and “Yellow apples”.

51
5. Computational cost

The response time is another issue related to image retrieval in general, most users cannot wait for a long deal of time to get the retrieval results. So that, it is quite important to take into account such an important aspect.

6. Degree of human intervention

Users tend to use systems in which they aren’t required to intervene in yielding results. Certain approaches such as those adopting the relevance feedback technique, involve the user in the retrieval process, which could be boring and time-consuming. Hence, designing fully automatic retrieval systems is highly recommended.

In the proposed approach, we conduct a supervised learning process which is composed of two main stages, training and retrieval. At first, we associate each concept with a set of images in which the concept is present, we refer to the set of images associated with the concept \( C_p \) as \( T_{C_p} \). Those images are gathered from the web using web search engines. With the aim of faithfully describing images, we extract diverse local and global visual features from the images. As web images could probably contain outliers, we use outlier detection methods in order to prune them. Visual appearances related to \( C_p \) are identified by clustering each \( T_{C_p} \), where each cluster includes images that have the same visual appearance i.e., visually similar images. Because the number of visual appearances may vary from a concept to another, Clustering Validity Indices (CVIs) are employed to automatically and accurately determine the number of clusters. For each \( C_p \), Probability density functions (Pdf) that correspond to the clusters of \( T_{C_p} \) are fused in a Gaussian Mixture Model (GMM) representing the concept’s visual model. We use, then, Expectation-Maximization (EM) algorithm for parameters estimation. During retrieval, user can naturally express his needs using a textual query; relevant images are the ones maximizing the posterior probability to the GMM corresponding to the query. Figure 17 shows the flowchart of the proposed approach.
From the description of our approach, it is apparent that we have taking into account all the issues we have cited. Our modeling way takes into account the intra-variation of images representing the same concept and the presence of outliers within training sets as well. The proposed approach has the advantage of retrieving images using a textual query in spite of the lack of textual descriptions with those images. Besides, it isn’t restricted to any predefined concept, as training images are automatically gathered from web search engines. In addition, it is fully automatic, as it doesn’t require any kind of human intervention. In addition, it is computationally fast, as we will show in the next chapter.
CHAPTER IV. A SUPERVISED PROBABILISTIC APPROACH TO FIGHT AGAINST THE SEMANTIC GAP

IV.3 Steps of the proposed approach for bridging the semantic gap

IV.3.1 Training images gathering

   To perform training, we assign each semantic concept with a set of images $T_{cp}$ that are labeled with this concept. Numerous images are available either in our personal collections or in the web. To facilitate our task and in order to fully automate the proposed approach, we leverage those images from the web. Nowadays, thanks to web search engines, an immense quantity of labeled images is available and accessible by users in the internet. From these engines, we selected Google Images, which assures a high accuracy in retrieval, as a source of the images. For each semantic concept, the top retrieved images are automatically downloaded and stored to be used for the training. Figure 18 shows the top retrieved images from Google Images for some semantic concepts.

IV.3.2 Features extraction

   Image representation is intrinsic component which extremely influences the performance of CBIR systems and the proposed methods for reducing the semantic gap as well. Image representation is very important in many other interesting applications such as robotics. Over the last decades, numerous traditional and sophisticated features have been proposed, and the competition is still opened to achieve the features that are capable to faithfully reflect the image semantics. In the following, we present the features we used in our approach.
IV.3.2.1 Color features

Color is one of the most important features describing images, and it is widely used in the CBIR systems. In the literature, several color descriptors were proposed (Swain and Ballard 1991, Flickner, Sawhney et al. 1995, Pass and Zabih 1996, Huang, Kumar et al. 1997), let us take a look on some of them:

IV.3.2.1.1 Color histogram

RGB-based color histogram represents the percentage of each color, from the RGB space, within an image. It has been used in many CBIR systems (Han and Ma 2002, Kherfi, Ziou et al. 2003). Taking into account that values of each channel range from 0 to 255, this, will therefore results in $256^3 = 16777216$ colors. Using a histogram with such a dimension is computationally expensive and impractical as well because of the “curse of the dimensionality” (Huang and Tran 2016). To cope with that, a quantization process, which consists in subdividing each of the three channels into a number of ranges, should be performed. For example, if each channel is divided into three ranges, this will yield a
CHAPTER IV. A SUPERVISED PROBABILISTIC APPROACH TO FIGHT AGAINST THE SEMANTIC GAP

histogram of 27 dimensions. Figure 19 shows three images and the 27-dimensions histograms extracted from each of them.

![Figure 19. RGB-based color histogram for three images](image)

Color histogram presents a multitude of advantages: it is straightforward, efficient, ease to be implemented and invariant to image size. However, it lacks the ability of describing color distribution within the image.

IV.3.2.1.2 Color statistical moments

Color moments are simple yet efficient features (ping Tian 2013) that are used in many CBIR systems (Vailaya, Figueiredo et al. 2001, Fan, Gao et al. 2004, Feng, Manmatha et al. 2004). The most common moments are mean ($\mu_t$), standard deviation ($\sigma_t$) and skewness($\gamma_t$) which are given by Eq.(32), Eq.(33) and Eq.(34)
\[ \mu_i = \frac{1}{M} \sum_{j=1}^{M} V_{ij} \]  

(32)

\[ \sigma_i = \left( \frac{1}{M} \sum_{j=1}^{M} (V_{ij} - \mu_i)^2 \right)^{\frac{1}{2}} \]  

(33)

\[ \gamma_i = \left( \frac{1}{M} \sum_{j=1}^{M} (V_{ij} - \mu_i)^3 \right)^{\frac{1}{3}} \]  

(34)

where \( V_{ij} \) is the color value of the \( i^{th} \) channel of the \( j^{th} \) image pixel and \( M \) is the sum of pixels in the image.

**IV.3.2.2 Texture features**

Texture may be defined as what constitutes a particular region, where the region’s structure consists in repetitive patterns that contain elements that are arranged on the basis of certain placement rule (Tamura, Mori et al. 1978). Numerous features have been proposed for texture description, they can roughly be classified into three approaches (Khalidi and Kherfi 2016), statistical (Haralick and Shanmugam 1973, Tamura, Mori et al. 1978), frequency (Daugman 1980, Daugman 1985) and geometrical (Zhu, Guo et al. 2005, Fan, Li et al. 2008). Figure 20 shows examples of texture images. Following, we provide explanations for texture features we used in this thesis.

**Figure 20.** Examples for texture images from the Outex dataset (Ojala, Maenpaa et al. 2002)
IV.3.2.2.1 Gray-Level Co-occurrence Matrix (GLCM)

The principle of the Gray-Level Co-occurrence Matrix (GLCM) (Haralick and Shanmugam 1973) is to count the appearance frequency of gray-level pairs within the image. Formally, suppose we are given a GLCM, referred to as $M$, then, each entry $M(i,j)$ represents the occurrence frequency of a gray-level $j$ followed by a gray-level $i$ according to the spatial relationship, denoted as $S(\Delta x, \Delta y)$.

Given an image $I$ and a spatial relationship $S(\Delta x, \Delta y)$, $M$ is calculated according to Eq.(35)

$$M(i,j) = \sum_{p=1}^{H} \sum_{q=1}^{W} \begin{cases} 1, & \text{if } I(p,q) = i \text{ and } I(p+\Delta x, q+\Delta y) = j \\ 0, & \text{otherwise} \end{cases}$$

(35)

where $H$ and $W$ are respectively the height and the width of $I$.

After having calculated $M$, several features can be extracted from this matrix, including the following:

$$Contrast = \sum_i \sum_j (i-j)^2 M(i,j)$$

(36)

$$Correlation = \sum_i \sum_j (i-\mu)(j-\mu) M(i,j)$$

(37)

$$Energy = \sum_i \sum_j M(i,j)^2$$

(38)

$$Homogeneity = \sum_i \sum_j \frac{1}{1+(i-j)^2} M(i,j)$$

(39)
CHAPTER IV. A SUPERVISED PROBABILISTIC APPROACH TO FIGHT AGAINST THE SEMANTIC GAP

IV.3.2.2.2 Local Binary Patterns (LBP)

Local Binary Patterns (LBP) (Ojala, Pietikainen et al. 2002) is a powerful feature which is widely used in several applications such as face recognition (Ahonen, Hadid et al. 2006). Figure 21 shows an example of the LBP calculation. The basic LBP operator is calculated by assigning each pixel within the image with a binary code of eight bits. This code is computed by considering the center pixel of a 3x3 neighborhood as a threshold value. The decimal value that corresponds to the binary one is then used to replace the original value of the center pixel. The appearance frequency of each code, called pattern, in the image is then computed to construct a histogram of 256 ($2^8$) dimensions. According to (Ojala, Pietikainen et al. 2002), there exist 58 amongst the 256 pattern that provide more information than others, which make possible to use only a small subset of patterns to describe image texture. These patterns are called uniform and contain at most two contiguous of bit suits. Later on, several LBP variants were proposed, an experimental evaluation of 13 variants has been conducted in (Hadid, Ylioinas et al. 2015).

![Local Binary Patterns Calculation Example](image)

**Figure 21.** Example of LBP calculation

IV.3.2.2.3 Weber Local Descriptor (WLD)

WLD (Chen, Shan et al. 2010) represents texture images as a histogram of differential excitation and orientation. Such that, the differential excitation is the function of the ratio between the intensity of the current pixel and the intensity differences of this pixel against its neighbors. The orientation component is the gradient orientation of the current pixel. WLD uses these two components to construct a concatenated histogram. WLD extracts for each pixel $I_c$, in a given image, two measures which are differential excitation $\xi(I_c)$ and orientation $\theta(I_c)$. These two measures are calculated according to Eq.(40) and Eq.(41), respectively.
\[ \xi(I_c) = \arctan \left( \frac{1}{p-1} \sum_{i=0}^{p-1} \left( \frac{I_i - I_c}{I_c} \right) \right) \] (40)

\[ \theta(I_c) = \text{median} \left( \frac{I_{R(i+4)} - I_i}{I_{R(i+6)} - I_{i+2}} \right) \] (41)

Such that \( i \in \{0, 1, 2, 3, \ldots, p - 1\} \) are indices of \( I_c \)'s neighbors.

Figure 22. Different steps to calculate the WLD (Chen, Shan et al. 2010)

After having \( \xi(I_c) \) and \( \theta(I_c) \) extracted from all pixels of the image, WLD subsample the value space then map each differential excitation and orientation to its appropriate bin as it is illustrated in Figure 22. The result is, therefore, a matrix \( M \) where each entry \( M(i,j) \) represents the occurrence frequency of a pixel that has value \( i \) as orientation and \( j \) as differential excitation. Finally, by combining all rows of this matrix (i.e. each row represents a
CHAPTER IV. A SUPERVISED PROBABILISTIC APPROACH TO FIGHT AGAINST THE SEMANTIC GAP

sub-histogram of some specific orientation), we get the final WLD histograms. Figure 23 shows the WLD histograms extracted from three different images.

![Figure 23. Example of WLD histograms for three different images](image)

IV.3.2.3 Shape features

Shape features are also quite important feature to be taken into consideration, especially for pattern recognition. In the following, we present some shape features

IV.3.2.3.1 Hu invariant moments

Hu moments (Hu 1962) are shape features which are invariant to scale, position and rotation. The geometric moment of \((p + q)\) order of an image \(I_{xy}\) is given by Eq.(42)

\[
m_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} x^p y^q * I_{xy}
\]

(42)
where $M$ and $N$ are the image dimensions and $I_{xy}$ denotes the value of the pixel has the coordinates $x$ and $y$.

The center of mass of $I_{xy}$ is defined by Eq.(43)

$$\overline{x}, \overline{y} = \left( \frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right)$$

(43)

The centralized moments of $I_{xy}$ are invariant under position. Those moments are given by Eq.(44)

$$\mu_{pq} = \sum_{x=1}^{M} \sum_{y=1}^{N} (x - \overline{x})^p * (y - \overline{y})^q * I_{xy}$$

(44)

The normalized central moments are scale invariant, they are defined as in Eq.(45)

$$\eta_{pq} = \frac{\mu_{pq}}{\mu_{00}^\gamma}$$

(45)

where $\gamma = \frac{p+q}{2} \forall p + q \geq 2$

The seven moments of Hu are given with

$$I_1 = \eta_{20} + \eta_{02}$$

(46)

$$I_2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2$$

(47)

$$I_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2$$

(48)

$$I_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} - \eta_{03})^2$$

(49)
\[ I_5 = (\eta_{30} + 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{21} - \eta_{03})^2 + (\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \] (50)

\[ I_6 = (\eta_{20} + \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})] \] (51)

\[ I_7 = (3\eta_{12} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \] (52)

Figure 24 shows the Hu moments extracted from the image at top and the same image rotated with 180 degree. We can notice that the values are the same, which confirm the invariance of Hu moments to the rotation.

**Figure 24.** Hu moments extracted from an image and the same image rotated with 180 degree

### IV.3.2.3.2 Basic geometrical shape features

There exists several geometrical features, here are some of them:

1. Minor axis length: it is the longest diameter within the shape.
2. Major axis length: it is the shortest diameter within the shape; it is perpendicular to the major axis.
3. Eccentricity: it is the ratio of the major axis to the minor axis.
4. Area: it is the number of pixels constituting the shape surface.
5. Perimeter: it is the number of pixels at the margin of the shape.
In order to calculate the minor, major axis length and eccentricity, we use the principal axes method. First, canny edge detector is applied to each image, where images are converted to gray level before feeding them to the detector. After that, the covariance matrix describing the spread of edge points is calculated, where the resulting matrix is of 2×2 dimensions.

Let \( C \) be a covariance matrix of the form:

\[
C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}
\]  

The eigenvalues of \( C \), denoted as \( \lambda \), can be calculated as in Eq.(54).

\[
(C - \lambda I) \cdot \nu = 0
\]

such that \( \nu \) is non-zero vector and \( I \) is the identity matrix.

Taking into consideration that \( \nu \) is non-zero vector, Eq.(54) is satisfied if \( |C - \lambda I| = 0 \)

\[
|C - \lambda I| = \begin{vmatrix} C_{11} - \lambda & C_{12} \\ C_{21} & C_{22} - \lambda \end{vmatrix} = (C_{11} - \lambda)(C_{22} - \lambda) - C_{12}C_{21} = 0
\]

Thus,

\[
\begin{align*}
\lambda_1 &= \frac{1}{2} \left( C_{11} + C_{22} + \sqrt{(C_{11} + C_{22})^2 - 4(C_{11}C_{22} - C_{12}C_{21})} \right) \\
\lambda_2 &= \frac{1}{2} \left( C_{11} + C_{22} - \sqrt{(C_{11} + C_{22})^2 - 4(C_{11}C_{22} - C_{12}C_{21})} \right)
\end{align*}
\]

The highest value from \( \lambda_1 \) and \( \lambda_2 \) corresponds to the major axis length, while the lowest one corresponds to the minor axis length (Figure 25). The eccentricity is given by Eq.(57).
\[ E = \frac{\lambda_1}{\lambda_2} \]  

(57)

**Figure 25.** Major and minor axis length for a date sample

### IV.3.2.4 Keypoint-based features

We use SIFT features (Lowe 2004) to extract and describe information of key-points within the image. Given an image \( I \), SIFT calculation consists mainly in four steps:

1. Constructing the scale space
   1.1- Gaussian blurring
      To get rid of some details in the image, the Gaussian filter is used to blur \( I \). \( I \) is blurred successively, then resized (to the half size) and successively blurred again.
   1.2- LoG estimation
      Laplacian of Gaussian (LoG) is useful to detect the key-points in \( I \) because applying LoG permit to show the image corners and edges. Nevertheless, calculating LoG for several images (in the different scales) is computationally intensive. One way to reduce such a cost is by calculating the Difference of Gaussian (DoG) between each two nearby scales in the space (see Figure 26).
2. Key-points localization

Once DoG are calculated, we need to determine the maxima and the minima within each scale (except the first and the last scale). Simply, a pixel is maxima (respectively, minima) if it is greater (respectively, least) than its 26 neighbors (Figure 27).

Then, Taylor series expansion of scales is used to obtain more precise location of the extrema (i.e., maxima or minima). The extrema having an intensity less than a specific threshold are rejected.
3. Assigning orientations for the key-points

Assigning an orientation to each key-point will assure a highly desirable property which is the invariance to rotation. This can be reached by calculating the gradient directions and magnitudes around each key-point. Then assign the key-point with the most prominent orientation.

4. Calculating the key-points descriptors

First, we consider a 16x16 window around each key-point, where this window is composed of eight (8) windows, each is of 4x4. For each 4x4 window, we calculate the gradient magnitude and orientations. By performing a simple binning process, the obtained orientations are put in a histogram of 8 bins, where the value added to each bin is relative to the magnitude of the gradient. Doing that for all the 4x4 windows, we get a histogram (i.e., feature vector) of 128 dimensions (i.e., 4x4x8) describes the key-point.

Because the number of key-points may differ from an image to another, the dimension of SIFT features will be unequal for two different images. We use the Bag of Words (BoW) technique to quantize the SIFT features and extract a feature vector with the same dimension for all the images. The principle of BoW is illustrated in Figure 28.
It is worth noting that we cluster the images into 100 clusters, thus each image will be represented with a feature vector of 100 dimensions.

**IV.3.3 Dimensionality reduction**

The dimension of the used features is relatively high, in order to reduce features dimension and pick up the most distinguishing ones, we use the Principal Component Analysis (PCA) method (Pearson 1901). Algorithm 1 demonstrates the steps of reducing the dimensionality using PCA.

---

**Figure 28.** Image representation using Bag of Words (BoW)

- Key-points extraction
- Clustering
- Calculate the centroids of clusters (visual words)
- Key-points extraction
- Assign each key-point to the nearest cluster
- Compute the number of points assigned to each cluster and extract the histogram of the image

---

68
Algorithm 1 Principal Component Analysis (PCA)

**Input:** $X$: Raw data of $N \times M$ dimensions, $D$: new dimension,

**Output:** $X_{new}$: New data with a reduced dimension

**Begin**

Calculate the mean of $X$

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Centralize the data on $\bar{X}$

$$X' = X - \bar{X}$$

Calculate the covariance matrix $C$, covariance of two variables $A$ and $B$ from $X'$ having $\bar{A}$ and $\bar{B}$ as mean, is given by

$$\text{Cov}(A, B) = \frac{1}{N-1} \sum_{i=1}^{N} (A_i - \bar{A})(B_i - \bar{B})$$

Calculate the eigenvalues of $C$ by solving the following equation

$$P(\lambda) = |C - \lambda I|$$

Calculate the eigenvector $v$ corresponds to each $\lambda$ by solving the following the equation

$$Cv = \lambda v$$

Sort the eigenvalues in decreasing order and keep the top $D$ ones having the highest values, and sort the eigenvectors accordingly, where $V$ denotes the set of top eigenvectors. Project $X'$ in the new reduced space

$$X_{new} = X' \times V$$

**End**
IV.3.4 Outlier detection

As reported above, training images could include some outliers which could negatively affect the learning process. Hence, in order to avoid falling in such a problem, we prune each training set $T_{cp}$ from the outlier ones. For this end, we use the Pruning Local Distance-based Outlier Factor (PLDOF) method proposed in (Pamula, Deka et al. 2011), as it is computationally fast. The core idea of the PDLOF is the calculation of a measure, which is called Local Distance-based Outlier Factor (LDOF), for some specific images. A high LDOF value for a sample indicates that it deviates from its neighbors and it is most likely to be an outlier.

Formally, let $X = \{x_i, i = 1, \ldots, T\}$ be a set of data points in the M-dimensional space. Let $N_{x_i}$ be the set of K-nearest neighbors of $x_i$, LDOF is defined as in Eq.(58)

$$LDOF(x_i) = \overline{d_p}/\overline{D_p}$$

where $\overline{d_p}$ represents the average distance from $x_i$ to the data points in $N_{x_i}$, it is given by Eq.(59).

$$\overline{d_p} = \frac{1}{|N_{x_i}|} \sum_{q \in N_{x_i}} d(x_i, q)$$

where $|N_{x_i}|$ is the magnitude of $N_{x_i}$ and $d(x_i, q)$ is the distance between $x_i$ and the data point $q \in N_{x_i}$.

$\overline{D_p}$ represents the average distance among data points in $N_{x_i}$, it is defined by Eq.(60).

$$\overline{D_p} = \frac{1}{|N_{x_i}|(|N_{x_i}| - 1)} \sum_{q, q' \in N_{x_i}, q \neq q'} d(q, q')$$

The remaining steps of the algorithm are summarized in the Algorithm 2.
Algorithm 2 Pruning Local Distance-based Outlier Factor (PLDOF)

Input: $X$: The set of data points, $Nc$: number of clusters, $it$: number of iterations, $N$: number of outliers,

Output: $N$ outlier points

Begin

\[ Y \leftarrow k\text{-means}(Nc, it, X) \]

for each cluster $C_j \in Y$ do

\[ \text{Radius}_j \leftarrow \text{radius}(C_j) \]

end for

if $|C_j| > N$ then

for each point $x_i \in C_j$ do

if $d(x_i, o_i) < \text{Radius}_j$ then

\[ \text{prune}(x_i) \]

else

Add $x_i$ to $U$

end if

end for

else

Add $x_i$ to $U$

end if

end for

end if
CHAPTER IV. A SUPERVISED PROBABILISTIC APPROACH TO FIGHT AGAINST THE SEMANTIC GAP

end if

for each point \( x_i \) to \( U \) do

calculate \( LDOF(x_i) \)

end for

Sort the data points according to their \( LDOF(x_i) \) values. Top \( N \) points with highest \( LDOF(x_i) \) values are considered as outliers

End.

where \( o_i \) and \( Radius_j \) are the centroid of the cluster \( C_j \) and its radius, respectively. Figure 29 show some images representing 'potato', which are detected as outliers.

![Figure 29. Images from the class 'potato', which are detected as outliers using PLDOF](image)

**IV.3.5 Detecting the optimal number of appearances**

In order to identify the different visual appearances within each set \( T_{Cp} \), we cluster those sets based on the features previously extracted. Because we use the k-means algorithm for clustering, the number of clusters is required to be supplied as an input parameter. To automatically and accurately determine the number of clusters, we employ the Clustering Validity Indices (CVIs). CVIs methods attempt to find the clustering solution in which the clusters are compact and well separated. In other words, the objective is to minimize the within cluster variance and maximize the between cluster variance. Hence, alternative clustering solutions with different input values for the number of clusters are compared. The maximum (or minimum) value of the index indicates the best solution. Calinski-Harabasz
(Caliński and Harabasz 1974) and Davies-Bouldin (Davies and Bouldin 1979) are two well-known indexes, while the first has shown the best performance in the experiments of Miligan and Cooper (Miligan and Cooper 1985).

IV.3.5.1 Calinski-Harabasz index

Formally, let us define a set \( X = \{x_i, i = 1, \ldots, N\} \) where \( x_i \) is a data point in the \( M \)-dimensional space \( R^M \). By clustering the set \( X \) into \( K \) clusters we get the set of clusters \( \mathcal{C} = \{c_k, k = 1, \ldots, K\} \). Let us denote the centroid of the cluster \( c_k \) by \( \bar{c}_k \) and the centroid of \( X \) by \( \bar{X} \). Then, \( \bar{c}_k \) is given by Eq.(61).

\[
\bar{c}_k = \frac{\sum_{i=1}^{N} x_i \mathbb{1}_{x_i \in c_k}}{|c_k|} \quad (61)
\]

Where \( |c_k| \) is the magnitude of cluster \( c_k \). \( \bar{X} \) is given by Eq.(62).

\[
\bar{X} = \frac{\sum_{i=1}^{N} x_i \mathbb{1}_{x_i \in X}}{N} \quad (62)
\]

Calinski-Harabasz index (CH) is given by Eq.(63).

\[
CH(K) = \frac{(SS_B/(K - 1))}{(SS_W/(N - K))} \quad (63)
\]

where \( SS_B \) denotes the overall between-cluster variance.

\[
SS_B = \sum_{k=1}^{K} |c_k| \|\bar{c}_k - \bar{X}\|^2 \quad (64)
\]
and $SS_W$ the overall within-cluster variance.

$$SS_W = \sum_{k=1}^{K} \sum_{i=1}^{N} \|x_i - \bar{c}_k\|^2$$  \hspace{1cm} (65)$$

For each training set $T_c$, CH index is calculated for $K = \{K_{\text{min}}, \ldots, K_{\text{max}}\}$. The k-means algorithm was used for clustering, based on the features previously extracted. The optimal number of clusters $K'$ is the one that maximizes the CH index. At the end of this step, for each concept $Cp$, we get the set of clusters $C'_{cp} = \{c'_{cpk}, k = 1, \ldots, k'_{cp}\}$ where each cluster includes visually similar images.

**IV.3.5.2 Davies-Bouldin index**

Davies-Bouldin index is defined as Eq. (66).

$$DB(K) = \frac{1}{K} \sum_{i=1}^{K} V_i$$  \hspace{1cm} (66)$$

where

$$V_i = \max_{j=1, \ldots, K, j \neq i} \left( \frac{\delta_i + \delta_j}{d(\bar{c}_i, \bar{c}_j)} \right)$$  \hspace{1cm} (67)$$

where $\delta_i$ and $\delta_j$ are the average distance of all data points in the clusters $c_i$ and $c_j$ to their cluster centroids $\bar{c}_i$ and $\bar{c}_j$, respectively, and $d(\bar{c}_i, \bar{c}_j)$ is the distance between these two latter centroids. $\delta_i$ (respectively $\delta_j$) is given by Eq.(68).
\[ \delta_i = \frac{\sum_{t=1}^{c_i} \| x_t - \bar{c}_i \|}{|c_i|} \] (68)

and \( d(\bar{c}_i, \bar{c}_j) \) is given by Eq.(69).

\[ d(\bar{c}_i, \bar{c}_j) = \| \bar{c}_i - \bar{c}_j \| \] (69)

For each training set \( T_c \), DB index is calculated for \( K = \{K_{min}, ..., K_{max}\} \), where the k-means algorithm is used for clustering. The minimum value of the index indicates the preferred solution. At the end of this step, for each concept \( C_p \), we get the set of clusters \( C'_{cp} = \{c'_{cpk}, k = 1, ..., k'_{cp}\} \) where each cluster includes visually similar images. Figure 30 shows the detected visual appearances for the semantic concept ‘potato’.

**Figure 30.** Detected visual appearances of the semantic concept ‘potato’
IV.3.6 Normality test using Mardia test

Since visually similar images are grouped in the same cluster, we assume that images within each cluster are normally distributed. We check this assumption using Mardia’s multivariate skewness and kurtosis tests (Mardia 1970).

Formally, suppose we are given a cluster \( c'_{cp_k} = \{x'_i, i = 1, \ldots, P \} \), where \( x'_i \) is a \( M \)-dimensional data point that belong to \( c'_{cp_k} \). Let \( \mu \) and \( \Sigma \) denote the mean of samples belonging to \( c'_{cp_k} \) and the samples covariance matrix, respectively. \( \mu \) and \( \Sigma \) are defined by Eq.(70) and Eq.(71):

\[
\mu = \frac{1}{P} \sum_{i=1}^{P} x'_i \tag{70}
\]

\[
\Sigma = \frac{1}{P} \sum_{i=1}^{P} (x'_i - \mu)(x'_i - \mu)' \tag{71}
\]

Mardia multivariate skewness and kurtosis are defined as follows:

\[
\beta_{2,M} = \frac{1}{P} \sum_{i=1}^{P} (x'_j - \mu)' \Sigma^{-1} (x'_j - \mu)^2 \tag{72}
\]

\[
\beta_{1,M} = \frac{1}{P^2} \sum_{i=1}^{P} \sum_{j=1}^{P} (x'_j - \mu)' \Sigma^{-1} (x'_j - \mu)^3 \tag{73}
\]

For a multivariate normal distribution:

1- The test statistic:

\[
Z_{1,M} = \frac{P}{6} \beta_{1,M} \tag{74}
\]
follows a chi-square distribution with $f = \frac{1}{6} M(M + 1)(M + 2)$ degrees of freedom.

2- The test statistic:

$$Z_{1,M} = \frac{\beta_{1,M} - M(M + 2)}{\sqrt{8 M(M + 2)/P}}$$

follows a standard normal distribution.

IV.3.7 Modeling of semantic concepts using Gaussian Mixture Models

Each cluster within $C'_v$ is normally distributed. Therefore, it is described by the probability density of an M-dimensional normal function (pdf) of the form:

$$G(x_i|\theta_k) = \frac{1}{\sqrt{|\Sigma_k|}} \frac{1}{(2\pi)^M} e^{-\frac{1}{2}(x_i - \mu_k)^T\Sigma^{-1}(x_i - \mu_k)}$$

where $x$ is M-dimensional data vector that belongs to the set $X$, and $\theta_k$ denotes the parameters of the Gaussian distribution that corresponds to the $k^{th}$ cluster, these parameters are:

$\mu_k$: The mean vector of the points belonging to the cluster.

$\Sigma_k$: The covariance matrix.

Probability density functions (pdf)$G_{Cp} = \{G_{Cp_k}, k = 1, ..., k'_{Cp} \}$, which correspond to the set of clusters $C'_{Cp} = \{c'_{Cp_k}, k = 1, ..., k'_{Cp} \}$ related to a concept $Cp$, are combined in a Gaussian Mixture Model (GMM) representing the visual model of $Cp$. Thus, our model is a GMM, which is a weighted sum of $k'_{Cp}$ component Gaussian densities, it is given Eq.(77) and Eq.(78)

$$P_{Cp}(x_i|\theta_{Cp}) = \sum_{k=1}^{k_{Cp}} w_k G (x_i/\theta_{Cp_k})$$
\[
\theta_{\text{cp}} = \{\theta_{\text{cp}k}, k = 1, \ldots, k'_{\text{cp}}\} \tag{78}
\]

where \( w_k \) denotes the weight of the \( k^{th} \) distribution, and \( \theta_{\text{cp}} \) the parameters of Gaussian component densities of the mixture \( P_{\text{cp}} \). The likelihood of the data that belong to \( C'_{\text{cp}} \) is given by:

\[
L = \prod_{i=1}^{N_{\text{cp}}} P_{\text{cp}}(x_i|\theta_{\text{cp}}) \tag{79}
\]

where \( N_{\text{cp}} \) denotes the number of data points that belong to \( C'_{\text{cp}} \).

Expectation-Maximization (EM) algorithm is then used to maximize \( L \) and estimate the parameters of \( \theta_{\text{cp}} \). The EM algorithm has been successfully applied to solve pattern recognition problems (Boussellaa, Bougacha et al. 2009, Isupova, Mihaylova et al. 2015) because it is computationally fast (Leung, Liang et al. 2009), doesn’t require a large storage space (Wu 1983), and each iteration of the algorithm increases the likelihood until a local maximum is achieved (Moon 1996).

Expectation step:

\[
y_{ij} = \frac{\frac{w_j}{\sqrt{\sum (2\pi)^m}} e^{-\frac{1}{2}(x_i - \mu_j)^T \Sigma^{-1} (x_i - \mu_j)}}{\sum_{i=1}^{N_{\text{cp}}} y_{ij}} \tag{80}
\]

Maximization step:

\[
w_j = \frac{\sum_{i=1}^{N_{\text{cp}}} y_{ij}}{N_{\text{cp}}} \tag{81}
\]

\[
\mu_j = \frac{\sum_{i=1}^{N_{\text{cp}}} x_i y_{ij}}{\sum_{i=1}^{N_{\text{cp}}} y_{ij}} \tag{82}
\]
\\sum_{i=1}^{N_{cp}} y_{ij} \left( x_i - \mu_j \right) \left( x_i - \mu_j \right)^T \sum_{i=1}^{N_{cp}} y_{ij} \\

These two steps are repeated for \( i = 1, \ldots, N_{cp} \) and \( j = 1, \ldots, k'_{cp} \) until convergence.

**IV.3.8 Image retrieval**

The previous steps represent the learning process which is accomplished offline, where the outcomes are the visual models corresponding to the different semantic concepts. Suppose we are given a database of images, denoted as, and user wants to retrieve images relevant to a particular concept \( Cp \). The user can naturally formulate his query using a textual query (i.e., semantic concept). The retrieval process is carried out according to Algorithm 3.

**Algorithm 3** Image retrieval procedure

**Input:** \( Db \): Database of images, \( q \): Textual query, \( P_q \) having the parameters \( \theta_q = \{ \theta_{q_1}, \ldots, \theta_{q_{k'}} \} \): the probability density function (pdf) related to \( q \).

**Output:** relevant images to \( q \)

**Begin**

**For** each image \( I \in Db \) represented by the feature vector \( x_I \) **do**

Calculate \( P_q(x_I | \theta_q) = \sum_{k=1}^{k'} w_k G(x_I | \theta_{q_k}) \)

Calculate the posterior probability \( P_q(\theta_q | x_I) \) given by

\[
P_q(\theta_q | x_I) = P_q(x_I | \theta_q) \times P_q(q)
\]

// Where \( P_q(q) \) represents the prior probability of the semantic concept \( q \).

**End For**
Sort images according to $P_q(\theta_q|x_i)$.

Display to the user the top images having obtained the highest probability scores.

**End.**

---

**IV.4 Application of the proposed approach to the task of pattern recognition**

From the detailed description of our approach, it is clear that we formulate the retrieval problem as a supervised classification problem. Therefore, the proposed approach can be applied to other applications, including those of pattern recognition. Hence, we apply the proposed approach to the task of date fruit recognition/classification, i.e., recognize for a particular date sample the variety it belongs to. Similar to the process of modeling semantic concepts such as “sun”, ”apple”, “traffic”, etc, we model the concepts representing the different date varieties such as “Tinisin”, “Ajina”, “Degla bayda”, etc.

In this case, a visual appearance could be a set of images at a specific maturity level. In addition, outlier images could be those deformed because of the transportation conditions. However, it should be noted that training images used for date varieties modeling are provided by a local training database, rather than the web images.
CHAPTER IV. A SUPERVISED PROBABILISTIC APPROACH TO FIGHT AGAINST THE SEMANTIC GAP

IV.5 Conclusion

In this chapter, we have presented our approach for narrowing the semantic gap using machine learning techniques. We have presented the methods we used to model the semantic concepts i.e., mapping low-level image features with high-level semantic concepts. We have shown that the proposed approach relies mainly on a process of probabilistic supervised learning, which takes into account the intra-variation of images representing the same semantic concept. We have also presented the methods we adopted to cope with the different issues that negatively affect the modeling process such as the presence of outlier images within the training sets. At the end of the chapter, we have demonstrated how we can apply the proposed approach for the recognition task, especially for date fruit recognition.
Chapter V. A COMPUTATIONAL MULTI-TASK SYSTEM FOR IMAGE RETRIEVAL AND PATTERN RECOGNITION

V.1 Introduction

This chapter is devoted to demonstrate and prove the effectiveness and the efficiency of the proposed approach. The chapter is divided into two parts, experimental setup and experimental results. In the first part, we present the experiment conditions on which we carried out the experiments, involving image datasets, performance metrics and parameters tuning. In the second part, we report our findings along with analyzing and discussing the obtained results. The second part, in turn, is divided into two sections, in first section we report the retrieval results, whereas in the second we review the recognition results.

V.2 Experimental setup

V.2.1 Datasets

V.2.1.1 Datasets used to test retrieval

Our dataset: we have collected an image database from the Internet, where the training images are automatically downloaded from Google Images. Our dataset is made up of 3200 images from 20 different classes (160 images per class). The classes
of are respectively, buildings, butterfly, carrots, cliff, cow, elephant, forest, grape, horse, laptop, lemon, moon, pigeon, plane, potato, snow, sunset, tiger, tomato and traffic. Eighty (80) images from each class were used for training purposes, whereas, the remaining (80) are intended for retrieval testing. Figure 31 shows some representative images from our dataset.

![Representative samples from the our dataset](image)

**Figure 31.** Representative samples from the our dataset

**IAPR TC-12:** This dataset (Makadia, Pavlovic et al. 2008) is made up of 19 627 images describing several aspects from the daily life such as people, landscape, sports, etc. Images within the dataset are labeled with a total number of 291 semantic concepts with an average of 5.7 concepts per image. 17 665 images were used for training, and the rest for testing. Figure 32 shows typical samples from the IAPR TC-12.
Coil-100: Columbia University Image Library (COIL-100) dataset (Nene, Nayar et al. 1996) is composed of 7200 images from 100 different object classes (72 images per class). Figure 33 shows some representative images from COIL-100. Sixty (60) images from each class were used for leaning purposes, whereas, the remaining (12) is intended for testing.
V.2.1.2 Dataset used to test recognition

**Date dataset:** In order to investigate the performance of the proposed approach in date fruit recognition and since no date benchmark is publicly available, we introduce a new benchmark that includes 11 varieties. These varieties are Ajina, Bayd hnam, Bouaarous, Deglat sdaya, Dfar Igat, Dgoul, Litima, Hamraya, Tarmount, Tantbucht and Tinisin, respectively. The total number of samples collected was 660 (i.e., 60 samples for each variety). We take images for all the samples using a camera at a resolution of 4128*3096. Figure 34 shows typical samples from each variety. We randomly selected 440 images (40 per variety) for training purposes, while the remaining 220 images (20 per variety) were intended for testing. Three main properties make our benchmark different than the ones used in the previous studies:

- It includes the highest number of varieties (11) and it has the biggest size compared to the related studies.
- It includes some varieties with a large intra-variation, as they comprise samples in the different maturity stages (i.e., immature, semi-mature, fully mature), hardness degree and shape (Figure 35). Besides, it contains some varieties with a small inter-variation because they highly resemble each other (Figure 36).
- It contains outlier samples such as those deformed because of the transportation conditions and those which grew much more compared to their corresponding ones (Figure 37). In brief, an outlier is a sample having visual features that deviate from the common features of the samples appertaining to the same variety.

![Figure 34. Typical samples from different varieties](image)

85
Figure 35. Typical varieties with a large intra-variation. Samples within those varieties are different in terms of maturity stages, hardness degree and shape (a) Bayd hمام (b) Hamraya and (c) Tinisin.

Figure 36. Typical varieties with a small inter-variation, we can notice the high visual resemblance (a) Three samples from Bayd hمام and three others from Litima, (b) Three samples from Dfar lgat and three others from Degla bayda and (c) Three samples from Tarmount and three others from Tantbucht.

Figure 37. Some outlier samples from different varieties.
V.2.2 Performance metrics

- Performance metrics for retrieval

We use three metrics to measure the performance of our method. These metrics are:

\[
\text{Precision} = \frac{\text{Numer of relevant retrieved images}}{\text{total number of retrieved images}}
\]

\[
\text{Recall} = \frac{\text{Number of relevant retrieved images}}{\text{total number of relevant images in the dataset}}
\]

Mean Average Precision (MAP) is a metric which takes into consideration the ranking order of the retrieved images. It is defined as

\[
\text{MAP} = \frac{\sum_{q=1}^{Q} \text{Avg}P(q)}{Q},
\]

Such that \(Q\) is the number of queries, and \(\text{Avg}P(q)\) is the average precision of the query \(q\).

- Performance metrics for recognition

We adopt the recognition accuracy as a performance measure. The recognition accuracy for a single variety is defined as:

\[
\text{Accuracy}_v = \frac{\text{Number of correctly recognized samples}}{\text{Number of samples assigned to the variety}} \times 100
\]

The average recognition accuracy over all the varieties is given by:

\[
\text{Accuracy}_o = \frac{\sum_v^{N} \text{Accuracy}_v}{N}
\]

where \(N\) is the number of varieties.
V.2.3 Parameters tuning

The parameters of the proposed approach are empirically tuned for each dataset. The values assigned to each parameter are shown in Table 2. It should be noted that for the IAPR TC-12 the outlier elimination method is not applied, as we restrict ourselves to the experimental setup of the previous works. Similarly, the outlier elimination is not applied for the COIL-100 dataset simply because it doesn’t contain any outliers. Empirically, we have used Davies-Bouldin index for the recognition, while Calinski-Harabasz index is applied for retrieval.

Table 2. summarizes the values assigned to the different parameters.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>K-nearest in the PLDOF</th>
<th>Number of outliers in the PLDOF</th>
<th>$K_{min}$ in DB index</th>
<th>$K_{max}$ in DB index</th>
<th>$K_{min}$ in CH index</th>
<th>$K_{max}$ in CH index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our dataset</td>
<td>9</td>
<td>5</td>
<td>/</td>
<td>/</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>IAPR TC-12</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>Coil-100</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Date fruit dataset</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>9</td>
<td>/</td>
<td>/</td>
</tr>
</tbody>
</table>

V.3 Experimental results

V.3.1 Retrieval

V.3.1.1 Retrieval results on the web dataset

The aim of this experiment is to assess the performance of the proposed approach from different aspects. The first aspect is the strength of each feature combination, while the second aspect is to compare the performance of our approach with others which are used in the previous works. As the processing time is an important criterion to be measured in
retrieval systems, we report, in details, the processing time required by each component in our approach. We also show how the increase of scope can influence the retrieval results quality.

V.3.1.1.1 First experiment: measuring the overall performance

The scenario of the experiment is started by supplying a textual query to the system, then the proposed approach is asked to retrieve 20 relevant images to the query. To do so, we perform the retrieval as described in the previous chapter. To assure reaching the highest performance, we test the retrieval strength of each feature combination. As it is very time-consuming to test all the possible combinations, we select the ones seem to be promising. Figure 38, Figure 39 and Figure 40 show the retrieval precision per class yielded by each feature combination.

![Figure 38](image-url)  
*Figure 38. Precision per class yielded by the first subset of feature combinations*
From the several combinations we examine, we observe that there is a significant difference between the precision yielded by each of them. By taking a look on the color features, we notice that they yield a high precision for the classes which can be distinguished by color such as Tomato and Moon. Nevertheless, this is not the case for some other classes such as Horse and Cow, as they strongly resemble other classes in terms of color. Texture
features, especially the WLD, have shown a very well precision for the most of classes. As for the SIFT feature, low precision has been reached in most of the classes, which confirm the need for combining this feature with others.

By combining individual features with each other we can see that the precision has been improved for the most of classes. In addition, we notice that two combinations, namely Color + SIFT + WLD + LBP and Color + SIFT + WLD + GLCM, have scored the best precisions for many classes.

By considering the two combinations cited above, we notice that a relatively high precision has been reached, even in certain classes which contain images with multiple objects. For example, the class Traffic contains images with complicated backgrounds, but it reaches a precision of more than 90%. We can see also that the class Horse has yielded a relatively low precision (about 40%) compared to the other classes. That’s may be because of the strong visual resemblance between images belonging to this class and the other images from the other classes such as Cow.

Figure 41, Figure 42 and Figure 43 show the retrieval recall per class yielded by each feature combination for our dataset.

![Figure 41](image-url)  
**Figure 41.** Recall per class yielded by the first subset of feature combinations

91
Figure 42. Recall per class yielded by the second subset of feature combinations

Figure 43. Recall per class yielded by the third subset of feature combinations

For the sake of clarity, we observe that the recall values seem to be low because, as we have mentioned, only twenty (20) images are returned to the user. From the Figures above showing the recall per class, we notice that the recall values are consistent with the precision values i.e., whenever the precision increases the recall increases and vice versa. In order to precisely determine the best combination to be adopted, we calculate the average precision
and average recall yielded for all the classes. Table 3 shows the average precision and recall for the different combinations.

**Table 3. Average precision and recall for all the combinations**

<table>
<thead>
<tr>
<th></th>
<th>Color</th>
<th>SIFT</th>
<th>LBP</th>
<th>WLD</th>
<th>Color + SIFT</th>
<th>Color + WLD</th>
<th>Color + SIFT + WLD</th>
<th>Color + SIFT + WLD + LBP</th>
<th>Color + SIFT + LBP</th>
<th>Color + SIFT + GLCM</th>
<th>Color + SIFT + WLD + GLCM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Precision (%)</strong></td>
<td>50.75</td>
<td>21.25</td>
<td>40.25</td>
<td>46.75</td>
<td>65</td>
<td>63</td>
<td>68</td>
<td>75.25</td>
<td>69.75</td>
<td>67.25</td>
<td>72</td>
</tr>
<tr>
<td><strong>Average Recall (%)</strong></td>
<td>12.69</td>
<td>5.31</td>
<td>10.06</td>
<td>11.69</td>
<td>16.25</td>
<td>15.75</td>
<td>17</td>
<td>18.81</td>
<td>17.44</td>
<td>16.81</td>
<td>18</td>
</tr>
</tbody>
</table>

From Table 3, we see that the combination denoted as Color + SIFT + WLD + LBP has achieved the best average precision and recall as well. To confirm the strength of the proposed approach we report the precision-scope curve, as shown in Figure 44.

![Precision-Scope curve of the proposed approach](image)

**Figure 44.** Precision-Scope curve of the proposed approach

From Figure 44, we notice that precision slightly decreases whenever the scope increases. For instance, at a scope equal to 100, we see that the average precision remains relatively high (42.85 %). Indeed, this confirms the powerfulness and the strength of our approach. Figure 45, Figure 46, Figure 47 and Figure 48 show the top five images retrieved for all the classes. Note that images having a red border are considered as irrelevant.
Figure 45. Top five retrieved images for the first subset of classes

Figure 46. Top five retrieved images for the second subset of classes
V.3.1.1.2 Second experiment: comparison with other conventional methods

This experiment is dedicated to compare the performance of the proposed approach with three conventional classifiers namely, KNN, SVM and DS. In fact, we choose those classifiers, as they are used in the work of (Liu, Xu et al. 2011). However, it is worth noting
that we have compared our approach with the recently proposed ones on the well-known benchmark IPRTC -12.

**Table 4.** Comparison of the proposed approach with the conventional classifiers

<table>
<thead>
<tr>
<th></th>
<th>SVM</th>
<th>KNN</th>
<th>DS</th>
<th>Proposed approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision (%)</td>
<td>70.50</td>
<td>43.5</td>
<td>51.25</td>
<td>75.25</td>
</tr>
<tr>
<td>Recall (%)</td>
<td>17.63</td>
<td>10.87</td>
<td>12.81</td>
<td>18.81</td>
</tr>
</tbody>
</table>

From the Table 4, we notice that the proposed approach has significantly outperformed the conventional classifiers in terms of both precision and recall.

**V.3.1.1.3 Third experiment: estimation of retrieval speed**

Most users tend to use retrieval systems which accurately retrieve their targets. High speed in retrieval is another extremely-recommended criterion because user cannot wait for a long deal of time to see the retrieval results. In some particular cases, such as in the medical and military applications, the response time becomes a crucial factor. In this experiment, we give details about the required time by each component in our approach (Table 5). The values of Table 5 present the processing time required for each class (160 images).

**Table 5.** Processing time required by each component in the proposed approach (for retrieval)

<table>
<thead>
<tr>
<th>Component</th>
<th>Processing time (per second and per class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features extraction</td>
<td>302.56</td>
</tr>
<tr>
<td>Detecting the optimal number of clusters</td>
<td>0.52</td>
</tr>
<tr>
<td>Remove outliers</td>
<td>0.57</td>
</tr>
<tr>
<td>Training (GMM and EM)</td>
<td>0.0047</td>
</tr>
<tr>
<td>Retrieval</td>
<td>0.02</td>
</tr>
</tbody>
</table>

The proposed approach is composed of two main steps: training and retrieval. Processing time required for retrieval is quite important because it is carried online, while processing time required for training is less important, as this process is conducted offline. By
summing up the values in Table 5, we find that 303.65 sec (i.e., 302.56 + 0.52 + 0.57 + 0.0047) is needed to train one class. For retrieval, only 0.02 second is needed to retrieve the relevant images to a given textual query. Clearly, our proposed approach would meet the requirement of users.

V.3.1.2 Retrieval results on COIL-100

The aim of this experiment is to assess the performance of the proposed approach on the COIL-100 dataset. Because all the images in this dataset are inliers and there are no outliers, we have applied our approach without taking into consideration the outlier elimination module. After the user submitted a textual query, ten images are retrieved. Note that color features, GLCM, WLD, LBP and Hu are used to describe the dataset images. Let us first take a look on the retrieval strength of the different combinations. Figure 49 shows the retrieval strength of the different combinations in terms of precision and recall.

![Figure 49](image)

**Figure 49.** Retrieval strength of the different combinations on the COIL-100

Even though that the COIL-100 dataset assembles 100 different class of images, the first observation we can make is the high precisions yielded by the most of combinations. As a second observation, we can see that color features have yielded the best performance (precision of 87.9% and recall of 73.25%) compared to the individual features (i.e., WLD, GLCM, Hu and LBP). This confirms the powerfulness of color as a decisive feature. For texture features, we notice that the LBP has yielded better results than the WLD and the GLCM, in contrast of the results on the IAPR TC-12 dataset. In addition, Hu moments have
yielded the lowest precision may be because of the high resemblance between the images from the different classes in terms of shape. However, by combining the three features, a high precision and recall of 96.30% and 80.25% have been achieved. Figure 50 shows the top five retrieved images for the first five classes.

![Top five retrieved images for the first five classes.](image)

**Figure 50.** Top five retrieved images for the first five classes.

Similarly as in IAPR TC-12 dataset, we report the precision in terms of scope, as shown in Figure 51.

![Precision-scope curve for the COIL-100 dataset](image)

**Figure 51.** Precision-scope curve for the COIL-100 dataset
From Figure 51, we notice that the precision has drastically decreased after retrieving 10 images. Obviously, this is because the dataset contains only 12 test images for each class. In contrast, we observe that the recall increases whenever the scope increases. In addition, we see that after the scope exceeds 20 the recall converges to 1, which means that most relevant images are successfully retrieved.

V.3.1.3 Comparison with the state of the art methods using IAPR TC-12 dataset

The aim of this experiment is to compare the performance of the proposed approach with three others proposed in (Zhang and Wu 2015), (Sun, Tang et al. 2014), (Wang, Yan et al. 2009) and (Makadia, Pavlovic et al. 2008). For a fair comparison, we strictly respect the experimental setup of those methods, including features extraction, performance metrics and the dataset splitting into training and testing. As for the visual features, a large variety of local and global features were extracted from images including SIFT, DenseSIFT (extracted from multi-scale grid and Harris-Laplacian points), histogram color (from each of RGB, HSV and LAB color spaces) and Gist. For the performance measuring, the MAP metric is used, where the dataset is splitted exactly as described in section V.2.3.1. However, as the feature vector extracted from each image is of about 37200 dimensions, we perform the PCA to reduce the vectors dimensionality to 100. Note that, similarly to the COIL-100 dataset, we haven’t applied the outlier elimination module. Table 6 outlines the results yielded by our approach and the three competing methods.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP</td>
<td>34.4</td>
<td>34.1</td>
<td>33.9</td>
<td>27</td>
<td>36.22</td>
</tr>
</tbody>
</table>

From Table 6, it is apparent that the proposed approach outperforms the related methods. It outperformed the methods in (Zhang and Wu 2015) and (Sun, Tang et al. 2014) (Wang, Yan et al. 2009) with 1.82%, 2.12% and 2.32%, respectively. In addition, it
significantly outperformed the method in (Makadia, Pavlovic et al. 2008) with 9.22%. These achievements are mainly due to the modeling way we adopt which takes into consideration the intra-variation within images of the same class. Nevertheless, it should be mentioned that it is very difficult task to model concepts of the IAPRTC-12 dataset because they are strongly confused with each other. Images within this dataset are quite complicated, as they contain about 5.7 concepts per image. Figure 52 and Figure 53 show the top retrieved images by the proposed approach for ten concepts from the IAPR TC-12.

![Figure 52](image)

**Figure 52.** Top five retrieved images from the IAPR TC-12 (first subset of concepts)
Figure 53. Top five retrieved images from the IAPR TC-12 (second subset of concepts)

From the two Figures above, we can confirm that certain annotations, which are manually assigned to some images within the IAPR TC-12, are incomplete, noisy and inconsistent. For instance, we see that the third image returned for the query ‘Cliff’ is considered as irrelevant despite that it really represents a cliff, simply because its annotations (i.e., the image) do not include the concept ‘Cliff’. Similarly for the fourth image returned for the query ‘Stone’ and the fifth image returned for the query ‘Bike’. Another example which proves the inconsistence of certain annotations is the images returned for the query ‘Palm’. The second and the third images are the same (i.e., duplicated images), however, the second is found to be relevant, whereas the third is not. This is because the second is annotated with ‘Palm’, while the annotations of the third lack this concept.

V.3.2 Recognition

From the description of the proposed approach, it is apparent that we formulate the retrieval issue as a supervised classification problem. We apply our approach for solving the problem of date fruit recognition. In this case, each component within the GMM corresponding to a particular variety would assemble visually similar images e.g., images in the same maturity stage. In addition, outlier samples could be for example the ones deformed
because of the transportation conditions. In the following, we provide details about the experiments we conducted to measure the different aspects of the recognition process.

V.3.2.1 First experiment: measuring the overall performance

By referring to Figure 16, we can see that the proposed approach includes two main stages, namely learning and testing. The learning stage, in turn, includes four stages which are features extraction, detecting the optimal number of clusters, eliminating outliers and modeling. After the learning process having finished, the testing process is launched in order to assign each test sample to the suitable variety.

The present experiment aims to determine the recognition accuracy that can be achieved by our approach. For the sake of clarity, we summarize our findings in the confusion matrix shown in Table. 7

Table 7. Confusion matrix

<table>
<thead>
<tr>
<th>Target class</th>
<th>Ajina</th>
<th>Bayd hma'm</th>
<th>Bouaaros</th>
<th>DeglaBayda</th>
<th>Dfar glat</th>
<th>Dgoul</th>
<th>Hamraya</th>
<th>Litima</th>
<th>Tantbucht</th>
<th>Tarmount</th>
<th>Tinisin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ajina</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bayd hma'm</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bouaaros</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>DeglaBayda</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dfar glat</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dgoul</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hamraya</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Litima</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tantbucht</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tarmount</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tinisin</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>0</td>
</tr>
</tbody>
</table>

97.83%
First of all, we notice the high recognition accuracy (97.83\%) reached by the proposed approach in spite of the issues cited above. Those issues are respectively, the existence of outlier samples, large variation, in terms of visual features, within some varieties and the high visual resemblance between some varieties. Indeed, these issues make the recognition task more complicated, as they potentially lead to confuse some varieties with others. Nonetheless, we see that in 7 from the 11 varieties, the recognition accuracy was 100\%. Those varieties are Ajina, Bouaarous, Dgoul, Litima, Tarmount, Tantbucht and Tinisin, respectively. For Ajina variety, this high rate may be due to its big size compared to the other varieties. Nevertheless, reaching an accuracy of 100\% in Tarmount and Tantbucht is, indeed, very encouraging because samples within these varieties are quite similar. In addition, it should be mentioned that the varieties that deeply look alike have achieved high recognition rates. For instance, most samples from Litima and Bayd ham varieties which cannot, at most, be distinguished even by humans, have successfully recognized (100\% and 95.2\%, respectively). Moreover, we notice that only one sample from each of Dfar and Bayda varieties has been wrongly assigned to each other. This slight confusion is totally justified by considering the extreme visual resemblance of these latter varieties. Despite the fact that fully mature samples from Tinisin are very similar, in terms of color, to those from Dgoul, our method has successfully recognized both of them. Actually, these high rates prove the distinguishing strength of our approach and its ability to deal with such confusing varieties.

**V.3.2.2 Second experiment: measuring the recognition speed**

In this experiment, we estimate the processing time needed for recognition. In most cases, it is boring for user to wait for a long deal of time in order to get the recognition results. Therefore, the processing time is amongst the most important criterions to be taken into consideration in assessing recognition systems. This criterion becomes quite crucial for the real-time recognition systems. Table 8 describes the major components of our system and the processing time required by each of them.
Table 8. Processing time required by the major components of the proposed system

<table>
<thead>
<tr>
<th>Component</th>
<th>Processing time (per second and per variety)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features extraction</td>
<td>6.71 (0.11 per image)</td>
</tr>
<tr>
<td>Detecting the optimal number of clusters</td>
<td>0.14</td>
</tr>
<tr>
<td>Remove outliers</td>
<td>0.15</td>
</tr>
<tr>
<td>Training (GMM+EM)</td>
<td>0.73</td>
</tr>
<tr>
<td>Recognition</td>
<td>0.00034 (for 20 test images)</td>
</tr>
</tbody>
</table>

The processing time required for training is less important than that of testing because the first is performed offline, while the second is performed online. By summing up the values in Table 8, we find that only 5.42 sec (i.e., \((0.11 \times 40) + 0.14 + 0.15 + 0.73\)) is needed to train one variety (i.e., 59.62 sec to train the 11 varieties). To recognize a sample, two processes are needed, features extraction which cost 0.11 seconds for one image and recognition, which cost 0.00001 sec (i.e., \(0.00034/20\)). Hence, our proposed approach would satisfy the exigencies of a real-time recognition system.

V.3.2.3 Third experiment: comparison with state of the art methods

Although the studies concerned with date recognition are very scarce, we devote this experiment to compare the proposed approach with three others, which are proposed in (Ghulam, 2015), (Haidar et al., 2012) and (Fadel, 2007). Table 9 outlines the three methods and their experimental settings.

Table 9. Experimental settings of the related methods

<table>
<thead>
<tr>
<th>Features</th>
<th>(Ghulam, 2015)</th>
<th>(Haidar et al., 2012)</th>
<th>(Fadel, 2007)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
<td>Weber Local Descriptor (WLD), Local Binary Patterns (LBP)</td>
<td>Color mean and standard deviation (from the 3 RGB channels), Area, Perimeter, Eccentricity, Major and Minor axis length, Entropy (from the 3 RGB channels) and Energy measures extracted from the Grey-Level Co-occurrence Matrix (GLCM)</td>
<td>Color mean and standard deviation, for each of the three RGB channels</td>
</tr>
<tr>
<td>Classifier</td>
<td>Support Vector Machine (SVM)</td>
<td>Artificial Neural Network (ANN)</td>
<td>Probabilistic Network (PNN)</td>
</tr>
<tr>
<td>Size of the used dataset</td>
<td>200 images</td>
<td>140 images</td>
<td>200 images</td>
</tr>
<tr>
<td>Number of varieties</td>
<td>4</td>
<td>7</td>
<td>5</td>
</tr>
</tbody>
</table>
From the table above, we can observe the substantial differences between our experimental settings and those of the mentioned methods. First, the number of images used in each of them is very limited, which is practically insufficient to really investigate the method performance. Second, the number of varieties is too little, while, in real life practical applications much more others have to be recognized. Table 10 shows the average recognition accuracy yielded by each method.

Table 10. Average recognition accuracy yielded by each method

<table>
<thead>
<tr>
<th>Method</th>
<th>(Ghulam, 2015)</th>
<th>(Haidar et al., 2012)</th>
<th>(Fadel, 2007)</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average accuracy (%)</td>
<td>55.63</td>
<td>93.27</td>
<td>64.20</td>
<td>97.83</td>
</tr>
</tbody>
</table>

From Table 10, we notice that our approach has significantly outperformed the methods of (Ghulam, 2015) and (Fadel, 2007) with 42.2% and 33.63%, respectively. In addition, it exceeds the method in (Haidar et al., 2012) with 4.56%. The noticeable difference in the recognition accuracy between the previous methods and ours can be attributed to several reasons. First of all, we believe that the main reason behind that is the inability of these methods to properly deal with the emerging challenges from the dataset we introduced in this thesis. Those methods aren’t capable to deal with the confusing varieties as well as the large variation within some varieties. Besides, they don’t take into account the presence of outlier samples, which could decrease the recognition accuracy.

To check the above interpretations, we report the accuracy per variety yielded by each method, as shown in Figure 54. For the method of (Haidar, 2012), we note that the lowest accuracies are yielded by Bayd hhamm, Dfar Igat, Tarmount and Litima varieties. Similarly, in the method of (Fadel, 2007), we observe that the lowest accuracies are those yielded by Bayd hhamm, Degla bayda, Dfar Igat and Tarmount. Thus, the problem arises with the varieties having roughly the same visual characteristics. Therefore, these methods failed to distinguish between the highly similar varieties, which confirms our interpretations. For the method of (Ghulam, 2015), the accuracies yielded by most of the varieties are comparable and very low, which may be because of the lack of color descriptors with the used features.
CHAPTER V. A COMPUTATIONAL MULTI-TASK SYSTEM FOR IMAGE RETRIEVAL AND PATTERN RECOGNITION

Figure 54. The accuracy per variety yielded by each method

V.3.2.4 Fourth experiment: test the recognition strength of different combinations

The current experiment aims to examine the recognition strength of each of the features we have used. The accuracy per variety yielded by each feature combination is shown in Figure 55. The first note we can make is that the fusion of all the features has yielded, in most of the varieties, the highest accuracy. In addition, we can see that Ajina variety has achieved a rate of 100% for all the possible combinations. This is because the samples appertaining to this variety have particular shape, texture and size. Color features have shown a well performance for most of the varieties, which proves the ability of color as decisive features. However, color features have yielded a low accuracy for Dfar lgat (57.58%) because using color alone is not enough to distinguish this variety from the deeply similar varieties like Degla Bayda. For the GLCM, we note that the accuracy scored by Litima was 65.52%, which can be explained by the fact that texture of this latter variety is analogous with that of some others such as Bayd hnam. Meanwhile, GLCM has reached a high accuracy in Tarmount (100%) and Dfar lgat (90%).
C H A P T E R V. A Computational Multi-Task System for Image Retrieval and Pattern Recognition

Figure 55. The accuracy per variety yielded by each combination.

Shape features have performed well in some varieties such as Tantbucht (100%), but fail in some others such as Degla Bayda (59.26%) and Dgoul (55.56%), as samples appertaining to the both varieties have often the same shape and size. By incorporating the GLCM, the accuracy has jumped to 90.48% and 100%, respectively.

Table 11 shows the average accuracy of each combination. From this table, we observe that the lowest accuracy is that of shape features (85.18%). Meanwhile, color features have demonstrated their importance by reaching an average accuracy of 86.97%. In addition, combination of color with shape has yielded better than the combination of color with the GLCM. By combining all the features we gain a recognition rate of 97.83%.

Table 11. Average recognition accuracy yielded by each combination

<table>
<thead>
<tr>
<th>Combination</th>
<th>Color</th>
<th>GLCM</th>
<th>Shape</th>
<th>Color+GLCM</th>
<th>Color+Shape</th>
<th>GLCM+Shape</th>
<th>All features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average accuracy (%)</td>
<td>86.97</td>
<td>85.89</td>
<td>85.18</td>
<td>92.66</td>
<td>95.79</td>
<td>93.50</td>
<td>97.83</td>
</tr>
</tbody>
</table>

V.4 Conclusion

In this chapter, we have tested the effectiveness of the proposed approach and we have reported the experimental results. We have firstly started by reviewing the experimental setup including the used datasets, performance metrics and the parameters tuning. Then, we
have reported our findings on both image retrieval and pattern recognition. After conducted several experiments, the reported results prove several facts. First, the efficiency of our approach has been clearly shown. Second, both retrieval and recognition have been performed with a high speed. Third, the proposed approach has significantly outperformed the state-of-the-art methods. Fourth, by testing several features combinations, the robustness of the proposed approach has been successfully demonstrated.
Content-Based Image Retrieval (CBIR) systems use low-level image features in order to identify the relevant images for a particular query. However, matching images based on their visual features could probably lead to misunderstanding the user intention, as visual features solely are not capable to reflect the semantics contained in the images. This misunderstanding or contradiction is commonly-known as the semantic gap issue. In the literature, several methods have been proposed for mitigating this gap. Each of which adopt a distinct technique to beat the problem. In spite of the growing number of works concerned with reducing the semantic gap, much more works are remains to be accomplished.

In this thesis, we have detailed the background of our work involving features extraction, query formulation and similarity measures. We have deeply studied the three main issues of CBIR, namely, feature selection, page zero and the semantic gap issue. In order to efficiently reduce the semantic gap, we have studied and reviewed, in details, the existing related works. We have also shown that those works belong to three different categories, namely, region-based, auto-annotation and relevance feedback.

In this thesis, we have proposed a novel approach for automatic mitigating the semantic gap in CBIR. We have used various and diverse techniques for this end. To perform retrieval, training images are gathered from the web. We have utilized several local and global visual features to describe images. To alleviate the negative effect of the existence of outliers, we have eliminated them using the PLDOF method. Then, we have used clustering validity
indices to discover the visual appearances within images representing the same concept. A Gaussian Mixture Model (GMM) is used to model each semantic concept.

The main contributions of the proposed approach reside mainly in taking two quite important aspects into account, which are the presence of outlier images and the intra-variation of images representing the same concept. Moreover, the proposed approach presents diverse pros: it is fully automatic, it is not limited to any per-defined semantic concepts, it allows the user to use a textual query in retrieving images from the unlabeled image collections.

We formulate the retrieval as a supervised probabilistic classification problem. Hence, it is possible to apply the approach to solve pattern recognition problems. We apply the proposed approach to the task of date fruit recognition/classification, i.e., recognize for a particular date sample the variety it belongs to. Indeed, the process of modeling semantic concepts such as sun, apple, cow, is analogous to modeling the concepts representing the different date varieties such as Tantbucht and Ajina.

To prove the efficiency of the proposed approach, we have conducted extensive experiments. By viewing the experimental results, we can make the following observations:

- The proposed approach has successfully treated the problem of outlier images.
- The proposed approach has improved the modeling process by considering the intra-variation of semantic concepts.
- The proposed approach performs both retrieval and recognition with a high speed.
- The proposed approach has achieved a high retrieval precision as well as a high recognition rate.
- The proposed approach has proven its strength against several state-of-the-art methods.

Further investigations can be carried out to improve the proposed approach. Because the approach is composed of different component, improvements can be done on the level of each component. For instance, it is well-known that the K-means converges slowly, thus using other clustering algorithms could speed up the process of modeling. In addition, searching for other feature combinations could improve the retrieval and the recognition rates.
Furthermore, one can investigate the possibility of retrieving images using a set of concepts instead of only one concept. This can, for example, be done by combining the GMMs representing the different semantic concepts.
Bibliography


Gong, T., S. Li and C. L. Tan (2010). A semantic similarity language model to improve automatic image annotation. 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, IEEE.


Khaledi, B. and M. Kherfi (2016). "Modified integrative color intensity co-occurrence matrix for texture
image representation." Journal of Electronic Imaging 25(5).


Pearson, K. "On lines and planes of closest fit to points in space." Philosophical Magazine. v2: 559-572.


BIBLIOGRAPHY


Xu, H., P. Pan, Y. Lu, C. Xu and D. Chen (2014). Improving automatic image annotation with google semantic link. Semantics, Knowledge and Grids (SKG), 2014 10th International Conference on, IEEE.


